UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
DOUTORADO EM CIÊNCIA DA INFORMAÇÃO

GUSTAVO DINIZ DO NASCIMENTO

DIRETRIZES PARA A INDEXAÇÃO SEMIAUTOMÁTICA DE PRODUÇÕES CIENTÍFICAS POR MEIO DE SINTAGMAS NOMINAIS

JOÃO PESSOA
2025

GUSTAVO DINIZ DO NASCIMENTO

DIRETRIZES PARA A INDEXAÇÃO SEMIAUTOMÁTICA DE PRODUÇÕES CIENTÍFICAS POR MEIO DE SINTAGMAS NOMINAIS

Tese apresentada ao curso de Doutorado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba (PPGCI/UFPB), como requisito para obtenção do título de Doutor em Ciência da Informação.

Linha de Pesquisa: Organização, acesso e uso da informação

Orientadora: Prof.ª Dr.ª Maria Elizabeth Baltar Carneiro de Albuquerque

Coorientadora: Prof.ª Dr.ª Raimunda Fernanda dos Santos

JOÃO PESSOA
2025

Catalogação na publicação
Seção de Catalogação e Classificação

N244d Nascimento, Gustavo Diniz do.
Diretrizes para a indexação semiautomática de publicações científicas por meio de sintagmas nominais / Gustavo Diniz do Nascimento. - João Pessoa, 2025.
211 f. : il.

Orientação: Maria Elizabeth Baltar Carneiro de Albuquerque.
Coorientação: Raimunda Fernanda dos Santos.
Tese (Doutorado) - UFPB/CCSA.

1. Indexação semiautomática. 2. Sintagmas nominais. 3. Representação temática da informação. I. Albuquerque, Maria Elizabeth Baltar Carneiro de. II. Santos, Raimunda Fernanda dos. III. Título.

UFPB/BC CDU 025.4(043)

Elaborado por RUSTON SAMMEVILLE ALEXANDRE MARQUES DA SILVA - CRB-15/0386

GUSTAVO DINIZ DO NASCIMENTO

DIRETRIZES PARA A INDEXAÇÃO SEMIAUTOMÁTICA DE PRODUÇÕES CIENTÍFICAS POR MEIO DE SINTAGMAS NOMINAIS

Tese apresentada ao curso de Doutorado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba (PPGCI/UFPB), como requisito para obtenção do título de Doutor em Ciência da Informação.

BANCA EXAMINADORA

Prof.ª Dr.ª Maria Elizabeth Baltar Carneiro de Albuquerque (Orientadora)
Universidade Federal da Paraíba (UFPB)

Prof.ª Dr.ª Gracy Kelli Martins (Examinadora interna)
Universidade Federal da Paraíba (UFPB)

Prof.ª Dr.ª Virgínia Bentes Pinto (Examinadora interna)
Universidade Federal da Paraíba (UFPB)

Prof.ª Dr.ª Rosane Suely Alvares Lunardelli (Examinadora externa)
Universidade Estadual de Londrina (UEL)

Prof. Dr. José Antonio Moreiro González (Examinador externo)
Universidade Carlos III de Madrid (UC3M)

Prof.ª Dr.ª Izabel França de Lima (Suplente interna)
Universidade Federal da Paraíba (UFPB)

Dr. Sale Mário Gaudêncio (Suplente externo)
Universidade Federal Rural do Semi-Árido (UFERSA)

JOÃO PESSOA
2025

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

ATA DE DEFESA DE TESE

Defesa nº 110

Ata da Sessão Pública de Defesa de Tese do(a) Doutorando(a) **GUSTAVO DINIZ DO NASCIMENTO** como requisito para obtenção do grau de Doutor(a) em Ciência da Informação, Área de Concentração em Informação, Conhecimento e Sociedade e com Linha de pesquisa em Organização, acesso e uso da informação.

No quinto dia do mês de fevereiro de dois mil e vinte e cinco (05/02/2024), das quatorze horas às dezessete horas e quinze minutos, na sala virtual do Google Meet, conectaram-se via videoconferência a banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação para avaliar o(a) candidato(a) ao Grau de Doutor(a) em Ciência da Informação na Área de Concentração Informação, Conhecimento e Sociedade, o(a) doutorando(a) **GUSTAVO DINIZ DO NASCIMENTO.** A defesa ocorreu de forma remota, com acesso por meio do link: meet.google.com/ajw-atpe-yxk**.** A banca examinadora foi composta pelos (as) professores(as): Dra. Maria Elizabeth Baltar Carneiro de Albuquerque – PPGCI/UFPB (Presidente/Orientadora), Dra. Raimunda Fernanda dos Santos – UFRN (Coorientadora), Dra. Gracy Kelli Martins Gonçalves – PPGCI/UFPB (Examinadora interna), Dra. Virgínia Bentes Pinto – PPGCI/UFPB (Examinadora interna), Dra. Rosane Suely Alvares Lunardelli - PPGCI/UEL (Examinadora externa), Dr. José Antonio Moreiro González - Universidad Complutense de Madrid (Examinador externo), Dra. Izabel França de Lima – PPGCI/UFPB (Suplente Interna) e Dr. Sale Mário Gaudêncio – UFERSA (Suplente Externo). Dando início aos trabalhos, o(a) Professor(a) Dr(a). Maria Elizabeth Baltar Carneiro de Albuquerque, Presidente(a) da Banca Examinadora, explicou aos presentes a finalidade da sessão e passou a palavra ao(à) discente para que fizesse oralmente a apresentação do trabalho de tese intitulado: "**DIRETRIZES PARA A INDEXAÇÃO SEMIAUTOMÁTICA BASEADA EM SINTAGMAS NOMINAIS: UMA PROPOSTA PARA ÁREA DA REPRESENTAÇÃO TEMÁTICA DA INFORMAÇÃO".** Após a apresentação, o(a) doutorando(a) foi arguido(a) na forma regimental pelos examinadores. Respondidas todas as arguições, o(a) Professor(a) Dr.(a). Maria Elizabeth Baltar Carneiro de

Albuquerque, Presidente(a) da Banca Examinadora, acatou todas as observações da banca e procedeu para o julgamento do trabalho, concluindo por atribuir-lhe o conceito:

(X) Aprovado ( ) Insuficiente ( ) Reprovado.

Observações da Banca: A banca deliberou pelo ajuste do título e apresentou sugestões nos pareceres em anexo.

Proclamados os resultados e encerrados os trabalhos, eu, Professor(a) Dr.(a) Maria Elizabeth Baltar Carneiro de Albuquerque, Presidente da Banca Examinadora, lavrei a presente ata que segue assinada por mim pelos(as) participantes da banca, juntamente com os pareceres de avaliação da tese e da defesa de tese do(a) doutorando(a), devidamente assinados por seus respectivos avaliadores e em formato digital.

João Pessoa, 05 de fevereiro de 2025.

gov.br Documento assinado digitalmente
MARIA ELIZABETH BALTAR CARNEIRO DE ALBU
Data: 05/02/2025 18:30:13-0300
Verifique em https://validar.iti.gov.br

Dra. Maria Elizabeth Baltar Carneiro de Albuquerque
Presidente/ Orientador (a) – PPGCI/UFPB

gov.br Documento assinado digitalmente
RAIMUNDA FERNANDA DOS SANTOS
Data: 10/02/2025 10:29:46-0300
Verifique em https://validar.iti.gov.br

Dra. Raimunda Fernanda dos Santos
Coorientador(a) – PPGCI/UFPB

gov.br Documento assinado digitalmente
GRACY KELLI MARTINS GONCALVES
Data: 12/02/2025 11:30:16-0300
Verifique em https://validar.iti.gov.br

Dra. Gracy Kelli Martins Gonçalves
Examinador (a) Interno (a) – PPGCI/UFPB

gov.br Documento assinado digitalmente
VIRGINIA BENTES PINTO
Data: 14/02/2025 09:56:13-0300
Verifique em https://validar.iti.gov.br

Dr. Virgínia Bentes Pinto
Examinador(a) Interno(a) – PPGCI/UFPB

ASSINADO DIGITALMENTE POR
safeID Rosane Suely Alvares Lunardelli
CPF: 439.202.419-91
ICP Brasil

Dra. Rosane Suely Alvares Lunardelli
Examinador (a) Externo (a) – UFPB

Firmado por MOREIRO GONZALEZ JOSE ANTONIO - ***8182** el día 06/02/2025 con un certificado emitido por AC FNMT

Dr. José Antonio Moreiro González
Examinador (a) Externo (a) – Universidad Complutense de Madrid

Dra. Izabel França de Lima
Suplente Interno(a) –PPGCI/UFPB

Dra. Sale Mário Gaudêncio
Suplente Externo (a) – UFPB

Documento assinado digitalmente
gov.br
GUSTAVO DINIZ DO NASCIMENTO
Data: 18/02/2025 14:50:57-0300
Verifique em https://validar.iti.gov.br

Gustavo Diniz do Nascimento
Doutorando(a)

AGRADECIMENTOS

Agradeço à professora Maria Elizabeth Baltar Carneiro de Albuquerque a orientação, ajuda, apoio, paciência, na trajetória acadêmica do doutorado e na construção desta tese. Não posso deixar de agradecer à professora Raimunda Fernanda dos Santos a coorientação, cuja participação foi essencial nessa trajetória acadêmica.

Meus agradecimentos se estendem aos professores do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba. Os diálogos e ideias em torno da Organização e Representação da Informação e do Conhecimento, durante o curso das disciplinas, permitiram a elaboração e publicação de produções científicas para a área e para a temática abordada nesta tese.

Agradeço às professoras Gracy Kelli Martins, Virgínia Bentes Pinto, Mariângela Spotti Lopes Fujita e ao professor José Antonio Moreiro González a confiança e a participação como membros da banca de qualificação desta pesquisa. Grato pelas pertinentes contribuições tão importantes para esta tese.

Meu honrado agradecimento às professoras Gracy Kelli Martins, Virgínia Bentes Pinto, Rosane Suely Alvares Lunardelli e ao professor José Antonio Moreiro González por não medirem esforços para se fazerem presentes e participarem como membros da banca de defesa desta tese. Grato por dividirem comigo este momento tão importante em minha trajetória acadêmica.

RESUMO

Este é um estudo que envolve aspectos concernentes à indexação semiautomática e aos sintagmas nominais com vistas a contribuir para as práticas de representação e recuperação da informação, bem como para as pesquisas no campo da Ciência da Informação e áreas afins. Nesse sentido, esta pesquisa tem como objetivo propor diretrizes para a indexação semiautomática de produções científicas por meio de sintagmas nominais. Para tanto, identificou-se na literatura científica nacional e internacional estratégias de indexação semiautomática; investigou-se a composição e as potencialidades dos sintagmas nominais enquanto unidades portadoras de informação; analisou-se o funcionamento de *software* PALAVRAS na identificação/extração de sintagmas nominais por meio da análise de artigos científicos da área de Representação Temática da Informação; e traçaram-se procedimentos para a identificação semiautomática de SN que funcionem como elementos descritores do domínio da Representação Temática da Informação. Utilizou-se como metodologia as pesquisas bibliográfica, documental, exploratória e descritiva com abordagem qualiquantitativa. O universo da investigação compreende os artigos científicos apresentados no GT2 do XXII Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, edição 2022. Buscou-se subsídio teórico-metodológico no potencial semântico dos sintagmas nominais aplicados na indexação semiautomática como unidades portadoras de significados precisos e específicos e que apresentam viabilidade no contexto da representação temática da informação. Como resultados, destacou-se que a semiautomação da indexação associada ao uso dos sintagmas nominais oferece uma alternativa viável no contexto de crescente produção de informação em ambiente digital e evidenciou-se a riqueza semântica dos sintagmas nominais, os quais podem ser utilizados como pontos de acesso ao documento de modo mais ágil e objetivo, proporcionando uma economia de tempo da prática da indexação, com o auxílio do processamento automático da informação, sem com isso eximir o olhar humano em tarefa subjetiva por natureza. Também, constatou-se que a associação da indexação semiautomática com o uso dos sintagmas nominais se mostrou alternativa promissora para a indexação de documentos, uma vez que propõe diretrizes que buscam nortear a construção de metodologias de indexação semiautomática que utilizam as potencialidades dos sintagmas nominais e apresenta um percurso para a análise temática semiautomática de textos, com base em critérios preestabelecidos e validados, com vistas a alcançar, em curtos espaços de tempo, descritores documentais representativos de documentos em unidades de informação e sistemas de recuperação da informação. Logo, a associação da indexação humana à automação e ao uso dos sintagmas nominais com unidades portadoras de informação se mostra uma alternativa promissora e viável para a representação temática da informação hodiernamente.

Palavras-chave: indexação semiautomática. sintagmas nominais. representação temática da informação.

ABSTRACT

This study involves aspects concerning semi-automatic indexing and nominal syntagms intending to contribute to information representation and retrieval practices and research in the field of Information Science and related areas. From this point of view, this research aims to propose guidelines for the semi-automatic indexing of scientific productions through nominal syntagms., considering the theoretical-methodological background of manual indexing and semi-automatic indexing and the potential of Nominal Syntagms. For that, semi-automatic indexing strategies were identified in the national and international scientific literature; the composition and potential of nominal syntagms as information-carrying units were investigated; the functioning of the PALAVRAS *software* in the identification/extraction of nominal syntagms was analyzed through the analysis of scientific articles in the area of Thematic Representation of Information; and procedures were outlined for the semi-automatic identification of NSs that function as descriptor elements of the domain of Thematic Representation of Information. The methodology used was bibliographical, documentary, exploratory and descriptive research with a qualitative and quantitative approach. The research comprises the scientific articles presented in TG2 of the XXII National Meeting of Research and Postgraduate Studies in Information Science, 2022 edition. The theoretical and methodological support was sought in the semantic potential of nominal syntagms applied in semi-automatic indexing as units that carry precise meanings and are viable in the thematic representation of information. As a result, it was highlighted that the semi-automation of indexing associated with the use of nominal syntagms offers a feasible alternative in the context of increasing information production in a digital environment and the semantic richness of nominal syntagms was evidenced, which can be used as access points to the document in a more agile and objective way, providing a reducing indexing time, with the help of automatic information processing, without freeing the human eye in a task that is subjective by nature. It was also found that the association of semi-automatic indexing with the use of nominal syntagms proved to be a promising alternative for document indexing since it proposes guidelines that seek to guide the construction of semi-automatic indexing methodologies that use the potential of nominal syntagms and presents a path for the semi-automatic thematic analysis of texts, based on pre-established and validated criteria, to achieve, in a short space of time, representative document descriptors of documents in information units and information retrieval systems. Hence, the association of human indexing with automation and the usage of nominal syntagms with information-bearing units proves to be a promising and effective alternative for the thematic representation of information in the digital age.

Keywords: semiautomatic indexing. nominal syntagms. thematic representation of information.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BDTD - BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES

BRAPCI - BASE DE DADOS REFERENCIAL DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO

CAPES - PORTAL DE PERIÓDICOS DA COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR

CDD - CLASSIFICAÇÃO DECIMAL DE DEWEY

CI - CIÊNCIA DA INFORMAÇÃO

ENANCIB - ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO

GT - GRUPO DE TRABALHO

KWOC - *KEYWORD-IN-CONTEXT* (KWIC) E *KEYWORD-OUT-OF-CONTEXT*

LD - LINGUAGEM (NS) DOCUMENTÁRIA (S)

NEPHIS - *NESTED-PHRASE INDEXING SYSTEM*

OC - ORGANIZAÇÃO DO CONHECIMENTO

OI - ORGANIZAÇÃO DA INFORMAÇÃO

PRECIS - *PRESERVED CONTEXT INDEX SYSTEM*

RC - REPRESENTAÇÃO DO CONHECIMENTO

RI - REPRESENTAÇÃO DA INFORMAÇÃO

SN - SINTAGMA(S) NOMINAL (IS)

SIBI/USP - SISTEMA INTEGRADO DE BIBLIOTECAS/UNIVERSIDADE DE SÃO PAULO

SOC - SISTEMAS DE ORGANIZAÇÃO DO CONHECIMENTO

SLIC - LISTAGEM SELETIVA EM COMBINAÇÃO

SPIRIT - *SYSTÉME SYNTAXIQUE ET PROBABILISTE D´INDEXATION ET DE RECHERCHE D´INFORMATICOS TEXTUELLES*

SRI - SISTEMA DE RECUPERAÇÃO DA INFORMAÇÃO

TGT - TEORIA GERAL DA TERMINOLOGIA

TTI - TRATAMENTO TEMÁTICO DA INFORMAÇÃO

UIS - UNIDADE E INFORMAÇÃO

UNISIST - SISTEMA MUNDIAL DE INFORMAÇÃO CIENTÍFICA

SUMÁRIO

1 INTRODUÇÃO

A atual conjuntura social é marcada pelo excesso de informação, mormente no ambiente virtual. A grande quantidade de informação disponibilizada à sociedade está associada ao desenvolvimento da ciência e da tecnologia, o que foi chamado por alguns autores de "explosão informacional" ou "*boom* informacional". Tal fato é caracterizado pelo surgimento de uma grande massa documental, especialmente após a segunda guerra mundial, momento em que a produção de informação se tornou elemento essencial à tomada de decisões.

Toda a conjuntura vivenciada após a segunda guerra mundial levou ao surgimento da Ciência da Informação (CI), a qual pode ser compreendida como um campo que se dedica ao estudo das propriedades e do comportamento da informação, bem como às forças que governam seu fluxo e os meios de processamento para viabilizar de forma eficaz a sua acessibilidade e utilização (Borko, 1968).

No contexto da recuperação da informação, sobretudo hodiernamente, com o aumento considerável da produção de informação em meio digital, um desafio emergente é organizar essa informação de tal forma que seja recuperada pelo usuário que dela necessita em curtos espaços de tempo. No contexto da CI, em especial no âmbito da Organização e Representação da Informação, destaca-se a representação temática da informação, a qual se centra na representação dos conteúdos dos objetos informacionais para fins de sua recuperação (Guimarães; Sales; Grácio, 2012). A CI, por sua vez, se dedica ao estudo dos diversos aspectos que envolvem a informação, abrangendo "a origem, coleta, organização, estocagem, recuperação, interpretação, transmissão, transformação e uso da informação" (Borko, 1968, p. 3).

Outro enfoque de estudo que tem sido trabalhado no contexto da Ciência da Informação e áreas como a Ciência da Computação é a Recuperação da Informação, haja vista o seu papel singular para os sistemas de informação. Enquanto processo, a Recuperação da Informação está relacionada a um conjunto de etapas que visam a busca e a localização de objetos informacionais para fins de satisfazer as necessidades dos usuários. A recuperação se executa, de modo geral, por meio da comparação entre as representações dos documentos e as representações dos usuários no momento da busca, ou seja, o cotejamento e, por conseguinte, a semelhança parcial ou total entre essas duas representações serão responsáveis pela recuperação ou não dos documentos, logo o sucesso nos Sistemas de Recuperação de Informação (SRIs) está diretamente atrelado à qualidade da representação feita no momento

de entrada de um documento no sistema, bem como na expressão de busca realizada pelo usuário.

É evidente que, não obstante a CI venha desempenhando estudos nessa área específica, a recuperação de informação na atual conjuntura ainda não é uma tarefa livre de dificuldades e de inconvenientes. Recuperar informação relevante na atual conjuntura se tornou uma atividade difícil devido à gama de informações que é disponibilizada nos vários suportes informacionais que, por sua vez, se encontram nas múltiplas instituições que guardam, organizam e disponibilizam informações tanto em meio físico como em meio virtual.

Dentre as atividades que contribuem para a representação e a organização da informação, destaca-se a "indexação[1]". Quando realizada manualmente, essa atividade é realizada por um profissional que se utiliza das etapas de análise de assunto e tradução do documento em termos que representam o seu conteúdo para fins de sua recuperação. Segundo Andrade, Cruz, Ferneda e Fujita (2022, p. 270), "[...] o sucesso da recuperação da informação é reflexo dos esforços realizados pelas bibliotecas no tratamento e organização da informação. A demanda informacional dos usuários deve ser compreendida e acatada para traçar planos para solucionar qualquer tipo de lacuna no sistema de organização". Nesse contexto, à indexação, segundo Andrade, Cruz, Ferneda e Fujita (2022, p. 270), "cabe a função de descrever o conteúdo dos documentos com coerência aos interesses organizacionais e da comunidade de usuários e, principalmente, concentrando-se na recuperação da informação".

Com o avanço na produção de informações em ambientes digitais, surgem algumas alternativas voltadas para o processamento e a representação do grande volume de informações, a exemplo das técnicas de indexação automática e semiautomática. Os estudos acerca da indexação automática, desenvolvidos inicialmente com base nas palavras isoladas que se encontravam nos textos, apresentaram alguns inconvenientes, algumas limitações, mormente, no que se refere a fenômenos específicos da língua, como a polissemia e a sinonímia.

Nesse contexto, desenvolveram-se estudos voltados para a indexação automática baseada nos Sintagmas Nominais (SN), os quais, diferentemente das palavras isoladas de um texto, se constituem nas menores partes dos discursos portadores de informação (Kuramoto, 2002). Para Perini (2010), o sintagma nominal, diferentemente de outros sintagmas da língua, possui potencial referencial, que é a sua propriedade semântica básica, condicionado ao modo

[1] O termo "indexação" é utilizado em vários domínios do conhecimento, por exemplo, na economia, na demografia, no comércio e na Ciência da Informação. Em nosso caso, é utilizado dentro do contexto da Ciência da Informação.

como ele é construído internamente. Assim, não é possível fazer referência a uma entidade do mundo usando a língua sem que se use algum sintagma nominal.

O Sintagma Nominal (SN) é a menor unidade linguística, com sentido específico, que faz referência a um ser, a um objeto, a um processo etc. Quando o SN é formado por mais de uma palavra, elas mantêm uma relação de dependência com o núcleo, assim, os sintagmas nominais são estruturas linguísticas que se encontram dentro das frases e textos e que possuem sentidos mais específicos quando comparados a palavras isoladas, soltas, descontextualizadas. Observe-se o exemplo da palavra "recuperação", isoladamente, pode-se interpretá-la dentro de vários contextos: saúde, economia etc. Não obstante, ao observar-se o sintagma nominal "a recuperação da informação", fica nítido que se estar a falar de recuperação em um contexto específico, seja em uma área do conhecimento ou em um ambiente informacional, por exemplo.

Pinto (2000, p. 65), acerca dos Sintagmas Nominais como índices, diz que "[...] os índices serão constituídos por passagens do texto portadoras de informação, neste caso pode-se ter uma representação mínima do conteúdo do documento à medida que esses grupos não são isolados do contexto no qual eles são inseridos (onde eles têm um valor referencial)".

Alguns estudos recentes dedicaram-se ao uso dos Sintagmas Nominais[2], considerando o potencial semântico dessas unidades, as quais vêm se mostrando uma alternativa promissora nas atividades de representação temática para recuperação de informação. Os seguintes autores trabalham direta e indiretamente com os Sintagmas Nominais em suas pesquisas sob a ótica da indexação e recuperação da informação ou na construção de métodos de identificação e extração desses sintagmas. São eles: Le Guern (1991), Kuramoto (1995; 2002), Miorelli (2001), Othero (2004), Souza (2005; 2006), Maia (2008), Morellato (2010), Corrêa *et al*. (2011), Lopes (2012), Silva (2014), Souza e Raghavan (2014), Silva e Correa (2015), Nascimento (2015), Corrêa e Bazílio (2017), Nascimento e Corrêa (2018; 2019), Corrêa e Celerino (2019), Silva e Corrêa (2020), entre outros.

De acordo com Corrêa e Lapa (2013) e Bandim e Corrêa (2018), os principais *softwares* de indexação automática utilizados no Brasil são o OGMA, o AUTOMINDEX e o SISA. Os dois primeiros extraem os termos presentes nos documentos, ou seja, realizam a

[2] Pinto (2005, p. 65) evidencia que "A representação do conhecimento registrado, tendo em vista a indexação de documentos, pode ser realizada tomando-se por base os conceitos / palavras-chave / unitermos, ou ainda, em uma visão mais moderna, os sintagmas nominais (proposta apresentada pelo grupo SYDO), ou frases (proposta de Alain F. Smeaton e Paraic Sheridan), ou ainda os sintagmas verbais (proposta de Geneviève Lallich e de Virginia SEM ACENTO? Bentes Pinto)".

indexação automática por extração. O *software* SISA, por sua vez, possibilita a realização da indexação automática por atribuição a partir do uso de um vocabulário controlado externo.

Na indexação automática, baseada na extração das palavras que compõem os documentos, há uma desconstrução do texto, do discurso do autor, uma vez que as palavras isoladas se tornam descontextualizadas, passando a não ter significado específico ou podendo ter vários, comprometendo, assim, a representação temática fidedigna dos documentos. Em contrapartida, os SN são passagens dos textos, carregam, portanto, sentido, visto que são unidades significativas portadoras de informação, que fazem referência a algo ou a alguém de modo mais preciso e específico.

São vários os fatores levantados por pesquisadores da área de CI que evidenciam a necessidade emergente de uma indexação automática: suprir a morosidade na indexação manual, dar conta da demanda de informação em meio digital, minorar os problemas suscitados pela subjetividade presente na indexação manual etc. Por outro lado, é imperativo ressaltar o quão pertinente é o papel do humano na atividade de indexação em unidades de informação[3] e sistemas de recuperação de informação, haja vista a sua capacidade de capturar nuances e relações entre conceitos que representam o conteúdo dos objetos informacionais.

Ao lado da indexação manual e da indexação automática, tem-se a indexação semiautomática, que combina a "indexação manual" com a "indexação automática". Sobre a indexação semiautomática, Pinto (2001, p. 227) ressalta que "[...] inicialmente o sistema faz uma indexação automática dos documentos levando em conta as ocorrências das palavras mais frequentes no texto. Em um segundo momento, o indexador humano refina a lista dos descritores propostos".

Corroborando essa concepção de Pinto (2001), Nascimento (2015, p. 41) evidencia que "[...] a indexação semiautomática refere-se àquela que ocorre em sistemas que indexam os documentos, mas que no final do processo os termos são validados pelo ser humano", já a indexação automática é a que é totalmente desenvolvida pelo computador. Lancaster (2004) e Moreiro González (2004) reforçam a ideia de que os *softwares* de indexação automática podem realizar a indexação dos documentos, ressaltando, entretanto, as limitações inerentes a tal processo, as quais deverão ser supridas pela parte humana, ou seja, pelo conhecimento do

[3] "Unidades de informação" é o termo utilizado neste trabalho para se referir a instituições como bibliotecas, arquivos, centros de documentação, bem como instituições do gênero.

bibliotecário indexador[4]. Desse modo, a indexação semiautomática busca equilibrar a eficiência da automatização com a precisão e a contextualização da intervenção humana.

O resultado da atividade de indexação, seja ela automática, manual ou semiautomática, é uma lista de descritores para representar o conteúdo de um determinado documento ou de uma coleção. Esses termos ou descritores serão utilizados não só para sinalizar os documentos em um sistema de recuperação de informação, mas também como linguagem de busca por parte dos usuários no momento dessa realização. A indexação semiautomática é um processo viável para dirimir as limitações e implicações negativas da indexação manual e da indexação automática, considerando, dentre outras questões: a possibilidade de combinar a eficiência da automatização com a precisão da intervenção humana; a viabilidade de realização de ajustes e refinamentos na indexação conforme necessário; a redução do tempo e do custo em comparação com a indexação manual e a melhoria da qualidade dessa operação, a partir de termos relevantes e precisos.

Considerando as limitações ainda presentes na indexação automática, bem como as características próprias da indexação manual, a indexação semiautomática apresenta-se como uma alternativa assertiva e promissora em um cenário onde é crescente a produção de informação, sobretudo, em meio digital. É sobre essa alternativa que se debruça a presente pesquisa, a qual conjuga a automação da indexação com o olhar do indexador humano.

Ao estudar a indexação semiautomática, este trabalho fez uso, em determinado momento, da indexação automática, a qual foi executada por meio do *software* PALAVRAS como instrumento automático para a identificação de sintagmas nominais. A escolha por este *software* fundamentou-se em outros estudos que utilizaram o referido software, bem como pesquisas que avaliaram o desempenho do PALAVRAS, como Silva e Corrêa (205). O referido software foi utilizado por Miorelli (2001), Souza (2005), Lopes (2012), Nascimento (2015), entre outros. Associada ao processamento automático está a parte manual da indexação, refletindo um olhar singular desempenhado pelo indexador no refinamento dos sintagmas candidatos a descritores documentais.

Explorar as possibilidades da indexação semiautomática baseada em unidades significativas, a exemplo dos SN para a construção de vocabulários de termos autorizados representativos de domínios específicos, se mostra como uma alternativa possível e promissora, uma vez que é necessário pensar em possibilidades teóricas, metodológicas e

[4] Neste trabalho sempre que se utilizar a expressão “indexador”, refere-se ao bibliotecário indexador, tendo em vista que outros profissionais, por exemplo, um especialista de domínio também realiza a indexação.

tecnológicas que deem conta do volume de informações que vêm sendo produzido, sem perder o aspecto semântico dos documentos e mantendo a qualidade nas etapas que compõem o processo de indexação, mormente, a parte humana, que é fundamental. Para tanto, surge a seguinte questão norteadora: como realizar a indexação semiautomática por meio da extração de Sintagmas Nominais de artigos científicos?

O contato com as pesquisas científicas, em âmbito nacional e internacional, acerca das possibilidades da indexação manual, automática e semiautomática, bem como o entendimento do papel singular dos sintagmas nominais, enquanto unidades portadoras de informação no âmbito da representação e recuperação da informação, nos levou a estabelecer o seguinte **objetivo geral**: propor diretrizes para a indexação semiautomática de produções científicas por meio de sintagmas nominais.

Ante o objetivo geral, os **objetivos específicos** consistem em:

a) Identificar, na literatura científica nacional e internacional, estratégias de indexação semiautomática;

b) Investigar a composição dos Sintagmas Nominais, enquanto unidades portadoras de informação;

c) Analisar o funcionamento do *software* PALAVRAS na identificação/extração de sintagmas nominais por meio da análise de artigos científicos da área de Organização e Representação do Conhecimento; e

d) Com base em diretrizes, traçar procedimentos para a identificação semiautomática de SN que funcionem como elementos descritores do domínio da Organização e Representação do Conhecimento.

A exploração da indexação semiautomática nos permite fazer uso dos benefícios das Tecnologias de Informação e Comunicação (TICs), sem com isso eximir o papel ímpar do indexador humano na construção de vocabulários representativos de áreas do conhecimento, sobretudo, quando se faz uso de unidades portadoras de informação, como é o caso dos Sintagmas Nominais. A indexação semiautomática, baseada em SN de artigos científicos da área de Organização e Representação do Conhecimento, pode contribuir para recuperação da informação de forma mais ágil nos diferentes domínios, apoiando-se nos benefícios da indexação automática, sem perder o olhar humano. Isto posto, a hipótese desta pesquisa é de que a indexação semiautomática, baseada em Sintagmas Nominais de artigos científicos, possibilita a identificação de termos representativos da área de Organização e Representação do Conhecimento. Assim, acredita-se que esta proposta, ajustada conforme o domínio representado, pode ser aplicada a outros domínios, evidenciando, desse modo, a viabilidade

da indexação semiautomática e do uso de SN para a construção de listas de termos autorizados dos documentos.

O contato com os autores Souza (2005), Maia (2008), Corrêa *et al*. (2011), Lopes (2012), Lapa (2014), Silva (2014), Souza e Raghavan (2006, 2014), e Martins (2014), entre outros, proporcionou subsídios para uma reflexão mais aprofundada acerca da indexação automática durante o curso de mestrado em Ciência da Informação. No doutorado, portanto, foi dado continuidade a esses estudos, refletindo, agora, sobre metodologias e estratégias de representação semiautomática de documentos que possam dar conta do volume de informações digitais que vem sendo produzido atualmente.

O interesse pessoal, enquanto pesquisador, pela prática da indexação, somado ao contato com as pesquisas de Kuramoto (1995, 1999), Corrêa *et al*. (2011), Lapa (2014), Maia (2008), Souza (2005), Nascimento e Corrêa (2019) etc. motivou o desenvolvimento deste estudo, o qual conjuga o entendimento mais aprofundado acerca das possibilidades oferecidas pela indexação semiautomática para a representação temática da informação e para a construção de listas de termos autorizados para esta área.

O estudo se mostra pertinente, uma vez que busca refletir sobre a indexação semiautomática de artigos científicos da área de Organização e Representação do Conhecimento apresentados no Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB 2022 – Grupo de Trabalho - GT 2 - Organização e Representação do Conhecimento. A escolha pelos trabalhos apresentados no referido GT se deu pela ementa do referido grupo de trabalho, que se relaciona diretamente com o interesse de estudo desta pesquisa. No tocante à escolha dos trabalhos dentro do referido GT, definiu-se a edição 2022 por conter as publicações mais recentes até o momento de execução desta pesquisa.

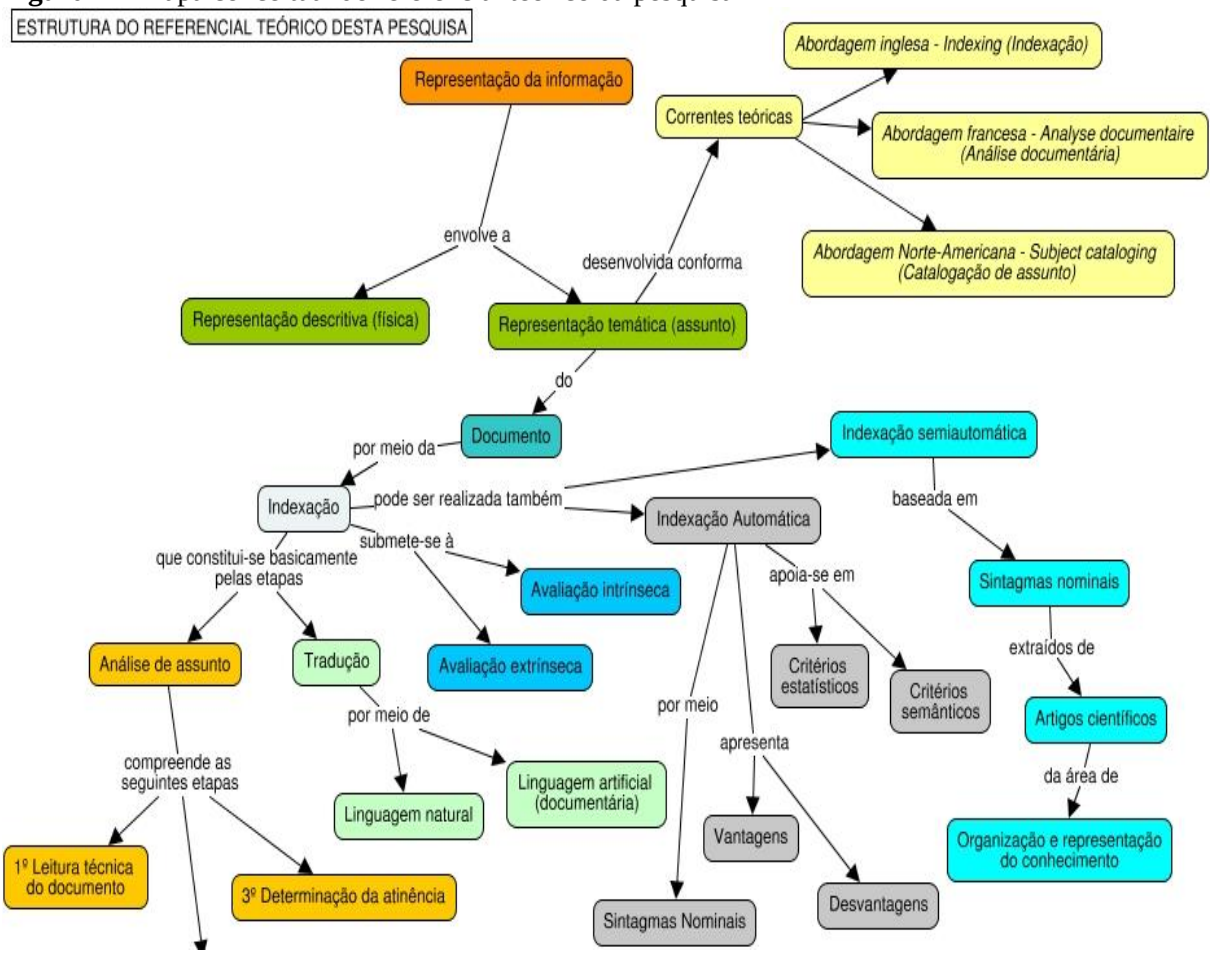
O caráter singular deste trabalho fica evidente, uma vez que é possível encontrar, na literatura nacional e internacional, pesquisas de indexação totalmente automáticas baseadas em palavras isoladas, bem como propostas automáticas baseadas em sintagmas nominais, mas não foram identificadas propostas de indexação semiautomática baseada em SN. Logo, o presente estudo trará contribuições, no plano teórico e prático, às atividades de representação temática da informação de forma mais ágil sem, com isso, perder a semântica dos documentos, uma vez que faz uso de SN como recursos para representação e recuperação documental.

Outrossim, acredita-se que a indexação semiautomática, por meio da extração de sintagmas nominais, se mostra relevante em diversas áreas e aplicações, com vistas à melhoria da precisão e eficiência na indexação e recuperação da informação.

Também, esta pesquisa mostra-se adequada aos temas e pesquisas estudados na Linha de Pesquisa “Organização, Acesso e Uso da Informação”, que contempla estudos no contexto das Teorias, metodologias e tecnologias voltadas à produção, à representação, à organização, apropriação, à democratização, aos usos e aos impactos da informação, do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba (UFPB). No âmbito da referida linha de pesquisa, este estudo reside na “Organização e Representação da Informação e do Conhecimento”.

Como forma de melhor visualizar os temas expostos na seção seguinte (referencial teórico), elaborou-se um mapa (Figura 1) com conceitos e temas necessários para o entendimento da proposta desta pesquisa.

Figura 1 – Mapa conceitual do referencial teórico da pesquisa



Fonte: Desenvolvido pelo autor, 2024.

Consoante à estruturação apresentada, a tese apresenta as seguintes seções: a primeira seção com esta introdução, que traz a apresentação geral da proposta de pesquisa, da problemática, dos objetivos, da hipótese, bem como a justificativa.

Na sequência, tem-se o referencial teórico, apresentado pelas segunda e terceira seções com suas respectivas subseções. A segunda seção traz reflexões acerca da área de Organização e Representação do Conhecimento, apontando conceitos, características e estratégias da prática da indexação manual, seus princípios e formas de avaliação. Ao lado dessas discussões, propõe-se uma reflexão acerca da indexação automática e semiautomática. Ainda no referencial teórico, reflete-se acerca dos Sintagmas Nominais como unidades portadoras de informação e suas potencialidades nas propostas de indexação automática. Ademais, faz-se uma discussão detalhada sobre as características e formas de constituição dos Sintagmas Nominais.

Em sequência, segue-se a terceira seção, dedicada à exposição do percurso metodológico a ser realizado, evidenciando a caracterização da pesquisa, uma síntese das referidas etapas, bem como o detalhamento das etapas da pesquisa empírica. A quarta seção analisa e discute os resultados experimentais. Depois disso, a quinta seção propõe as diretrizes à indexação semiautomática baseada em sintagmas nominais e, por fim, encontram-se, na sexta seção, as considerações finais.

2 ORGANIZAÇÃO E REPRESENTAÇÃO DA INFORMAÇÃO E DO CONHECIMENTO

Os seres humanos, desde os seus primórdios, desenvolvem estratégias de comunicação, das formas mais rudimentares às mais desenvolvidas, proporcionadas pelos avanços tecnológicos. Essas comunicações são mediadas por diferentes linguagens, as quais já faziam parte da realidade das populações pré-históricas, como o uso de representações simbólicas nas paredes das cavernas.

Nesse sentido, pode-se dizer que a humanidade, ao longo do tempo, tem se preocupado com os registros de seus conhecimentos, evidenciando, assim, a necessidade de organização desses registros, sobretudo em razão da exacerbada produção de conhecimento nos mais variados suportes informacionais. Isto posto, a organização e a representação da informação, materializada em distintos registros, tornaram-se essenciais às atividades sociais, às instituições de pesquisa, às instituições de disseminação da informação, bem como às várias áreas da sociedade.

Vinculadas diretamente aos processos comunicacionais, estão as atividades de organização, representação e recuperação da informação. Caixeta e Souza (2008, p. 35), em artigo sobre a representação do conhecimento, ressaltam que “o fenômeno da representação é tão antigo quanto qualquer forma de civilização”.

A organização, a guarda e a recuperação de informações já eram uma preocupação de distintas disciplinas que já existiam bem antes do surgimento da CI, entretanto as disciplinas que estavam voltadas para essas atividades se resumiam às técnicas e aos métodos de guarda e organização, enfatizando o armazenamento em detrimento da disponibilização do que estava preservado. Todavia é com a CI que a preocupação vai mais além das técnicas, uma vez que essa área de estudo se constitui como um conjunto de teorias e práticas que, consoante Saracevic (1996), é um campo dedicado a questões científicas e à prática profissional, voltadas aos problemas da eficaz comunicação do conhecimento e dos registros de conhecimentos entre os indivíduos, no contexto social, institucional ou individual do uso da informação e das necessidades informacionais.

Com o desenvolvimento da sociedade, os avanços tecnológicos e a produção vertiginosa de informações, as instituições, organizações e toda a sociedade se deparam com uma grande massa de registros informacionais, que, muitas vezes, não são relevantes para os indivíduos ou não se encontram recuperáveis. Nesse contexto de desenvolvimento científico e tecnológico e de crescente produção de informações, Guimarães (2006, p. 247) ressalta que:

> [...] o advento dos meios eletrônicos, que por sua vez, possibilitou uma efetiva explosão informacional e o desenvolvimento das tecnologias da informação, especialmente da internet, fez com que se "diminuíssem as distâncias" no mundo inteiro, trazendo uma maior preocupação com questões de organização e representação do conhecimento.

Ao se falar de representação temática da informação, é comum encontrar, na literatura da área, as expressões: organização da informação (OI), organização do conhecimento (OC), representação da informação (RI), representação do conhecimento (RC), contudo é pertinente que não os tome como sinônimos, uma vez que representam conceitos distintos. A esse respeito, Brascher e Café (2008, p. 2) ressaltam que há diferença entre os termos e que ela se sustenta "no pressuposto de que a informação e conhecimento são conceitos distintos e, portanto, OI e OC e RI e RC também o são". Essas autoras discutem as diferenças apontadas por alguns autores a fim de caracterizar os termos "informação" e "conhecimento" como conceitos distintos, possibilitando, assim, a distinção clara entre os referidos termos dentro do campo da CI.

Para as autoras citadas:

> A organização da informação é, portanto, um processo que envolve a descrição física e de conteúdo dos objetos informacionais. O produto desse processo descritivo é a **representação da informação**, entendida como um conjunto de elementos descritivos que representam os atributos de um objeto informacional específico. Alguns tipos de representação da informação são construídos por meio de linguagens elaboradas especificamente para os objetivos da OI (Brascher; Café, 2008, p. 5).

Uma reflexão das autoras que nos permite compreender de forma mais clara a distinção entre os referidos conceitos é a de que:

> No contexto da OI e da RI, temos como objeto os registros de informação. Estamos, portanto, no mundo dos objetos físicos, distintos do mundo da cognição, ou das ideias, cuja unidade elementar é o conceito. [...] Quando nos referimos à OC e à RC, estamos no mundo dos conceitos e não naquele dos registros de informação (Bracher; Café, 2008, p. 5-6).

Com base nas autoras, percebe-se que a OI está direcionada a objetos informacionais e a OC direcionada a unidades do pensamento (conceitos), sendo a OI também aplicada à organização de um conjunto de objetos informacionais, a exemplo de bibliotecas, arquivos etc. Por conseguinte, esses dois processos distintos envolvem formas distintas de representação, justificando, também, a interpretação de que a RI envolve atividades que descrevem física e semanticamente os objetos informacionais, enquanto a RC constitui uma

estrutura conceitual que representa modelos de mundo, permitindo-nos compreender os fenômenos nele presentes.

No contexto da Ciência da Informação, a organização do conhecimento lida com a estruturação e organização dos conceitos de domínios específicos, construindo, assim, representações conceituais no mundo das ideias. Silva (2017, p. 18), acerca da representação do conhecimento, ressalta que:

> A Organização do Conhecimento (OC) é a área que se ocupa, primordialmente, das representações do conhecimento; abordada em diversos campos do saber, como a Filosofia, Linguística, Computação, Lógica, Biologia, Sociologia, entre outros, mas que possui forte desenvolvimento na Biblioteconomia e na Ciência da Informação (CI).

A Organização do Conhecimento e a Representação do Conhecimento apoiam a organização da informação em diferentes ambientes informacionais. Tome-se, como exemplo, um tesauro, que é produto da Organização do Conhecimento e subsidia as atividades de indexação na Organização e Representação da Informação.

Quadro 01 – Organização da Informação e Organização do Conhecimento

	Organização da Informação (OI)	Organização do Conhecimento (OC)
Aplicação	Mundo dos objetos físicos	Mundo dos conceitos
Finalidade	Descrição física de objetos informacionais	Análise do conceito
Alguns resultados	Resumos Índices Fichas catalográficas	Tesauros Taxonomias Ontologias

Fonte: Vignoli, Souto e Cervantes (2013, p. 61) com base em Brascher e Café (2008).

Vignoli, Souto e Cervantes (2013, p. 61) ressaltam que "relacionados com a aplicação da OC e da OI, existem os Sistemas de Organização do Conhecimento (SOC), que visam à organização da produção intelectual humana". O tesauro é um exemplo de SOC que subsidia a indexação manual, semiautomática ou automática no contexto da Organização e Representação da Informação.

Para além das discussões do escopo de atuação de cada área de estudo, no contexto específico desta tese, e considerando que a proposta desta pesquisa se volta à descrição de objetos informacionais, ou seja, à indexação semiautomática de documentos digitais, parece lógico que se utilize a partir daqui as expressões "Organização da Informação" e "Representação da Informação" para se referir ao processo em estudo. Isto também porque se

estar lidando com a descrição de conteúdos de “Objetos Informacionais”, recursos informacionais em meio digital, por meio da atividade de indexação.

As atividades de representação da informação dentro de uma Unidade de Informação (biblioteca, centro de documentação, arquivo etc.) proporcionam a comunicação entre usuários (indivíduos que necessitam de informação) e sistema (coleção de documentos). Não basta que as instituições forneçam informação, faz-se necessário que esta esteja minimamente organizada para que assim seja encontrada pelo indivíduo que dela necessita e busca. Dentro desse contexto, algumas atividades, como a indexação e a classificação, são desempenhadas pela subárea Tratamento da informação dentro do campo maior e interdisciplinar que é a Ciência da Informação.

A organização da informação é fundamental no processo de recuperação informacional, principalmente para facilitar o acesso aos documentos, estejam eles em formato analógico ou digital. Isto porque é evidente que hodiernamente as informações em formato digital vêm crescendo de forma demasiada de modo mais avançado do que as em formato impresso. A preocupação não é mais apenas “ter a informação”, senão “achar a informação”. As atividades de organização, guarda e recuperação da informação, juntas, permitem aos usuários que as buscam, acharem de forma mais ágil.

O Tratamento da Informação no contexto da CI, aplicado a unidades de informação e a sistemas de recuperação de informação, envolve atividades agrupadas em dois grandes eixos, as atividades voltadas para a representação descritiva (também denominada de tratamento descritivo) dos recursos informacionais (documentos) e as atividades voltadas para a representação temática (também denominada de tratamento temático) dos recursos informacionais.

Cesarino e Pinto (1978, p. 269) ressaltam que os documentos, no contexto dos SRIs, podem ser analisados de duas maneiras:

> a) bibliograficamente ou objetivamente – este tipo de análise pretende a descrição do documento através de suas características físicas, com o objetivo de dar resposta à questão: <<Qual a aparência física deste documento?>> b) intelectualmente ou subjetivamente – este tipo de análise pretende a descrição do documento em termos de suas características de conteúdo, com o objetivo de dar resposta à questão: <<Sobre o que é este documento?>>.

Os dois aspectos de análise dos documentos dentro do Tratamento/Representação da Informação, expostos por Cesarino e Pinto (1978), “Bibliograficamente ou objetivamente” e

“intelectualmente ou subjetivamente” envolvem, dentro das UIs, as atividades de representação/descrição dos aspectos físicos dos documentos, por outro lado, os aspectos intelectuais/subjetivos envolvem as atividades que busquem a representação dos temas, conteúdos dos documentos (Catalogação por assunto, indexação, classificação de documentos).

Sobre essa temática, Dias e Naves (2007, p. 17) compreendem o “Tratamento da informação” como expressão que engloba todas as disciplinas, técnicas, métodos e processos relativos à “[...] descrição física e temática dos documentos numa biblioteca ou sistema de recuperação da informação”; ao “[...] desenvolvimento de instrumentos (códigos, linguagens, normas, padrões) a serem utilizados nessas descrições”; e à “[...] concepção/implantação de estruturas físicas ou bases de dados destinadas ao armazenamento dos documentos e de seus simulacros (fichas, registros eletrônicos, etc.)”. Compreende as disciplinas de classificação, catalogação, indexação, bem como as especialidades que delas derivam, ou terminologias novas nelas aplicadas, tais como metadados, e ontologias, entre outras.

Fujita, Rubi e Boccato (2009, p. 22), acerca das atividades que constituem o Tratamento da Informação, ressaltam que:

> O tratamento descritivo refere-se propriamente à catalogação, ou seja, à representação descritiva da forma física do documento (autor, título, edição, casa publicadora, data, número de páginas etc.). O tratamento temático, em bibliotecas, diz respeito ao assunto tratado no documento, ou seja, compreende a análise documentária como área teórica e metodológica que abrange as atividades de classificação, elaboração de resumos, indexação e catalogação de assunto, considerando as diferentes finalidades de recuperação da informação.

A esses dois eixos de atividades de tratamento da informação, Ruiz Perez (1992) utiliza a nomenclatura: “análise documental de forma” e “análise documental de conteúdo”. Adentrando especificamente nas atividades voltadas para o *tratamento temático* dos documentos encontram-se as seguintes atividades: a indexação, a catalogação por assunto, a classificação e a confecção de resumos (Dias e Naves, 2007; Guimarães, 2009; Redigolo, 2010).

O Tratamento Temático da Informação ocupa lugar ímpar na CI, uma vez que proporciona a mediação entre a informação produzida e os usuários que dela necessitam. Essa afirmação está de acordo com a posição de Guimarães (2008, p. 78), ao observar que o:

> [...] Tratamento Temático da Informação (TTI) nela ocupa (como se pode observar tanto na literatura quanto nas distintas práticas profissionais) um

> espaço nuclear, visto revelar a mediação entre a produção e o uso da informação, entre elas tecendo a mais sólida ponte: a que dá acesso ao conteúdo informacional.

No contexto da área de Organização do Conhecimento e Ciência da Informação, Barité (1998, p. 124) esclarece que o Tratamento Temático da Informação - TTI se volta para a "[...] análise, descrição e representação do conteúdo dos documentos, bem como suas inevitáveis interfaces com as teorias e sistemas de armazenamento e recuperação da informação".

Ainda acerca da área de Representação Temática da Informação, na CI, Dantas, Sampaio e Albuquerque (2020, p. 75) ressaltam que:

> A representação temática é uma área que envolve a organização de assuntos, independentemente dos suportes de registro, pois visa o tratamento dos materiais, a partir de técnicas específicas, de forma a identificar os descritores representativos do tema em pauta, para que seja possível a recuperação precisa, de interesse de um determinado grupo.

Partindo do propósito maior desta tese, adentramo-nos nas atividades específicas de Representação Temática da Informação nas Unidades de Informações e nos Sistemas de Recuperação de Informação, ressaltando que são encontrados, na literatura da área, autores utilizando as expressões "tratamento temático da informação" e "representação temática da informação". Essa variedade terminológica pode ser justificada pelas correntes teóricas sob as quais as atividades, voltadas para a indicação de conteúdos dos documentos, foram se desenvolvendo. Além da catalogação por assunto, classificação e indexação, *a* elaboração de resumos também constitui uma atividade de representação temática da informação.

Estas atividades apresentam uma semelhança: objetivam proporcionar a recuperação da informação, todavia elas diferem no tocante ao enfoque que é dado a cada atividade, bem como aos resultados alcançados por cada uma, consoante ver-se-á a seguir. Há uma variação terminológica que envolve as atividades de indexação e catalogação de assunto, conforme já ressaltada por Silva e Fujita (2004).

A esse respeito, Rubi (2008, p. 39) expõe, como está sintetizado no Quadro 2, os elementos que diferenciam a indexação da catalogação de assunto:

Quadro 2 - Elementos diferenciadores entre indexação e catalogação de assunto

CARACTERÍSTICAS	**INDEXAÇÃO**	**CATALOGAÇÃO DE ASSUNTO**
EQUIVALENTE EM INGLÊS	Indexing	Subject cataloguing
ORIGEM	Inglesa	Norte-americana
DEFINIÇÃO	"A ação de descrever e identificar um documento de acordo com seu assunto." (UNISIST, 1981, p. 84)	"A disciplina ou conjunto de disciplinas que tratam da representação, nos catálogos de bibliotecas, dos assuntos contidos no acervo." (Fiúza, 1985, p. 257)
AMBIÊNCIA	Serviços de indexação e resumos Sistemas de informação especializado	Bibliotecas
PROFISSIONAL RESPONSÁVEL/FORMAÇÃO INICIAL	Indexador/Especialista no assunto	Catalogador/Bibliotecário
PROCEDIMENTOS: ETAPAS	1. Exame do documento e estabelecimento do assunto de seu conteúdo; 2. Identificação dos conceitos presentes no assunto; 3. Tradução desses conceitos nos termos de uma linguagem de indexação (Associação Brasileira de Normas Técnicas, 1992)	1. Análise individual dos documentos para determinação dos conceitos expressos; 2. Representação dos conceitos em linguagem do sistema; 3. Determinação das entradas para os conceitos identificados; 4. Relacionar termos correlatos aos grupos de documentos familiares que poderão ser recuperados. (Connell, 1996).
LINGUAGENS DOCUMENTÁRIAS	Tesauros	Lista de cabeçalhos de assunto
O ASSUNTO É DEFINIDO POR	Termos/Descritores	Cabeçalhos de assunto
PRODUTO FINAL	Índices	Catálogos
REVOCAÇÃO DA RECUPERAÇÃO	Baixa	Alta
PRECISÃO NA RECUPERAÇÃO	Alta	Baixa

Fonte: (Rubi, 2008, p. 39).

Guimarães (2009) ressalta que as atividades de representação dos conteúdos dos documentos constituíram-se ao longo do tempo por meio de três correntes teóricas, a saber: Abordagem Norte-Americana - *subject cataloging* (Catalogação de Assunto); Abordagem Inglesa- *Indexing* (Indexação); e Abordagem Francesa -*Analyse Documentaire* (Análise Documentária). Esse autor, sobre a primeira corrente teórica (**Catalogação de Assunto – Abordagem Norte-Americana**), diz que:

> [...] observa-se uma primeira abordagem a partir da ótica do subject cataloguing, voltada diretamente para a atividade profissional em bibliotecas e sob forte influência da Escola de Chicago. Essa concepção decorreu diretamente dos princípios de catalogação alfabética de Cutter e da tradição de cabeçalhos de assunto da Library of Congress, cuja ênfase reside no catálogo enquanto produto do tratamento da informação em bibliotecas (Guimarães, 2008, p. 82).

Conforme exposto, essa abordagem desenvolve-se especificamente no âmbito das bibliotecas, com interesse na construção de produtos, no caso, os catálogos presentes nas bibliotecas como instrumento de representação temática e de recuperação da informação.

A segunda corrente teórica de desenvolvimento da área de Tratamento Temático da Informação (**Indexação – Abordagem Inglesa**) desenvolve-se a partir da percepção da **indexação**, essa concepção abrange as bibliotecas, centros de documentação, bem como o universo editorial, por meio da qual os índices, como produtos do TTI, são resultados da utilização de linguagens de indexação, como os tesauros. Aqui há uma preocupação mais teórica no que se refere à construção dessas linguagens. Alguns trabalhos que se destacaram nessa segunda corrente de desenvolvimento do TTI foram os de Foskett, Austin, Farradane, Metcalfe, Aitchinson, Gilchrist e Lancaster (Guimarães, 2008).

A terceira corrente de desenvolvimento da área de TTI (**Análise Documentária – Abordagem Francesa**), diferentemente das correntes anteriores, volta-se para a compreensão e o desenvolvimento dos referenciais teóricos para os processos de TTI, enquanto as correntes anteriores estavam preocupadas com produtos, aqui a preocupação é com o processo em si que resulta em produtos. Nessa concepção, segundo Guimarães (2008, p. 83), "[...] o foco centra-se no próprio processo de TTI, vale dizer, na explicitação dos procedimentos voltados para a identificação e seleção de conceitos para posterior representação e geração de produtos".

Na literatura da área de representação temática da informação, no âmbito da CI, encontram-se expressões, a exemplo de "indexação" e "análise documentária", ora como

sinônimas, ora como processos distintos e com abrangência distinta. Segundo Silva e Fujita (2004, p. 137), "A existência de diferentes correntes teóricas explica o uso de termos como análise de assuntos, análise de conteúdos documentários e análise documentária". Essa variedade pode ser explicada pela existência das correntes teóricas mencionadas anteriormente e que embasam essas atividades, conforme ressaltam Silva e Fujita (2004, p. 136):

> Esses estudos, porém, estão claramente inseridos em correntes teóricas e é fácil confundir, na literatura, a função da indexação perante a necessidade de análise de conteúdo. Na literatura observa-se a existência de duas correntes teóricas: a francesa e a inglesa.

Assim, tem-se a expressão "Análise documentária[5]" que é adotada com dois sentidos distintos. Esta expressão foi conceituada por Gardin *et.al*. (1981, p. 29) como "um conjunto de procedimentos efetuados com a finalidade de expressar o conteúdo de documentos científicos, sob formas destinadas a facilitar a recuperação da informação". Nesse contexto, a corrente francesa compreende a "Análise documentária" como um macro universo, no qual a indexação está inserida. Essa posição é ressaltada pelas autoras Silva e Fujita (2004, p. 138), ao dizerem: "[...] compreendemos a análise documentária como área teórica e metodológica com o objetivo de tratamento temático de documento que abrange as atividades de indexação, classificação e elaboração de resumos [...]".

Já os autores espanhóis compreendem que a expressão "Análise documentaria", conforme Silva e Fujita (2004, p. 137),

> [...] comporta dois níveis de divisão: o da forma – análise descritiva ou bibliográfica – refere-se ao tratamento físico da informação ligado com o suporte, e o do conteúdo, que se refere ao tratamento temático da informação e destina-se à representação condensada do assunto intrínseco ou extrínseco tratado em um determinado documento.

Assim, o termo "Análise documentária", na concepção dos autores franceses, abrange exclusivamente o tratamento temático dos conteúdos dos documentos. A esta concepção, autores como Chaumier, Kobashi, Smit, Tálamo, Ginez de Lara, Cintra, Cunha, Guimarães, Fujita, Gil Leiva, Ruiz Perez, Pinto Molina, entre outros, estão de acordo (Silva; Fujita, 2004). Entretanto, na concepção dos teóricos espanhóis, o referido termo abrange dois

[5] É possível encontrar na literatura da área variação para esse termo, ora o uso do termo "análise documentária", ora "análise documental". Esta variação está associada à tradução, consoante ressaltam Guimarães, Nascimento e Moraes (2005).

grandes processos, a representação descritiva (descrição física) e a representação temática (descrição temática) de um determinado documento.

Já a expressão "Indexação" é utilizada pelos teóricos da corrente inglesa, como, Foskett, Lancaster, Campos, Van Slype, Farrow, entre outros, como processo (Silva; Fujita, 2004). Nessa concepção, "análise documentária" e "indexação" compreendem processos idênticos, sendo esse processo constituído pelas etapas de análise de assunto e tradução[6].

No que se refere às correntes teóricas de desenvolvimento do TTI, sempre que necessário, utiliza-se nesta pesquisa a expressão "Indexação", concebida como processo pelos autores da corrente inglesa. Essa escolha se justifica porque a presente pesquisa trata de estratégias de indexação semiautomática, as quais buscam propor índices representativos dos conteúdos documentais para fins de recuperação da informação.

Nesse contexto, a indexação se mostra como uma atividade essencial para garantir o propósito de qualquer instituição de informação que vise a disseminação de informação, uma vez que serve como ponte, permitindo a comunicação entre o usuário e a informação. Torna-se pertinente que se compreenda a distinção básica entre as três maneiras de execução dessa atividade que, consoante Pinto (2001), pode ser realizada das seguintes formas: manual, chamada também de intelectual, que é feita pelo humano, no caso, bibliotecário indexador; mecânica, feita por recursos computacionais, conhecida também por **indexação automática** e uma análise que combina as duas: humana e mecânica.

Esta última é chamada por Pinto (2001, p. 227) de **indexação semiautomática** ou assistida pelo computador e realizada, também de acordo com a autora, da seguinte maneira:

> [...] inicialmente o sistema faz uma indexação automática dos documentos levando em conta as ocorrências das palavras mais frequentes no texto. Em um segundo momento, o indexador humano refina a lista dos descritores propostos pelo sistema fazendo os ajustes e/ou complementações necessárias.

Em revisão de literatura sobre a automatização no processo de indexação, Gil Leiva (1999, p. 57; 2008, p. 320) pontua que foram identificadas várias expressões acerca da associação da indexação com a automatização, as quais evidenciaram a existência de três conceitos:

[6] Embora a indexação de modo geral seja compreendida como sendo constituída por duas etapas, há autores que expõem esse processo em mais etapas que são discutidas em seção específica sobre a Indexação Manual.

> a) aquele que envolve os programas (softwares) que realizam o processo de armazenamento dos termos de indexação obtido por um indexador humano, chamado **de indexação assistida por computado**r;
> b) o segundo entendimento engloba os programas que realizam a análise dos documentos de modo automático com vistas a identificar os possíveis termos descritores de um documento e, se necessário, os referidos termos são validados por um profissional, chamado de **indexação semi-automática** e;
> c) por fim, o terceiro conceito diz respeito aos programas computacionais que realizam o processo de análise dos documentos e não ocorre validação por profissionais, ou seja, a **indexação automática** propriamente dita.

Enquanto Pinto (2001) utiliza os termos "indexação semiautomática" e "indexação auxiliada por computador" como sinônimos, Gil Leiva (1999) diferencia, conforme a citação acima, a indexação auxiliada pelo computador, sendo aquela em que se utiliza o computador para gravar os dados e os termos de indexação; já a indexação semiautomática envolve uma atividade mais complexa pelo computador que é capacitado para realizar análises linguísticas e estatísticas nos textos, selecionando os possíveis termos representativos dos documentos.

Nesta pesquisa, utiliza-se a expressão "Indexação semiautomática" para se referir à indexação que é realizada em dois momentos: pelo computador e pelo bibliotecário indexador. O primeiro momento se caracteriza pela realização da análise do texto e da identificação dos possíveis termos descritores e o segundo momento é realizado pelo homem, no contexto desta pesquisa, o bibliotecário indexador, validando ou não os termos. Assim, é importante deixar clara a distinção que permeia uma e outra maneira de indexação. Segue-se o modelo tradicional, ou seja, a indexação manual[7].

2.1 INDEXAÇÃO MANUAL

Nesta subseção, analisa-se a atividade de indexação de documentos a partir dos seguintes aspectos: definição, o processo e as etapas da indexação, seus princípios e as medidas utilizadas para mensurar esse processo.

É crescente o número de pesquisas acerca da atividade de indexação em bibliotecas, unidades de informação e SRIs, cada uma da sua maneira. Contudo as definições encontradas na literatura especializada comungam com a ideia central desta atividade, a saber: representar/sinalizar o conteúdo dos documentos nas UIs e nos SRIs a fim de permitir, de forma eficiente, a recuperação da informação.

[7] Encontraram-se, na literatura da área, autores que utilizam outros termos sinônimos para se referirem à indexação manual, como Pinto (2000, p. 67), que usa "manual, intelectual, ou humana"; e Maimone, Kobashi, e Mota (2016, p. 80), "operação intelectual humana".

A indexação é uma atividade que pressupõe um processo que, como tal, envolve etapas as quais variam conforme a visão de cada autor: algumas propostas detalham mais esse processo, outros a apresentam de modo mais generalizado. No entanto, todas buscam o mesmo propósito: representar o conteúdo dos documentos para posterior recuperação.

O Sistema Mundial de Informação Científica - UNISIST (1981, p. 84), o qual foi responsável por elaborar um documento contendo "os princípios da indexação", sendo esse documento a primeira tentativa internacional de se normalizar o processo de indexação, expõe que:

> A indexação é vista como a ação de descrever e identificar um documento de acordo com o seu assunto. [...]. Durante a indexação, os conceitos são extraídos do documento através de um processo de análise, e então traduzidos para os termos de instrumentos de indexação (tais como tesauros, listas de cabeçalhos de assunto, esquemas de classificação etc.).

Tálamo (1987) propõe que o processo de indexação consiste na identificação do tema de um documento, ou seja, de seu assunto, por meio de um conjunto de perguntas que o indexador faz ao documento, com vistas a identificação do assunto, a saber: quem? (ser); o quê? (tema); como? (modo); onde? (lugar); e quando? (tempo). Consoante a referida autora, por meio da aplicação dessas questões, é possível identificar o tema e o objetivo do texto presentes no documento. Outros autores como Kobashi (1994), Lasswell (1971), Garcia Gutierre e Lucas (1987) também estudaram propostas de indexação baseadas em perguntas e respostas.

Van Slype (1991) compreende a indexação como a operação que enumera os conceitos sobre os quais trata um documento, representando-os por meio de uma linguagem combinatória, a saber, lista de descritores livres, lista de autoridades e o thesaurus de descritores.

Para a Associação Brasileira de Normas Técnicas (ABNT), órgão responsável pela normalização técnica de documentos científicos no Brasil e que fornece insumos ao desenvolvimento tecnológico brasileiro, a indexação consiste no "Ato de identificar e descrever o conteúdo de um documento com termos representativos dos seus assuntos e que constituem uma linguagem de indexação" (ABNT, NBR 12676, 1992, p. 2).

Para Pinto Molina (1993, p. 208), a indexação "[...] é a técnica de caracterizar o conteúdo de um documento (...) retendo as ideias mais representativas para vinculá-las a termos de indexação adequados".

A indexação se propõe a descrever e determinar os conceitos presentes em um documento, tendo em vista a sua importância para a Unidade de Informação ou SRIs. Esta atividade engloba o conhecimento do assunto do documento, bem como uma definição precisa do nível de informação a ser preservado de forma a responder às necessidades dos usuários (Guinchat; Menou, 1994).

Pinto (2000, p. 65), baseada em Gardin (1974), diz que a "indexação documentária" pode ser compreendida como "um conjunto de atividades que consiste em identificar, nos documentos, os seus 'Traços Descritivos' (TD's) ou macroproposições e em seguida extrair os elementos/descritores (sintagmas) indicadores do seu conteúdo visando a sua recuperação posterior". Essa autora já evidenciava o potencial da indexação baseada em unidades significativas dos documentos, ou seja, os sintagmas nominais (proposta apresentada pelo grupo SYDO), ou frases (proposta de Alain F. Smeaton e Paraic Sheridan), ou ainda os sintagmas verbais (proposta de Geneviéve Lallich e de Virgínia Bentes Pinto).

Lancaster (2004, p. 1, grifo nosso), no início de sua obra, ressalta a diferença existente entre as práticas de: catalogação descritiva, indexação e elaboração de resumos. Para o referido autor, "Os processos de catalogação descritiva identificam autores, títulos, fontes, e outros elementos bibliográficos; **os processos de indexação identificam o assunto de que trata o documento**; e o resumo serve para sintetizar o conteúdo do item".

Segundo Robredo (2005, p. 165), a indexação "[...] consiste em indicar o conteúdo temático de uma unidade de informação, mediante a atribuição de um ou mais termos (ou códigos) ao documento, de forma a caracterizá-lo de forma unívoca".

Para Dias e Naves (2007), a indexação possui dois sentidos, o primeiro, em sentido mais amplo, se refere à atividade de criar índices (de autor, de título, de assunto etc.); o segundo, de forma mais restrita, refere-se exclusivamente à indexação de conteúdo dos documentos.

Para Gonçalves e Souza (2008, p. 4), a indexação é:

> [...] o processo de atribuir - enumerar conceitos que sintetizem e representem o conteúdo de um documento, de acordo com as políticas e especialidade (domínio do assunto) a representar em um sistema de recuperação da informação.

No que se refere à definição desta atividade, Araújo e Oliveira (2011) defendem que "Indexação" consiste na descrição dos conteúdos dos documentos e possui como maior propósito a recuperação da informação necessitada por parte do usuário. Essas autoras

observam ainda que a indexação "é uma das principais atividades desenvolvidas numa Biblioteca ou Unidade de Informação" (Araújo; Oliveira, 2011, p. 41).

Para Gil Leiva e Fujita (2012, p. 31), em uma perspectiva cognitiva, "a indexação gera palavras-chave, índices ou os cabeçalhos de assunto de um documento. Para obtê-los, previamente é desencadeada uma sucessão interativa e simultânea de processos mentais que têm a ver com a percepção da informação, da memória e da compreensão". Troitiño Rodriguez *et al*. (2016) consideram a indexação como um processo de Organização do Conhecimento que possui uma metodologia de representação da informação, garantindo, assim, o acesso aos documentos e a recuperação destes.

Pret e Cordeiro (2019, p. 167) apontam que a indexação, como processo de análise conceitual e representação da informação do documento que possibilita a sua recuperação futura, "tem sido incorporada pelos SOC a fim de sistematizar conceitos de um determinado campo do saber, sistematização que permitiria a recuperação dos documentos, atendendo as necessidades de uso de suas comunidades discursivas.".

Considera-se, a partir das definições expostas nesta discussão, que o processo de indexação realizado em unidades de informação (bibliotecas tradicionais, digitais, híbridas, centros de documentação, arquivos etc.) envolve a representação dos conteúdos de itens informacionais (livros, recursos digitais, imagens, fotografias etc.) para fins de recuperação, funcionando como uma ponte entre o usuário e o documento em distintos contextos informacionais.

Compreender a atividade de indexação como um processo pressupõe a existência de etapas dentro dessa prática. Embora os autores da área explicitem o processo de indexação constituído por duas três etapas, por exemplo, em alguns casos, desdobramentos de propostas mais gerais, a essência das propostas dos autores continua a mesma, a saber: a representação do conteúdo dos documentos para fins de recuperação. Essa quantidade de etapas que compõe o processo de indexação varia consoante cada autor, segundo pode ser visto no Quadro 3:

Quadro 3 – Etapas da Indexação manual

AUTORES	ETAPAS/ESTÁGIOS	
UNISIST (1976)	***1ª Etapa/estágio***	Análise - estabelecimento dos conceitos tratados num documento, isto é, o assunto;
	2ª Etapa/estágio	Tradução dos conceitos nos termos da linguagem de indexação.
Chaumier (1988, p. 64-65)	***1ª Etapa/estágio***	Conhecimento do conteúdo do documento. Faz-se através de uma leitura rápida, ou "leitura em diagonal" do documento;
	2ª Etapa/estágio	Escolha dos conceitos. Exige uma verdadeira análise conceitual do documento. Isso pode resultar da análise realizada para condensação se o documento puder ser resumido;
	3ª Etapa/estágio	Tradução dos conceitos escolhidos. Essa tradução se faz, na fase que segue, nos termos da linguagem documentária utilizada pelo serviço de

		documentação.
Van Slype (1991 apud RUBI, 2008, p. 26)	***1ª Etapa/estágio***	Conhecimento do conteúdo do documento;
	2ª Etapa/estágio	Escolha dos conceitos a serem representados, baseando-se na aplicação da regra da seletividade e exaustividade;
	3ª Etapa/estágio	Tradução dos conceitos selecionados da forma em que aparecem impressos no documento, para os descritores do tesauro aplicando a regra da especificidade.
	4ª Etapa/estágio	Incorporação dos elementos sintáticos.
Associação Brasileira de Normas Técnicas NBR 12676 (1992)	***1ª Etapa/estágio***	Exame do documento e estabelecimento do assunto de seu documento;
	2ª Etapa/estágio	Identificação dos conceitos presentes no assunto;
	3ª Etapa/estágio	Tradução desses conceitos nos termos de uma linguagem de indexação.
Chu & O'Brien (1993 apud DIAS; NAVES, 2007, p. 28)	***1ª Etapa/estágio***	Análise de Assunto do texto;
	2ª Etapa/estágio	Expressão do conteúdo do assunto nas palavras dos indexadores (Linguagem Natural);
	3ª Etapa/estágio	Tradução para um vocabulário de indexação;
	4ª Etapa/estágio	Expressão do assunto em termos do índice.
Pinto (2000)	***1ª Etapa/estágio***	Análise conceitual;
	2ª Etapa/estágio	Tradução;
	3ª Etapa/estágio	Controle de qualidade.
Fujita (2003, p. 63)	***1ª Etapa/estágio***	Analítico: em que é realizada a compreensão do texto como um todo, a identificação e a seleção de conceitos válidos para a indexação;
	2ª Etapa/estágio	Estágio de tradução, que consiste na representação de conceitos por termos de uma linguagem de indexação;
Lancaster (2004)	***1ª Etapa/estágio***	Análise conceitual;
	2ª Etapa/estágio	Tradução.
Robredo (2005 apud RUBI, 2008, p. 27)	***1ª Etapa/estágio***	Análise conceitual do conteúdo do documento envolvendo sua compreensão; a identificação dos conceitos que representem o conteúdo e a seleção dos conceitos tendo em vista a recuperação da informação;
	2ª Etapa/estágio	Expressão dessa análise por meio de códigos, palavras ou frases representativas do assunto;
	3ª Etapa/estágio	Tradução das descrições dos assuntos para a linguagem de indexação;
	4ª Etapa/estágio	Organização das descrições de acordo com a sintaxe da linguagem de indexação.
Naves e Kuramoto (2006, p. 104)	***1ª Etapa/estágio***	Análise do documento e estabelecimento do seu assunto;
	2ª Etapa/estágio	Identificação dos principais conceitos do documento;
	3ª Etapa/estágio	Tradução dos conceitos identificados em termos de uma linguagem de indexação.

Fonte: Adaptado de UNISIST (1976), Lancaster (2004), Fujita (2003), Associação Brasileira de Normas Técnicas NBR 12676 (1992), Chaumier (1988), Naves e Kuramoto (2006), Pinto (2000), Van Slype (1991), Chu & O'Brien (1993) e Robredo (2005).

Embora os autores visualizem a indexação, no que tange à quantidade de etapas, de forma variada, o propósito maior permanece o mesmo: a representação dos conteúdos dos documentos para que aquela funcione como ponte entre estes e os usuários, ou seja, representar para recuperar. Considera-se nesta pesquisa a indexação na visão de Lancaster (2004), o qual compreende que tal operação é realizada por meio de duas etapas: análise de assunto (ou análise conceitual) e tradução.

A **análise de assunto**, por ser uma atividade eminentemente desempenhada pelo ser humano, é uma etapa subjetiva. Essa subjetividade evidencia-se pelos próprios elementos

envolvidos em tal processo, os quais, conforme ressalta Fujita (2003), são: *leitor*, *texto* e *contexto*, uma vez que é desenvolvida pelo indexador humano. As autoras Strehl (1998) e Pinheiro (1978) reforçam o caráter subjetivo da indexação, quando, esta última esclarece que essa atividade envolve o "julgamento", o qual possui concordâncias e discrepâncias. Esta etapa da indexação, que objetiva a compreensão do conteúdo de um documento e a identificação e seleção de seus conceitos representativos, pode ser realizada sob diferentes pontos de vista, segundo Albrechtsen (1993), a saber: concepção simplista, concepção orientada para o conteúdo e concepção orientada pela demanda.

A concepção simplista tem como foco a análise objetiva do documento, por meio das entidades objetivas absolutas, ou seja, com base nas entidades linguísticas isoladas do documento, e com base na soma dessas entidades. Aqui a preocupação é com o que as entidades linguísticas mostram explicitamente. Segundo Albrechtsen (1993), essa concepção pode facilmente ser automatizada, uma vez que aqui a análise é feita na superficialidade, ou seja, no que as palavras e grupos de palavras mostram aos usuários.

Já a concepção orientada para o conteúdo vai além do que está explícito, do que está posto a princípio, buscando compreender o que está nas entrelinhas, no subjetivo. Aqui a identificação do conteúdo busca os conceitos que estão além da estrutura superficial do texto, evidenciando o implícito, o conteúdo latente.

Por fim, a concepção orientada pela demanda compreende o assunto como o recurso necessário para acesso ao conhecimento registrado. Assim, os assuntos são instrumentos que permitem a construção de conhecimento por parte dos usuários que necessitam de informações. Como a orientação é direcionada para o usuário, o indexador se coloca no lugar do usuário e se questiona como poderia tornar o conteúdo de um documento visível a um usuário, ou seja, o indexador busca identificar o conteúdo que possivelmente satisfaça as necessidades dos usuários, bem como os termos que o represente.

Fujita (2003) ressalta que cada concepção apresenta vantagens e desvantagens. A concepção simplista resulta em uma indexação automática e, com isso, a diminuição de custos para essa atividade, uma vez que dispensa o indexador. Contudo essa concepção apresenta a limitação de identificar os conteúdos de forma superficial, não evidenciando a transferência de conhecimento. A concepção orientada para o conteúdo volta-se exclusivamente para a representação dos documentos, sem se preocupar com as possibilidades de usos desses documentos, ou seja, sem considerar o propósito maior que é a satisfação da necessidade informacional do usuário.

Já a concepção orientada pela demanda apresenta como ponto positivo que é a possibilidade de permitir o acesso à disseminação da informação, uma vez que a indexação é baseada em perguntas que possam satisfazer as necessidades informacionais dos usuários de um determinado sistema. Porém, é preciso enfatizar que essa concepção exige mais responsabilidade e comprometimento por parte do indexador na priorização de assuntos relevantes a usuários potenciais. Fujita (2003) conclui que a concepção orientada para o conteúdo, bem como a orientada pela demanda são intrínsecas, uma vez que a primeira deve orientar a identificação dos conceitos dentro do processo de leitura documentária, e a segunda orienta a seleção de conceitos.

Fidel (1994) emprega a expressão 'indexação centrada no usuário' como princípio da indexação que se baseia nas solicitações que são esperadas de determinada clientela. Estudos como os de Fujita (2003) e Boccato e Fujita (2006) mostram a necessidade de uma indexação baseada nessas duas concepções. Na mesma linha de pensamento, Blair (1990) ratifica a importância de o indexador alcançar um equilíbrio entre uma análise voltada para o conteúdo do documento, bem como para o uso potencial desse documento, ou seja, uma análise orientada para os seus usuários.

Para Lancaster (2004, p. 9), a análise conceitual "implica decidir do que trata um documento", isto é, qual o seu assunto. O referido autor ressalta que o indexador deve tentar responder às perguntas "1. De que trata? 2. Por que foi incorporado a nosso acervo? 3. Quais de seus aspectos serão de interesse para nossos usuários?". Para Dias e Naves (2007, p. 5), a análise de assunto "é o processo de ler um documento para extrair conceitos que traduzam a essência de seu conteúdo". Vieira (1988) compreende a "análise intelectual" (análise de assunto) como sendo constituída por três fases: 1ª - compreensão do conteúdo do documento através de leitura deste; 2ª - identificação dos conceitos principais; e 3ª - seleção dos conceitos, levando em consideração a exaustividade, a especificidade e a consistência. Verifica-se que, apesar da existência de algumas diferenças em termos de nomenclatura, essas três atividades mencionadas por Vieira são semelhantes às três colocadas por Dias e Naves (2007), as quais são: "Leitura técnica do documento; Extração dos conceitos e Determinação da atinência".

Lancaster (2004) emprega os termos em inglês '*about e aboutness*' com a seguinte tradução, respectivamente, "trata de" e "atinência[8]". Como as traduções podem variar,

[8] Verifica-se que são encontrados, na literatura especializada, autores que traduzem o termo "aboutness" para "tematicidade", outros para "atinência". Há ainda tradução do referido termo para "concernência, conforme pode ser visto em Baranov (1983). Medeiros (1986) e Fujita (2003), por exemplo, adotam o termo

encontram-se, na literatura da área, autores que utilizam traduções desses termos como "tematicidade", "temática", "assunto", "aboutness" e atinência. Em relação a este último termo, Lancaster (2004, p. 14) ressalta que "o tema da atinência está relacionado muito de perto com o da relevância – isto é, a relação entre um documento e uma necessidade de informação ou entre um documento e um enunciado de necessidade de informação (uma consulta)".

Nesse contexto, Dias e Naves (2007) ressaltam que o termo atinência é utilizado para designar aquilo de que trata o conteúdo substantivo de uma obra, ao invés de se ater tanto, a princípio, à forma ou ao suporte em que a informação está registrada. Esses autores evidenciam que o texto possui uma atinência relativamente estável, permanente, em contrapartida, esse mesmo texto possui distintos significados, os quais variarão conforme o usuário e o momento histórico em que se encontra esse usuário, por exemplo, esse texto poderá ter um significado A para um usuário X em um determinado tempo de sua vida e possuir significado B, totalmente diferente, em outra época para o mesmo usuário X. Em suma, a atinência permanece inalterada, permanente, já o significado dessa atinência variará conforme a percepção que o usuário terá dessa atinência.

Ainda sobre o assunto substancial ou o conteúdo relevante e nessa seara das diferentes traduções do termo em inglês *aboutness*, Begthol (1986) traz uma distinção entre *aboutness* e *meanings*: aquele sendo compreendido como o conteúdo intrínseco ao documento, o qual independe do tempo em que seja utilizado por um usuário, sendo estável; e *meanings* compreende os diferentes significados que os usuários podem fazer desse conteúdo intrínseco. Nesse contexto, Fujita (2003, p. 80) expõe que "a tematicidade sempre será o conteúdo relevante do documento, porém, algumas variáveis irão influenciar na determinação desse conteúdo, como os interesses informacionais dos usuários, [...], entre outras". Cavalcanti (1989), no contexto da linguística, ressalta que "tematicidade intrínseca" é o tema relevante na concepção do autor, já a "tematicidade extrínseca" é o tema relevante para o leitor, as quais, para a autora, são chamadas, respectivamente, de "saliência autor" e "relevância leitor".

Expostas as diferentes traduções e acepções acerca do termo "aboutness" dentro do processo de indexação, pontua-se que se utiliza com mais frequência no decorrer desta pesquisa a expressão "atinência" em vez de "tematicidade", compreendendo-a como "de que

---

"tematicidade", por considerarem que esse termo esteja mais relacionado com a ideia de "tema" do documento. Já Lancaster (2004), Dias e Naves (2007), Naves (1996, 2000), por exemplo, adotam a expressão "atinência". E Todd (1992) ressalta que, na literatura mais recente, é recorrente o termo *aboutness* (atinência) como sinônimo do termo *subject* (assunto) de um documento.

trata um documento", conforme usado na tradução do livro de Lancaster (1993), por ser o mais utilizado na literatura especializada sobre o tema.

Embora Lancaster (2004, p. 15) exponha a "Análise conceitual" como "nada mais do que a identificação dos assuntos estudados ou representados num documento", torna-se evidente, mormente com as pesquisas que vêm sendo desenvolvidas nessa vertente, que essa atividade envolve certa complexidade, sobretudo em sua primeira fase, a qual se desdobra em outras etapas.

Em relação às etapas ou estágios que compõem a análise de assunto, Dias e Naves (2007) elencam: Leitura técnica do documento; Extração dos conceitos e Determinação da atinência. Segundo Fujita (2003, p. 64), baseada nos "Princípios de Indexação" e em uma análise orientada para o conteúdo e pela demanda, ou seja, o usuário, a análise de assunto é subdividida em três estágios: "compreensão do conteúdo do documento; identificação dos conceitos que representam este conteúdo e seleção dos conceitos válidos para recuperação". Reflita-se, portanto, sobre cada uma dessas etapas, fazendo uso da nomenclatura das etapas utilizadas por estes últimos autores nos próximos parágrafos.

Nesse primeiro momento da análise de assunto, **Leitura técnica do documento**[9], o profissional irá fazer uma leitura do documento, buscando informações substanciais que possam indicar o conteúdo da obra. A esse respeito, Dias e Naves (2007, p. 27) ressaltam que:

> No exame do documento, as práticas foram inicialmente desenvolvidas em bibliotecas e principalmente voltadas para livros, que era o tipo de documento predominante. Daí nasceu a expressão *leitura técnica do documento*, significando uma forma de leitura do conteúdo do documento que fosse apropriada para a realização das demais tarefas da análise de assunto: identificação dos conceitos; seleção dos conceitos; e expressão do(s) assunto(s) do documento na forma de uma frase, ou frases, de indexação.

Fujita (2003, p. 84) ressalta que "essa leitura, a documentária, difere da leitura comum porque exige outros procedimentos, ainda que os conhecimentos necessários para um bom entendimento de um texto sejam comuns a ambas".

Em relação à primeira etapa da indexação, ou seja, a análise de assunto, Lancaster (2004, p. 24) recomenda que o indexador faça "um misto de ler e 'passar os olhos' pelo texto", tendo em vista que não há tempo suficiente para que o indexador dedique um maior tempo para a indexação de um único documento, por exemplo. Neste momento, o indexador

[9] Pode-se encontrar, na literatura da área, expressões como "Leitura Técnica do Documento", "Leitura Documentária", "Leitura pelo Indexador" etc.

deve se atentar às partes mais importantes do documento, as quais, consoante o referido autor (2004, p. 24) são "autor, título, resumo, sinopse e conclusões".

Fujita (2003, p. 82), sobre a leitura técnica do documento, ressalta que, "[...] quando falamos em leitura para fins de indexação, podemos dizer que o indexador necessita compreender o texto para identificar e selecionar conceitos". Lancaster (2004, p. 20-21), acerca dessa leitura do documento em indexação, evidencia alguns fatores que afetam diretamente essa leitura por parte do indexador. Isso acontece ao evidenciar que "Ao indexador raramente é dado o luxo de poder ler um documento do começo ao fim".

Com base no exposto, compreende-se que a leitura técnica do documento envolve características próprias, como: a leitura rápida de modo geral de toda a obra; a leitura detalhada de partes específicas consideradas potenciais para a identificação de conceitos e do conteúdo substancial; e uma leitura rápida, tendo em vista as demandas de processamento dentro das UIs. Assim, conforme Fujita (2003), o leitor-indexador utilizará estratégias próprias de leitura documentária com vistas a facilitar a identificação do conteúdo informativo de um documento para fins de representação e posterior recuperação.

Após a leitura técnica do documento, passa-se para o segundo momento da análise de assunto, **Extração de conceitos**, sendo viável antes refletir sobre o entendimento do termo "conceito" utilizado neste trabalho, uma vez que a noção de "conceito" varia conforme o domínio. Segundo a Teoria Geral da Terminologia – TGT, "conceito" é o objeto de estudo da Terminologia, já os termos e os símbolos são as formas designativas dos conceitos. O "conceito" para a Linguística é diferente do entendimento dentro da Terminologia, bem como difere também do uso na Biblioteconomia e Ciência da Informação.

Maculan e Lima (2017, p. 57) ressaltam que, "dependendo da visão epistemológica adotada, o conceito de 'conceito' difere, o que justifica um exame mais reflexivo sobre esse tema". Dubois *et al*. (1997, p. 135) esclarecem que na Linguística "dá-se o nome de conceito a toda representação simbólica, de natureza verbal, que tem uma significação geral conveniente a toda uma série de objetos concretos que possuem propriedades comuns". Nascimento (2008) ressalta que para Saussure o conceito é sinônimo de significado. É a partir dos estudos de Ferdinand Saussure que a Linguística passa a ser considerada uma disciplina científica autônoma (Fiorin *et al*., 2003).

Para Saussure (1996), a língua é constituída por signos, os quais possuem duas faces: o significado (conceito) e o significante (imagem acústica). Na Terminologia[10], a partir dos

[10] Para Dubuc (1999), a Terminologia permite identificar e analisar o vocabulário de uma determinada área do conhecimento e, se necessário, possibilita a criação e normalização dos termos que ocorrem nesse domínio.

estudos de Wuster[11], que iniciou o que se chamou de Teoria Geral da Terminologia - TGT[12], o conceito é considerado, segundo Boutin-quesnel (1985, p. 18), como "uma unidade de pensamento constituída por um conjunto de características atribuídas a um objeto ou a uma classe de objetos e que pode se exprimir por um termo ou por um símbolo". Nessa área, conforme Maculan e Lima (2017, p. 64), o conceito (noção) "é convencionado no contexto de uso, articulado pela comunidade que o compartilha, para, depois se proceder à sua designação por meio de um termo que o represente, numa monossignificação ou monorreferencialidade (um termo só pode representar um conceito), típicas do discurso científico".

Nesta pesquisa, a discussão apoia-se na Teoria Analítica do Conceito, desenvolvida por Ingetraut Dahlberg (1978) que desenvolveu princípios para a identificação e o entendimento dos conceitos na perspectiva da recuperação da informação por eles representados. Para Dahlberg (1978, p. 147):

> [...] conceito é uma unidade do conhecimento, compreendendo afirmações verdadeiras sobre um dado item de referência, representado numa forma verbal [sendo que:] afirmação verdadeira é a componente de um conceito que expressa um atributo do seu item de referência; item de referência é o componente de um conceito para o qual sua afirmação verdadeira e sua forma verbal estão diretamente relacionadas, sendo assim seu 'referente'; forma verbal (termo/nome) de um conceito é o componente que resume convenientemente ou sintetiza e representa um conceito com o propósito de designar um conceito em comunicação.

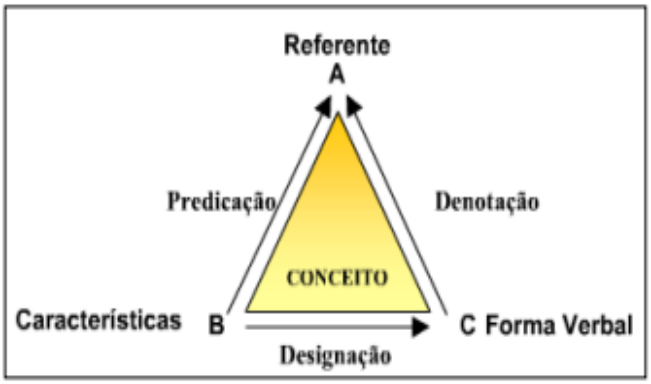
Assim, para Dahlberg (1978), o conceito é compreendido como sendo constituído por três elementos: o *referente* (um objeto, uma atividade, um procedimento etc.), *as características (predicações)* desse objeto e a *forma verbal (termo)* que representa o conceito. É possível visualizar essa tríade que constitui o conceito por meio da Figura 2:

Logo, essa área de estudo denominada Terminologia estuda teoricamente os termos, seus conceitos, os sistemas de conceitos, bem como as suas representações.

[11] Wuster foi fundador da Escola Terminológica de Viena em 1931. Com sua tese de doutorado intitulada *A normalização internacional da terminologia técnica*, inaugura os estudos sobre a chamada Teoria Geral da Terminologia – TGT. Disponível em: http://www.gel.hospedagemdesites.ws/estudoslinguisticos/volumes/32/htm/mesaredo/mr004.htm?/estudoslinguisticos/volumes/32/htm/mesaredo/mr004.htm

[12] O desenvolvimento dos estudos nessa vertente fez surgir críticas à TGT e, por conseguinte, novas teorias evidenciaram perspectivas mais pragmáticas, sociais, como a Teoria Comunicativa da Terminologia – TCT, considerando o termo, além de linguístico e cognitivo, como comunicativo. Em 2000, surge a Teoria Sociocognitiva da Terminologia – TST, de Rita Temmerman. (Maculan; Lima, 2017).

Figura 2 – Triângulo do conceito



Fonte: Dahlberg (1978, p. 149).

Após a compreensão do "conceito" (constituído pelos elementos: referente, características e forma verbal) no contexto da recuperação da informação para a CI, segue-se com as etapas de identificação e extração dos conceitos presentes nos documentos a serem indexados.

Embora as etapas aqui expostas sejam estudadas separadamente, na prática elas são desenvolvidas quase que concomitantemente, uma vez que nenhum indexador irá parar a leitura e dizer a si mesmo "agora começarei a identificar os conceitos". Essas etapas da análise de assunto são desenvolvidas de modo natural por meio de processos cognitivos pelo indexador.

Terminada a "*Leitura Técnica do Documento*", o indexador irá selecionar os conceitos que representam o conteúdo, ou seja, o assunto tratado na obra. Em Dias e Naves (2007), encontra-se a nomenclatura "Extração de Conceitos" para esse segundo momento da análise de assunto. Já, para os "Princípios de Indexação" formulados pela *World Information Systen for Science and Technology* (1981), esta subetapa da análise de assunto é nomeada de "Identificação de conceitos", quando o indexador passa a ter um olhar mais racional e lógico para o conteúdo da obra, por meio da leitura técnica que fez anteriormente. Para Dias e Naves (2007, p. 55), "A identificação de conceitos é uma etapa que, embora logicamente bem próxima da atividade de leitura, tem, entretanto, características próprias. A habilidade principal exigida nesta atividade é a de síntese".

Após a extração/identificação dos conceitos que identificam o conteúdo de um determinado documento, guiada pela compreensão do assunto e do contexto do documento, passa-se para o terceiro momento da análise de assunto, **a determinação da atinência**. Enquanto a "extração/identificação de conceitos" fundamenta-se em uma concepção orientada para o conteúdo, este terceiro momento baseia-se em uma análise de assunto orientada pela

demanda, considerando o assunto como instrumento para transferência de conhecimento (Fujita, 2003, p. 60). Aqui, o indexador irá selecionar, dentre os conceitos extraídos na etapa anterior, aqueles que realmente representam, sintetizam o assunto de um documento. Essa etapa requer do indexador uma capacidade de síntese ainda mais aguçada, tendo em vista que ele escolherá os conceitos que podem ser considerados como representativos do assunto de um documento.

Sobre esta terceira etapa da análise de assunto, Dias e Naves (2007, p. 55) ressaltam que:

> Finalmente, a etapa em que os conceitos identificados são selecionados para efetiva representação no sistema de informação exige do analista de assunto ainda outros tipos de conhecimentos. Os dois mais importantes são o conhecimento dos usuários e o conhecimento da coleção. Como esses são conhecimentos necessários em diversas outras áreas das unidades de informação, certamente que um aspecto importante a ser investigado diz respeito ao que especificamente interessa ao analista de assunto nesses dois tópicos.

Faz-se necessário que se exponha aqui a percepção que Beghtol (1986) acerca da atinência de um documento, uma vez que a autora distingue claramente os termos "atinência" e "significado[13]". Para ela, quando se fala em atinência de um documento, estar a falar do conteúdo relativamente permanente desse documento, ou seja, o seu conteúdo, o qual permanece estável. Em contrapartida, esse mesmo documento possui um número variado de significados, conforme a percepção do usuário. A atinência refere-se ao conteúdo imutável, estável, já o significado moldar-se-á conforme os usuários ou o usuário que faz uso dessa atinência.

Para Dias e Naves (2007, p. 42), "isso varia de acordo com o uso que a pessoa pode encontrar da atinência do documento numa certa época, e o mesmo documento pode ter diferentes significados para o mesmo leitor em diferentes épocas, mas documento, este, imutável, possui uma atinência fundamental".

A "Análise de Assunto", seja na etapa de leitura do documento, na seleção de conceitos ou na escolha dos conceitos representativos do conteúdo de um documento, caracteriza-se como uma atividade interdisciplinar, uma vez que recebe influências de fatores

[13] Tinker (1966) diz que o "significado" pode ser compreendido como a relevância de uma palavra para o conceito que ela rotula. Dias e Naves (2007) acrescentam que "Considera-se significado como a representação, na linguagem, de um significante, correspondendo o primeiro ao conceito ou à noção, e o segundo à forma". Pode-se compreender a ideia de significante como a "imagem acústica" percebida ao proferir-se um palavra com seu significado (Saussure, 1996).

linguísticos, cognitivos e lógicos (Dias; Naves, 2007). Essas influências ficam nítidas quando se percebe que todo o processo envolve o uso da linguagem, seja no momento de analisar e reconhecer as estruturas linguísticas que compõem um determinado gênero documental, seja na elaboração de linguagens documentárias.

A partir do momento que o indexador analisa um documento, por meio de uma linguagem, para representá-lo também por meio de uma linguagem, ele está recebendo influências da Linguística. Entretanto, não são todas as áreas desse campo de estudo que interessam à indexação de documentos, mais especificamente são interessantes a semântica, a sintaxe e a morfologia. Segundo Baranow (1983), as contribuições dessas áreas mencionadas são mais evidentes na prática da indexação automática.

Nesse contexto, têm-se contribuições da Linguística tradicional e da Linguística textual à análise de assunto: aquela com seus aspectos semânticos e sintáticos, bem como com os aspectos linguísticos-pragmáticos; e esta com seus esquemas formais que caracterizam cada tipo de texto. A Linguística textual preocupa-se com textualidade do que vem a ser chamado de texto. Para Fávero e Koch (1988, p. 14), essa é a área responsável por "determinar o que faz com que um texto seja um texto e diferenciar as várias espécies de texto". Logo, parece-nos lógico que um indexador terá mais facilidade de indexar um documento, acerca do qual já se conheça minimamente sua organização interna, suas partes etc.. Além disso, o conhecimento dessa sintaxe própria de cada espécie de texto permitirá ao indexador dedicar-se à leitura das partes potencialmente informativas, conforme já demonstrado por Fujita (2003). Essa forma global de organização do texto, diferenciando as espécies de texto, de acordo com suas formas de organização própria, é chamada de superestrutura por Van Dijk e Kintsch (1983).

Outros aspectos interferem na prática da Análise de assunto, conforme ressaltam Silva e Fujita (2004, p. 147) ao evidenciarem que "é aqui que os aspectos lógicos, linguísticos e cognitivos, envolvidos na indexação, representam fatores de interferência, cabendo ao indexador a habilidade necessária para poder realizar a análise conceitual efetiva do documento". A atividade de analisar um documento, com vistas à identificação do conteúdo para posteriormente representá-lo e permitir a sua recuperação em outro momento, envolve um processamento mental.

Ao observar a percepção de Vygotsky (1991) sobre as funções psicológicas superiores, ficam claras as interferências cognitivas que estão presentes na análise de assunto. Para o autor, as funções psicológicas (ou psíquicas) superiores são aquelas funções mentais que

caracterizam o comportamento consciente do homem: a percepção, a atenção voluntária, a memória e pensamento.

Para Allen (1991, p. 13), os processos cognitivos são “atividades mentais como pensamento, imaginação, lembrança e solução de problemas”. Logo, observam-se várias atividades caracterizadas como processos mentais na prática da análise de assunto, a saber: a memorização ativa, a elaboração conceitual, o uso da linguagem, o raciocínio dedutivo etc. Como exemplo, Dias e Naves (2007, p. 48) expõem que “a compreensão de um texto implica na identificação dos conceitos essenciais nele contidos, e essa compreensão necessita da interação de processos intelectuais complexos, que envolvem a memória do indivíduo e a ativação das estruturas cognitivas”.

Silva e Fujita (2004, p. 153) consideram que os processos cognitivos utilizados pelo leitor durante a leitura são:

> [...] o seu conhecimento sobre a estrutura textual, visando identificar a informação que considera relevante; o conhecimento prévio sobre o assunto do texto e a recuperação de esquemas de compreensão formados com sua experiência de vida que o permite inferir sobre o assunto abordado.

Essas atividades são encontradas, ditas de outra maneira na percepção de Vygotsky (1991), ao estudar as funções psicológicas ou psíquicas superiores. Destarte, os aspectos cognitivos estão presentes em todo o processo de indexação, uma vez que o indexador, a todo momento, está processando informação em sua mente, relembrando assuntos, compreendendo estruturas de texto, diferenciando textos etc.

Além das influências linguísticas e cognitivas expostas, a análise de assunto ainda recebe contribuições da lógica. Conforme Dias e Naves (2007, p. 50), “o uso da linguagem cria vários problemas do ponto de vista lógico, como, por exemplo, a homonímia (várias coisas o mesmo nome), a sinonímia (vários nomes para a mesma coisa), e recorre-se à lógica para resolver esses tipos de obstáculos”. Acrescenta-se aqui outros dois fenômenos a que está sujeito o indexador ao analisar um texto: a polissemia e a ambiguidade. Segundo esses autores, a lógica “consiste numa operação mental que possibilita, através do raciocínio, o surgimento de novas proposições através de proposições já existentes”.

Na prática da análise de assunto, segundo Dias e Naves, (2007, p. 50):

> A importância da lógica formal para a análise de assunto vem do fato de que símbolos lógicos, ao contrário dos lingüísticos, têm um significado perfeitamente exato. Uma das atividades do cientista é o uso de raciocínios; como estes para se explicitarem empregam a linguagem, são ditos

> discursivos. Diante de uma série de fatos, o cientista muitas vezes acredita poder obter conclusões através de argumentos em que usam palavras e expressões como ora, portanto, por conseguinte, etc. Quando entra nesse processo, está raciocinando sobre os fatos que analisou e, desta forma, procurando tirar conclusões (inferindo) a partir deles.

Isto posto, é possível visualizar em maior ou em menor grau a coexistência de fatores ora linguísticos, ora cognitivos e lógicos, o que se faz compreender a análise de assunto como uma atividade complexa, sobre a qual vários estudos vêm sendo desenvolvidos com o intuito de tornar mais claras as operações mentais que constituem essa atividade. Terminada esta etapa, o indexador caminha para a segunda e última etapa da indexação: a tradução.

A **tradução**, consoante Lancaster (2004, p. 18), "envolve a conversão da análise conceitual de um documento num determinado conjunto de termos de indexação". O referido autor faz a distinção entre a "indexação por extração (indexação derivada)" e a "indexação por atribuição". Enquanto na primeira as palavras, expressões, grupos de palavras que ocorrem no documento são extraídos e utilizados para representarem o conteúdo do documento, na segunda, há a atribuição de termos, expressões, grupos de palavras provenientes de outra fonte que não seja o próprio documento, podendo ser extraídas de um Vocabulário Controlado (VC), por exemplo.

Segundo Lancaster (2004, p. 19), "são três os tipos principais de vocabulários controlados: esquemas de classificação bibliográfica (como a Classificação Decimal de Dewey), listas de cabeçalhos de assuntos e tesauros". Ainda segundo esse autor, os vocabulários controlados exercem basicamente três funções: controlar sinônimos, optando por uma forma padrão; diferenciar homógrafos e reunir ou ligar termos cujos significados apresentem uma relação mais estreita entre si.

Segundo Sousa e Fujita (2014, p. 25):

> O segundo estágio da indexação, representação de conceitos por termos de uma linguagem de indexação, direciona a tradução dos assuntos selecionados e centra nos instrumentos de indexação, para assegurar os conceitos de forma útil e acessível. Assim encontramos: os instrumentos verbais, representados por tesauros e listas de cabeçalhos de assunto, entre outros; os instrumentos simbólicos, onde os conceitos são representados por símbolos de classificação.

Assim, a tradução constitui-se da conversão do resultado da Análise de Assunto em termos ou símbolos que representem o assunto. A tradução na atividade de classificação de um livro, conforme a Classificação Decimal Universal-CDU, por exemplo, será por meio de uma notação que, nesse caso específico, compõem-se de símbolos, letras e números. Já a

tradução na atividade de indexação, por exemplo, fará uso de instrumentos verbais, os quais envolvem qualquer recurso que faça uso de termos, palavras, grupos de palavras e conceitos. Independente do instrumento utilizado para a tradução, a tradução objetiva converter o resultado da etapa anterior em elementos sinalizadores, representantes dos conteúdos dos documentos.

De forma semelhante ao que se encontra na área de análise de assunto, há uma variação terminológica quando se refere às linguagens (ou instrumentos) utilizados nesta etapa. Conforme já ressaltado anteriormente, essa variação pode estar relacionada às influências das correntes teóricas que envolvem a atividade de indexação de modo geral, bem como suas etapas e instrumentos utilizados em tal operação. Assim, encontram-se os seguintes termos: vocabulários controlados, linguagens de indexação, linguagens artificiais, linguagens documentárias, linguagem controlada etc.

Inicie-se com uma definição básica de Cintra *et al*. (2002, p. 34), para quem "as LDs são, pois, instrumentos intermediários, ou instrumentos de comunicação, através dos quais se realiza a 'tradução' da síntese dos textos e das perguntas dos usuários". A tradução também ocorre no momento da busca, pois a formalização das questões dos usuários é feita na linguagem utilizada pelo sistema, ou seja, por meio dos termos utilizados pelo sistema da Biblioteca, ou Arquivo etc. É nesse sentido que Cintra *et al*. (2002) falam em "instrumentos de comunicação documentária", permitindo assim que o usuário se comunique com o sistema e, por conseguinte, com o documento de que necessita.

Boccato (2009, p. 119) também utiliza a expressão linguagem documentária, ressaltando que

> [...] são linguagens estruturadas e controladas, construídas a partir de princípios e de significados advindos de termos constituintes da linguagem de especialidade e da linguagem natural (linguagem do discurso comum), com a proposta de representar para recuperar a informação documentária.

Embora a Linguagem documentária se constitua de termos de um determinado campo científico e da linguagem natural (do discurso comum), consoante ressaltado por Boccato (2009), essa linguagem apresenta estrutura e sistemas de relacionamentos precários quando comparados com a linguagem natural, a qual é dinâmica e reconfigura-se a todo momento conforme as necessidades humanas. Entretanto, essas características não tornam as linguagens documentárias ineficientes, pelo contrário, elas permitem a comunicação eficiente entre usuário e sistema por meio de uma linguagem uniforme.

Para Gardin *et al*. (1968), uma LD deve integrar três elementos básicos: um léxico, compreendido como uma lista de elementos descritores, os quais devem ser filtrados; uma rede paradigmática com o intuito de estabelecer certas relações essenciais entre os elementos descritores, cujas relações são, convencionalmente estáveis e podem ser compreendidas como classificações; por fim, uma rede sintagmática com o intuito de representar as relações contingentes entre os elementos descritores, relações essas que são evidentes e válidas no contexto particular em que se encontram. Essa rede sintagmática (nominais, verbais, adjetivais etc.) é construída por meio de regras sintáticas destinadas a estabelecer relações de dependência entre os termos que dão conta do tema.

Lancaster (2004, p. 19), ao se referir à LD por meio do termo "Vocabulário Controlado", diz que "um vocabulário controlado é essencialmente uma lista de termos autorizados. Em geral, o indexador somente pode atribuir a um documento termos que constem da lista adotada pela instituição para a qual trabalha". Contudo, embora possa parecer uma simples lista, os vocabulários controlados se mostram mais complexos, pois envolvem uma estrutura semântica, a qual se destina a controlar termos sinônimos, diferenciar homógrafos e estabelecer relações entre termos que, por exemplo, façam parte de uma mesma esfera semântica.

Rowley (2002, p. 168) define uma linguagem de indexação como "[...] sendo os termos ou códigos que podem ser usados como pontos de acesso num índice". Enquanto Rowley (2002) faz referência ao vocabulário controlado como sendo de dois tipos: *as linguagens alfabéticas de indexação* e os *sistemas de classificação*, Lancaster (2004) expõe três tipos de vocabulários controlados, a saber: esquemas de classificação bibliográfica (como a Classificação Decimal de Dewey – CDD), listas de cabeçalhos de assuntos e tesauros.

Rowley (2002) evidencia que as linguagens de indexação utilizadas na tradução, durante o processo de indexação, são distribuídas em três categorias: *Linguagens controladas de indexação ou linguagens documentárias*, *Linguagens naturais de indexação e Linguagens livres de indexação.*

Para a referida autora, as *Linguagens naturais de indexação* constituem-se de termos que se encontram nos textos, nos discursos dos autores, aqui todos os termos são extraídos do próprio documento. Logo, qualquer termo presente no documento poderá servir como um descritor para o documento do qual foi extraído e *Linguagens livres de indexação*: aqui os termos utilizados, como o próprio nome já diz, são livres, ou seja, podem ser provenientes de qualquer fonte. Silva e Brito (2018, p. 96) definem a linguagem de indexação como um "conjunto de termos padronizados, que representam um conceito da linguagem natural, com

vistas a uma maior padronização dos sistemas de indexação". Barité (2011) e Dodebei (2002) se referem a este instrumento como uma linguagem controlada, artificial, que se constitui ora de termos de domínio específico, ora de termos da linguagem natural com o intuito de representar de forma padronizada os conteúdos presentes nos documentos.

Diferentemente da virtualidade da Linguagem Natural, a linguagem de indexação possui relações estáveis e nitidamente explícitas. Esse controle presente nas linguagens de indexação se faz necessário, uma vez que a linguagem natural apresenta fenômenos que podem prejudicar a comunicação entre usuários e entre usuários e documentos: a sinonímia, a homonímia, a polissemia e a ambiguidade.

Por meio das definições expostas, ora chamadas de Linguagem Documentária, ora de Linguagem de indexação, dentre outras nomenclaturas, foi possível perceber que a expressão "Linguagem de Indexação" é utilizada para se referir a uma linguagem controlada, ou seja, uma linguagem construída para fins documentários.

Assim, se, por exemplo, um bibliotecário, ao indexar um documento opta por utilizar termos que ocorrem no próprio documento, sem nenhuma forma de controle ou de termos previamente estabelecidos, ele estará fazendo uso de uma Linguagem natural de indexação, conforme exposto por Rowley (2002). Como apenas a expressão "Linguagem de indexação" pode ser interpretada como representando tanto uma linguagem controlada como uma linguagem natural, as nomenclaturas adotadas por Rowley (2002) evidenciam essa distinção de forma mais clara ao utilizar os termos: Linguagem controlada de indexação e Linguagem natural de indexação.

2.1.1 Princípios da Indexação: especificidade, exaustividade, revocação e precisão

Uma das definições para o termo "Princípios", que se encontra nos dicionários de Língua Portuguesa, diz respeito "ao conjunto de proposições fundamentais e diretivas que servem de base e das quais todo desenvolvimento posterior deve ser subordinado". No contexto da indexação, os princípios podem ser compreendidos dessa forma, ou seja, como "elementos básicos" que fundamentam o processo de indexação.

No tocante aos princípios que permeiam a indexação, Rubi (2009, p. 82, grifo nosso) ressalta que "esses princípios deverão influenciar o bibliotecário na sua decisão sobre a determinação de conceitos cujo resultado será observado pelo usuário na recuperação da informação. São eles: *especificidade, exaustividade, revocação e precisão*".

No tocante à **Exaustividade**[14], Lancaster (2004, p. 27) ressalta que "corresponde, grosso modo, ao número de termos atribuídos em média [...]. A indexação exaustiva implica o emprego de termos em número suficiente para abranger o conteúdo temático do documento de modo bastante completo". Sobre esse princípio, Rubi (2009, p. 85) ratifica esse entendimento acerca da exaustividade ao afirmar que:

> A exaustividade diz respeito ao número de termos atribuídos como descritores do assunto do documento, ou seja, em que medida todos os assuntos discutidos no documento são reconhecidos durante a indexação e traduzidos na linguagem documentária da biblioteca. Quanto mais exaustiva for a indexação, mais termos ela vai empregar. É indicada, por exemplo, em bibliotecas de público variado e de diferentes perfis, que podem buscar a mesma informação com termos diferentes.

A quantidade de termos a serem utilizados, um número mínimo, um máximo, ou uma média frequente são definidos nas *políticas de indexação*, as quais são elaboradas, segundo Carneiro (1985, p. 221), com base nos seguintes fatores: características e objetivos da organização, determinantes do tipo de serviço a ser oferecido; identificação dos usuários, para atendimento de suas necessidades de informação; e recursos humanos, materiais e financeiros, que delimitam o funcionamento de um sistema de recuperação de informações.

Não é objetivo aqui adentrar, de forma exaustiva, na temática 'Políticas de indexação', mas, apenas expô-la como instrumento que norteará a indexação e, consequentemente, o grau de exaustividade adotado por unidades de informações ou SRIs. Sobre as políticas de indexação, dentre os vários trabalhos publicados acerca dessa temática, temos Rubi; Fujita (2003), Fujita (2006), Leiva; Fujita (2012), Oliveira (2017) etc.

A exaustividade, consoante Soergel (1994), pode ser vista de duas formas: a exaustividade de pontos de vista e a exaustividade de importância (relevância). A primeira garante que as facetas ou os pontos de vista, julgados como úteis para a representação proposta pelas LD, serão disponíveis para a recuperação da informação. Já a segunda diz respeito ao nível de importância (de relevância) dos descritores propostos pelas regras de indexação.

Em relação a este princípio, Pinto (2000) evidencia que ele pode ser responsável pelo surgimento do *ruído*. A autora utiliza a categorização "ruído ou barulho" para se referir ao excesso de documentos propostos pelos sistemas de recuperação da informação como resposta a uma demanda, mas que na realidade não correspondem ao assunto demandado. Esse *ruído*

[14] Encontra-se na literatura da área o termo "profundidade" para se referir à exaustividade, como pode ser visto em Lancaster (2004).

está diretamente ligado ao grau de exaustividade de determinada indexação, quanto mais termos utiliza-se para representar um documento, maior as chances de aumentar o ruído.

Segundo Lancaster (2004, p. 29), "quantos mais assuntos forem incluídos mais *exaustiva* será a indexação. Por outro lado, quanto menos assuntos forem incluídos mais **seletiva** será a indexação. Evidentemente a indexação exaustiva exigirá o emprego de maior número de termos". Um dos fatores que afetam diretamente a decisão de fazer uso de uma ou de outra indexação (exaustiva ou seletiva) diz respeito ao custo demandado pelas UIs e pelos SRIs, pois os indexadores precisarão de mais tempo, caso optem pela indexação exaustiva. Outro fator que afeta decisivamente essa decisão é o público-alvo.

A **especificidade** diz respeito ao ato de ser o mais específico possível na representação temática de um determinado documento, de forma análoga ao princípio da exaustividade. Aqui não se busca generalizar um determinado conteúdo por meio de um termo genérico, senão escolher termos específicos, utilizando termos isolados ou, quando necessário, realizando combinações entre termos por meio de adjuntos adnominais, por exemplo.

Acerca desse princípio, Lancaster (2004, p. 34) ressalta que "o princípio que, isoladamente, é o mais importante da indexação de assuntos, e que remonta a Cutter (1876), é aquele segundo o qual um tópico deve ser indexado sob o termo mais específico que o abranja completamente". Por exemplo, uma obra que trate da "Pedagogia Libertadora", deveria ser indexada, segundo esse princípio, sob o termo "Pedagogia Libertadora" e não sob "Tendências Pedagógicas", visto que este último é bem mais genérico, uma vez que envolve várias concepções, como: a concepção tradicional, a libertadora, a libertária, o tecnicismo educacional etc.

Para Lancaster (2004, p. 34), nesse contexto, "seria melhor utilizar vários termos específicos, ao invés de um termo que seja mais genérico". Outro ponto levantado por esse autor diz respeito à "redundância" que alguns indexadores podem cometer, ao atribuir termos genéricos e termos específicos ao mesmo tempo, acreditando que está cobrindo todo o conteúdo, mas na verdade, com essa prática, ficará difícil diferenciar os artigos específicos de artigos genéricos.

Para Rubi (2009, p. 85):

> [...] a especificidade está relacionada ao nível de abrangência que a biblioteca e a linguagem documentária permitem especificar os conceitos identificados no documento. Exemplo: um livro cujo assunto seja especificamente sobre "tilápias" será indexado sob o assunto "peixes". Essa situação é característica de bibliotecas que optam por uma baixa

> especificidade nos assuntos que, por sua vez, trará como resultados na recuperação uma alta revocação.

Acerca desse princípio, Pinto (2000, p. 66) diz que "esta maneira de indexar diz respeito à profundidade com a qual o conteúdo de um documento é tratado. Se, de uma parte, ela favorece a precisão, de outra, contribui para aumentar o silêncio na recuperação da informação". Essa autora, utiliza o termo "**silêncio**" para se referir à "ausência de documentos que responderiam às necessidades dos usuários, mas na verdade não foram encontrados, mesmo fazendo parte da coleção". Assim, estar-se diante de medidas inversas, pois, se por um lado, ao ser específico, alcança-se uma melhor precisão, ao ser genérico, perde-se em precisão e alcança-se uma melhor revocação. O caminho, talvez, seja alcançar um equilíbrio entre essas duas medidas, uma vez que são inversas.

Em relação à especificidade, Lancaster (2004, p. 35) ressalta ainda que "o indexador deve ter em mente que é possível conseguir especificidade mediante combinações de termos. Se não houver nenhum termo que sozinho possa representar o tópico, busca-se uma combinação apropriada de termos do vocabulário controlado". Assim, um ponto importante que deve permear sempre o processo de indexação, no que tange a esses princípios, é a *ambiguidade*, que pode ocorrer durante o processo, uma vez que o bibliotecário indexador pode ingenuamente acreditar que seja importante ser generalista e específico ao mesmo tempo, afetando profundamente as medidas de revocação e precisão nas buscas. Para se evitar isso, faz-se necessário que o indexador repense continuamente o objetivo da instituição, as características de sua coleção e de seu público.

Conforme já mencionada, a **capacidade de revocação** diz respeito à competência de um sistema de recuperação de informação para recuperar documentos relevantes, úteis à determinada necessidade informacional. Carneiro (1985, p. 234) ressalta que a revocação "se relaciona com a capacidade do sistema em assegurar a recuperação de um número desejável de documentos relevantes". Rubi (2009, p.85) evidencia que a capacidade de revocação "pode ser mensurada por meio da relação entre o número de documentos relevantes sobre determinado tema, recuperados pelo sistema de busca, e o número total de documentos sobre o tema, existentes nos registros do mesmo sistema".

Quanto à **capacidade de precisão**, Lancaster (2004, p. 4) utiliza os termos revocação e precisão, para se referir, respectivamente, "à capacidade de recuperar documentos úteis" e "a capacidade de evitar documentos inúteis". A expressão "precisão" pode ser encontrada na literatura da área por meio do sinônimo "relevância". Para Rubi (2008, p. 85-86), a

capacidade de precisão, ou relevância, "está relacionada ao número de documentos recuperados para atendimento das solicitações encaminhadas pelo usuário".

De forma semelhante à mensuração da revocação, a capacidade de um sistema de ser preciso pode ser avaliada por meio da relação entre os documentos relevantes recuperados e o número total de documentos recuperados. Por exemplo, se em uma busca, o sistema recuperou 155 documentos, dos quais apenas oito são considerados pelo usuário como documentos relevantes àquela busca específica, a precisão do sistema será baseada na relação entre esses dois dados.

Por se tratarem de medidas inversas, uma vez que, ao se aumentar a revocação, diminui-se automaticamente a precisão, é preciso que o sistema tenha seus objetivos bem definidos em uma Política de Indexação, a qual norteará os graus de exaustividade e de especificidade, utilizados no processo de indexação, os quais afetam diretamente as capacidades de revocação e de precisão do sistema.

Rubi (2009, p. 86) esclarece que:

> [...] quanto mais exaustivamente um bibliotecário indexa seus documentos, maior será a revocação na recuperação da informação buscada e, inversamente proporcional, a precisão será menor. E quanto mais especificamente um bibliotecário indexar, menor será a revocação, porém a precisão será maior.

Carneiro (1985) ressalta que a decisão de dar ênfase a um ou outro princípio será baseada nos usuários do sistema, visto que esses "podem ter diferentes exigências quanto aos níveis de revocação e precisão. Assim, usuários mais gerais se satisfarão com uma alta revocação, enquanto que uma clientela bem mais específica, obviamente, se satisfará com um sistema que, embora ofereça poucos documentos como resposta a uma busca, recupere documentos precisos, muito próximos da demanda dos usuários.

Há também usuários que podem se satisfazer com um equilíbrio entre a revocação e a precisão oferecido pelo sistema.

2.1.2 Avaliação da Indexação

Os elementos envolvidos na atividade de indexação: leitor, texto e contexto (Fujita, 2003) caracterizam-na como uma atividade subjetiva, sobretudo por ser uma operação realizada pelo ser humano, embora este tente ser o mais objetivo possível, imparcial, não tendencioso, o processo não deixa de ser subjetivo.

A esse respeito, Lancaster (2004, p. 68) evidencia que:

> É mais do que evidente que a indexação é um processo subjetivo e não objetivo. Duas (ou mais) pessoas possivelmente divergirão a respeito do que trata uma publicação, quais aspectos merecem ser indexados, ou quais os termos que melhor descreve os temas selecionados. Ademais, uma mesma pessoa decidirá de modo diferente quanto à indexação em momentos diferentes. A coerência na indexação refere-se à extensão com que há concordância quanto aos termos a serem usados para indexar o documento.

Essa subjetividade faz com que um mesmo profissional/indexador indexe um mesmo documento de forma completamente diferente em momentos distintos ou que diferentes indexadores divirjam na atividade de indexação. Essa coerência mencionada por Lancaster (2004) refere-se ao grau de concordância entre um mesmo indexador ou diferentes indexadores ao atribuir os termos representativos do conteúdo de um documento. O referido autor (2004, p. 68), acerca dessa coerência, expõe que "talvez a medida mais comum seja a simples relação *AB(A=B)*, onde A representa os termos atribuídos pelo indexador *a* e B representa os termos atribuídos pelo indexador *b*, e AB representa os termos com os quais *a* e *b* concordam". Essa consistência é medida por diferentes equações, por exemplo, a equação de Hooper (1965) e Holling (1981). Hooper (1965) refere-se a essa coerência entre um mesmo indexador ou entre indexadores como *pares de coerência (PCs)*, os quais serão calculados pela comparação entre os termos semelhantes atribuídos pelos indexadores.

Lancaster (2004, p. 68) esclarece que "a coerência *interindexadores* refere-se à concordância entre indexadores, enquanto a coerência *intraindexador* refere-se à extensão com que um indexador é coerente consigo mesmo". Para Cooper (1969), a coerência interindexadores é avaliada com base no nível do termo, ou seja, a coerência é verificada por meio do grau de concordância dos indexadores em utilizar um determinado termo, por exemplo, o termo X a um determinado documento. Assim, a coerência interindexadores é definida como a proporção de indexadores que atribuem o termo em questão, menos a proporção daqueles que não o atribuem.

Os estudos de avaliação de coerência na indexação baseiam-se, em sua maioria, na semelhança entre os termos atribuídos, embora outras possibilidades sejam passíveis, por exemplo, a atribuição de pesos maiores à coerência na atribuição de termos considerados mais importantes para a representação do conteúdo de determinado documento.

Outro aspecto a ser considerado como estudos de coerência diz respeito à semelhança parcial entre os termos, uma vez que se podem ter indexadores que atribuam um mesmo

cabeçalho principal e divirjam no subcabeçalho. Nessa vertente de avaliação da coerência, há de se mencionar a quantidade de termos atribuídos no momento da indexação, conforme ressalta Lancaster (2004), ao dizer que "[...] é certo que haverá mais concordância quanto aos tópicos do documento considerados principais do que quanto aos tópicos considerados de menor importância que mereçam ser incluídos". O referido autor, supondo que o fato de os termos serem atribuídos em ordem de prioridade, levanta "a hipótese de que a concordância atingirá o nível mais alto no nível de dois termos e em seguida começará a cair gradualmente até o ponto onde tenham sido atribuídos tantos termos que a concordância voltará a aumentar" (2004, p. 70). Alguns possíveis fatores que afetam a coerência na indexação são levantados pelo autor em questão, a saber:

Figura 3 – Possíveis fatores que influem na coerência da indexação

1. Quantidade de termos atribuídos
2. Vocabulário controlado *versus* indexação com termos livres
3. Tamanho e especificidade do vocabulário
4. Características do conteúdo temático e sua terminologia
5. Fatores dependentes do indexador
6. Instrumentos de auxílio com que conta o indexador
7. Extensão do item a ser indexado

Fonte: Lancaster (2004, p. 71).

Diversos fatores podem afetar os resultados alcançados pela indexação. Lancaster (2004), inspirado pelas ideias de Oliver *et al*. (1966), explica que a indexação pode ser afetada por fatores ligados ao indexador, por exemplo, conhecimento do assunto, experiência etc., fatores ligados ao vocabulário, especificidade, qualidade do vocabulário etc., fatores ligados ao documento, linguagem, extensão, fatores ligados ao processo, tipo de indexação, nível de exigência feita aos indexadores e fatores ambientais, aos quais se podem citar: ambiente onde é realizada a indexação, excesso de barulho etc.

Nesse contexto, Narukawa, Gil Leiva e Fujita (2009, p. 110) condensam os fatores que mais incidem sobre o produto da indexação em três aspectos, os quais são "o próprio indexador, o objeto analisado e o contexto em que se concretiza".

Dentre os elementos ligados ao próprio indexador, pode-se citar o domínio da temática, uma vez que bibliotecários especialistas em domínios específicos se comportam de maneira bem diferente de bibliotecários indexadores generalistas, que lidam com várias temáticas em suas práticas diárias. A formação profissional também se vincula ao próprio indexador. Elementos como complexidade, tamanho, propriedades, ligados ao *objeto* a ser

indexado, afetam os resultados alcançados. Ao *contexto* pode-se associar os aspectos ligados às políticas de indexação, as necessidades dos usuários etc.

Por ser resultado de um processo subjetivo, a indexação não possui um resultado considerado correto ou errado em sentido absoluto, como ressalta Lancaster (2004, p. 86), ao dizer que "não existe nenhum conjunto 'melhor' de termos. Alegar que tal conjunto existe implica uma presciência de todos os pedidos que serão feitos à base de dados na qual o documento de acha representado". Não obstante, há estratégias que permitem verificar a coerência entre os termos atribuídos a determinados documentos, avaliando os termos que foram atribuídos por um mesmo indexador, ou por indexadores distintos ou metodologias de indexação e, ainda, avaliando a indexação por meio de buscas simuladas por parte de usuários.

Sobre esse último método de avaliação, cujo processo de avaliação busca avaliar se os documentos indexados são de fato recuperados pelos usuários que deles necessitam, Lancaster (2004, p. 87) ressalta que

> [...] é possível testar o trabalho dos indexadores de uma maneira mais rigorosa do que simplesmente passando os olhos pelos termos atribuídos, que é o máximo que se pode esperar de uma operação rotineira de checagem. O método mais evidente consiste em realizar uma simulação de uma avaliação real.

Para o referido autor, esse método consiste nos seguintes passos:

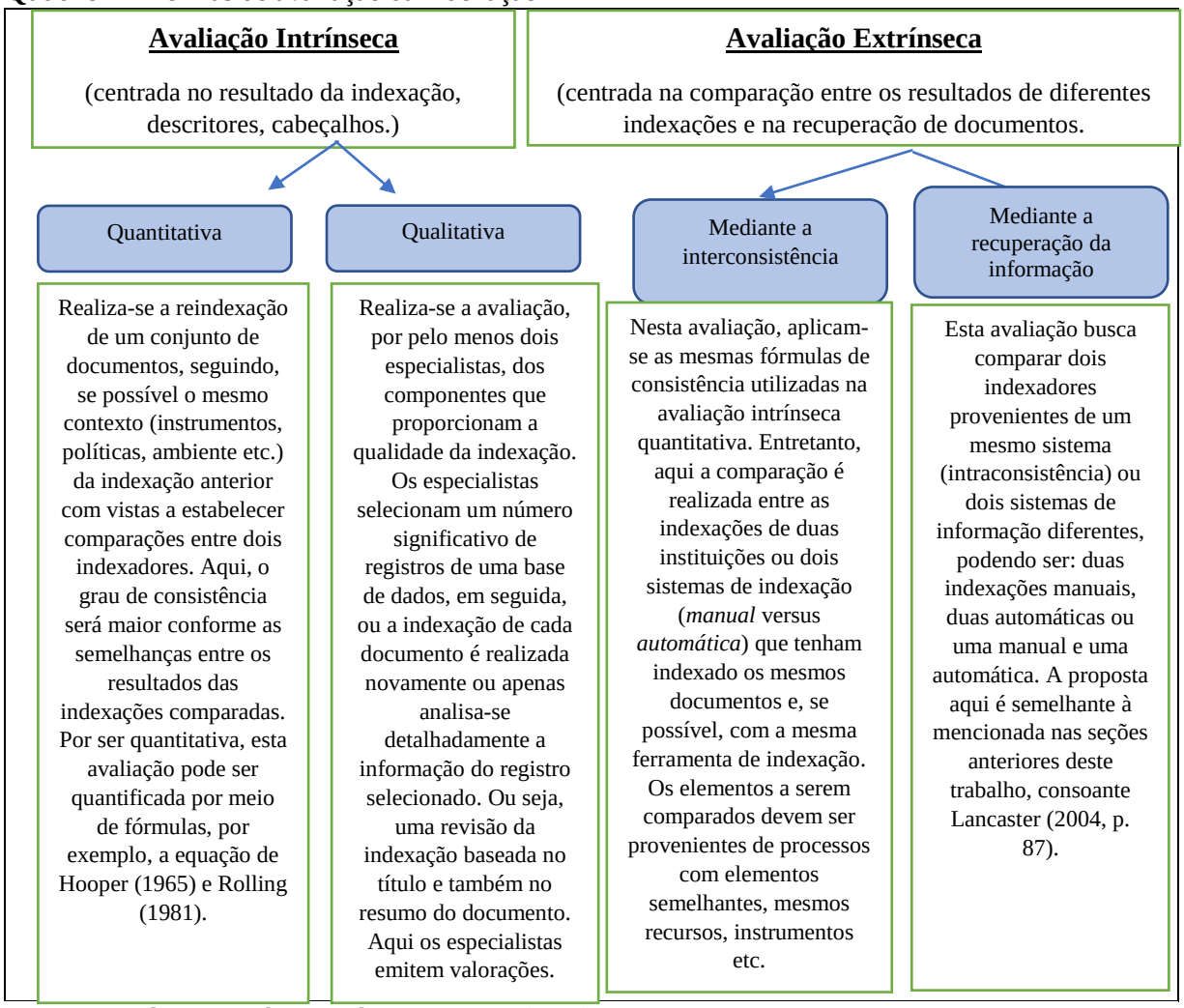
> 1. Selecione um grupo de documentos dentre os que compõem o fluxo normal de entrada, antes que cheguem às mãos dos indexadores.
> 2. Para cada documento elabore, digamos, três questões para as quais o item seja considerado uma resposta importante. Uma das questões se basearia no tema central do documento enquanto as outras estariam centradas nos temas secundários, mas ainda assim importantes.
> 3. Faça com que experientes analistas de buscas elaborem estratégias de busca para cada uma dessas questões. É claro que esses analistas não devem ser as mesmas pessoas cuja indexação estará sendo examinada.
> 4. Faça com que os itens sejam indexados da forma rotineira.
> 5. Compare a indexação com as estratégias de busca, a fim de determinar se os itens relevantes são recuperáveis ou não com os termos atribuídos (Lancaster, 2004, p. 87).

O método de avaliação exposto por Lancaster (2004), apresentado anteriormente, encaixa-se no que Narukawa, Gil Leiva e Fujita (2009) chamam de avaliação extrínseca da indexação, uma vez que possui o propósito de testar a função da indexação na recuperação dos documentos indexados. Para os referidos autores, a avaliação da indexação se dá de duas

maneiras gerais: a avaliação intrínseca e a avaliação extrínseca. Naquela avalia-se o resultado (produto) da indexação, ou seja, os cabeçalhos, subcabeçalhos, rótulos, buscando verificar a qualidade desses descritores atribuídos. Nesta, a avaliação extrínseca, realizam-se comparações entre os resultados das indexações feitas entre diferentes unidades de informação, ou tipos de indexação (manual e automática, por exemplo), além de testar a indexação na recuperação dos documentos indexados.

Na figura abaixo, é possível visualizar de forma detalhada essas duas formas de avaliação propostas por Narukawa, Gil Leiva e Fujita (2009).

Quadro 4 – Formas de avaliação da indexação



Fonte: Baseado em Narukawa, Gil Leiva e Fujita (2009).

Complementando o Quadro 4, no tocante à *avaliação intrínseca qualitativa*, Narukawa, Gil Leiva e Fujita (2009, p. 108, grifo nosso) ressaltam que os especialistas

> [...] devem emitir valorações e conseguir consensos ao menos nestes componentes inerentes a uma indexação de qualidade: **Exaustividade** (Que

> se extrai todos os conceitos caracterizadores do conteúdo íntegro dos documentos), **Especificidade** (Que exista uma relação exata entre as unidades conceituais escolhidas e o termo ou os termos eleitos para representá-la), **Correção** (Que não se produza erros de inclusão (um termo que não corresponde) nem erros de omissão (a exclusão de um termo que corresponde) e **Perspectiva do usuário** (Que se considere os interesses e necessidades do usuário).

No tocante à avaliação intrínseca quantitativa, a qual visa comparar os resultados de indexações em diferentes momentos, buscando identificar os índices de consistência entre as indexações realizadas em diferentes momentos, ou seja, a intraconsistência, de acordo com Narukwa, Gil Leiva e Fujita (2009, p. 109), executa-se "quando um profissional indexa novamente um documento transcorrido um tempo (seis ou doze meses para comprovar se se produzem variações com respeito à primeira indexação".

Essa avaliação quantitativa pode ser operacionalizada por meio das equações apresentadas no Quadro 5:

Quadro 5 - Equações de índices de Consistência

Hooper (1965)	Rolling (1981)
$\frac{C}{A+B-C}$ Uma variação desta equação é: $\frac{100C}{C+A+B}$ onde, C= Termos comuns nas duas indexações A= Termos usados na indexação A mas não em B B= Termos usados na indexação B mas não em A	$\frac{2C}{A+B}$ onde, C= Termos comuns nas duas indexações A= Termos usados na indexação A B= Termos usados na indexação B

Fonte: Gil Leiva (2008, p. 386).

Essa fórmula de Hooper é utilizada por Gil Leiva (1999, 2003 e 2008), como *Índice de Consistência*, o qual se executa da seguinte forma:

Quadro 6 – Índice de consistência

$C_i = \frac{T_{co}}{(A+B)-T_{co}}$	Tco = Número de termos comuns nas duas indexações A= Número de termos usados na indexação A B= Número de termos usados na indexação B Como se pode verificar, com esta equação, os índices de consistência oscilam entre os valores 0 e 1 e depois é possível multiplicar o resultado por cem para obter o percentual.

Fonte: Adaptado de Narukara, Gil Leiva e Fujita (2009, p. 109).

No que diz respeito à Avaliação Extrínseca da Indexação, tem-se a avaliação mediante a interconsistência, a qual busca comparar as indexações de diferentes unidades de informação ou ainda diferentes formatos de indexação, uma indexação manual e automática, por meio das mesmas fórmulas mencionadas acima.

Já, na avaliação extrínseca, mediante a recuperação, a indexação é avaliada por meio de buscas nas quais os documentos estejam indexados, com vistas a identificar se tais documentos são ou não recuperados. Para isso, segundo Narukawa, Gil Leiva e Fujita (2009, p. 110), procede-se da seguinte forma:

> A avaliação extrínseca mediante a recuperação [15]consiste em interrogar duas bases de dados que contém *(sic)* os mesmos campos e idênticos conteúdos salvo os campos que armazenam a indexação. Com os resultados obtidos, se encontram os índices de exaustividade e precisão na recuperação. Este método de avaliação é custoso, mas proporciona resultados claros, mediante o uso de fórmulas de exaustividade e precisão na recuperação: Exaustividade = Número de documentos relevantes recuperados Número de documentos relevantes na coleção Precisão = Número de documentos relevantes recuperados Número total de documentos recuperados.

A avaliação da indexação apresenta diversas utilidades dentro de uma unidade de informação, permitindo mensurações periódicas entre um mesmo indexador, ou diferentes indexadores, ou ainda diferentes sistemas de indexação. Seja qual for o propósito final, a avaliação da indexação sempre trará benefícios, ou apontando os intervenientes ou evidenciando os pontos positivos. Por ser uma atividade composta por elementos subjetivos e sofrer interferências de variados fatores, é pertinente que a avaliação seja uma prática constante nas unidades de informação. Por exemplo, se se pensar em uma instituição que, devido à demanda, queira incorporar um sistema de indexação semiautomático em determinada coleção digital, a avaliação entre a indexação manual e a automática permitirá à unidade de informação verificar se o sistema produz resultados próximos ao da indexação manual, fazendo com que se opte ou não pelo sistema.

Vistas as formas de avaliação da indexação, mormente, as formas de comparação entre indexações realizadas por diferentes instituições, bem como distintas metodologias, por exemplo, uma indexação manual e uma automática, passe-se agora a discutir sobre as propostas de indexação automática.

[15] É possível encontrar, em Lancaster (2004, p. 87) e em Narukawa, Gil Leiva e Fujita (2009, p. 110), o passo a passo para se realizar a avaliação extrínseca mediante a recuperação.

2.2 INDEXAÇÃO SEMIAUTOMÁTICA

Como já mencionado, a indexação semiautomática agrega os benefícios advindos do processamento automático de textos digitais à avaliação manual feita pelo indexador, a qual se mostra fundamental, tendo em vista que os *softwares*, mesmo com todos os avanços, apresentam limitações.

Assim, a indexação semiautomática envolve dois momentos gerais: a análise e extração de termos (descritores) dos documentos ou atribuição de termos (descritores) provenientes de Linguagens Documentárias dos documentos analisados e, em seguida, a avaliação desses descritores por um indexador humano, o qual busca validar os termos selecionados automaticamente, como evidenciam Lancaster (2004) e Moreiro González (2004).

As propostas semiautomáticas se diferenciam das automáticas pelo fato de que naquelas há a participação do indexador humano ao final da identificação e seleção dos termos pelos programas computacionais, enquanto nestas dispensa-se a validação realizada pelo homem. É notório o avanço nas pesquisas sobre a indexação automática e, com isso, os benefícios advindos dessas propostas, ao mesmo tempo em que a literatura da área consolida a indexação manual como parâmetro e referência à avaliação das propostas automáticas. É perceptível, na literatura da área, de um lado a existência de estudos voltados à indexação manual e à busca constante por compreender os processos mentais envolvidos em tal processo e, por outro lado, os estudos que buscam automatizar por completo esse processo, ou seja, os estudos e experiências de indexação automática.

A indexação semiautomática envolve uma etapa inicial feita pelo *software* e um segundo momento realizado pelo indexador humano, o qual terá a incumbência de avaliar os termos e decidir sobre quais termos serão descritores documentais representativos do documento do qual foram retirados. Assim, consoante Pinto (2001, p. 227), essa indexação é realizada da seguinte maneira:

> [...] inicialmente, o sistema faz uma indexação automática dos documentos levando em conta as ocorrências das palavras mais frequentes no texto. Em um segundo momento, o indexador humano refina a lista dos descritores propostos pelo sistema fazendo os ajustes e/ou complementações necessárias.

Já se evidenciou neste trabalho a importância da indexação manual realizada por um profissional qualificado para tal, bem como a complexidade de tal procedimento, o que, por

sua vez, justifica as limitações ainda presentes em propostas automáticas. Assim, parece-se lógico estudar propostas que se beneficiem ora dos avanços da indexação automática, ora das singularidades que apenas o humano consegue dar conta. Nessa perspectiva, propostas de indexação semiautomática podem ser bastante frutíferas e, consequentemente, se mostrar eficientes, mormente, na atual conjuntura de crescimento exponencial de produção de informação.

Como proposta de indexação semiautomática, cite-se o SISA desenvolvido na Espanha por Isidoro Gil Leiva (1999, 2008). O SISA foi inicialmente proposto à área de Biblioteconomia e Documentação, contudo sua flexibilidade permite adaptação a outras áreas, desde que possua uma Linguagem Documentária da área. De modo geral, o processo de indexação desenvolvido por este *software* é executado por meio de três módulos: no *módulo 1* há o pré-processamento do documento a ser indexado, o qual é preparado e sinalizado para que seja lido pelo software. Essa sinalização envolve a marcação das partes que serão indexadas. Os marcadores #CTI#, por exemplo, representando o "começo do título", #FTI# representando o "fim do título" e dessa forma acontece com as outras partes indexadas pelo SISA. O referido *software* utiliza as seguintes fontes na indexação: o texto completo (título, resumo e texto), uma lista de palavras vazias (*stoplist*) e uma linguagem documentária, todos em formato txt. (Narukawa, Gil Leiva, Fujita, 2009). De acordo com os referidos autores, nesse módulo (2009, p. 106) "ocorre também a eliminação das palavras vazias mediante a comparação com a lista de palavras vazias e então é computado o total de palavras das fontes título, resumo e texto".

Após o processamento das fontes nesse primeiro módulo, ocorre no *módulo 2*, a "análise de conteúdo", que, conforme Narukawa, Gil Leiva e Fujita (2009, p. 106), consiste no procedimento em que um algoritmo realiza a:

> [...] busca e seleciona os termos preferidos que são os coincidentes com os termos da linguagem documentária; [...] os termos não preferidos que são os termos sinônimos, por isso não podem ser utilizados e remetem aos preferidos e; os termos construídos sintaticamente de forma diferente dos termos preferidos que são as palavras semivazias, aquelas que o sistema julga importante, mas não se enquadram nas anteriores.

O último módulo consiste em estabelecer a valoração e ponderação dos termos por meio da aplicação de um critério de avaliação dos termos para que assim o sistema possa selecionar os termos representativos do conteúdo do documento indexado. Esse processo se

mostra necessário para indicar os termos possivelmente mais representativos, senão o sistema selecionaria todos os termos da Linguagem Documentária que coincidissem com os das fontes (título, resumo e texto). Os critérios aplicados nesse módulo são basicamente dois: posição do termo e frequência de ocorrência.

Ainda conforme Narukara, Gil Leiva e Fujita (2009, p. 106/107), para que o SISA selecione e proponha os termos finais para indexação, aplicam-se os seguintes critérios:

> [...] o termo é apresentado como termo de indexação se, um termo autorizado aparece na fonte título e na fonte resumo, ou se um termo autorizado aparece na fonte título e na fonte texto, ou se um termo autorizado aparece na fonte resumo e na fonte texto. No entanto, os termos não autorizados ou semivazios são apresentados como candidatos à indexação se, a palavra semivazia aparece no título, resumo e texto, ou se aparece no texto dez vezes ou mais, além de aparecer em oito parágrafos diferentes ou mais.

Por último, por se tratar de uma proposta de indexação semiautomática, há a participação do indexador humano, o qual possui a incumbência de analisar os termos de indexação propostos pelo *software*, bem como os termos semivazios candidatos a descritores, apresentados em lista, que podem ou não ser incluídos como descritores documentais, ou seja, termos de indexação. O sistema permite ainda a possibilidade de inclusão de termos da linguagem documentária. Nesse momento, o indexador, levando em consideração as peculiaridades do sistema de informação, ao qual está vinculado, pode tomar a decisão de adicionar ou suprimir termos.

Os *softwares*, nas mais variadas áreas, vêm alcançando desempenhos satisfatórios, contudo, ainda apresentam limitações, quando comparados à indexação realizada pelo humano. Em pesquisa realizada por Narukawa, Gil Leiva e Fujita (2009), foi avaliada a indexação automatizada do SISA, sob os aspectos: análise da indexação com o SISA, consistência da indexação: comparação da indexação automática com a indexação manual e avaliação da exaustividade e precisão na recuperação com o SISA.

Alguns fatores que interferiram no desempenho do *software*, segundo Narukawa, Gil Leiva e Fujita (2009, p. 115), foram:

> a) A dificuldade de o SISA atribuir termos compostos: o SISA atribui mais termos simples ao invés de termos compostos e a estratégia de busca por termo composto torna inviável a recuperação de artigos científicos. Exemplos: O SISA atribuiu os termos neoplasias e glândulas salivares, enquanto a estratégia de busca se constituiu do termo neoplasias das glândulas salivares. Neoplasias e boca atribuídos pelo SISA e a estratégia de

> busca neoplasias bucais. Os termos materiais para moldagem odontológica, assistência odontológica para pessoas portadoras de deficiência, cisto odontogênico calcificante dificilmente são atribuídos pelo SISA, ao contrário da indexação manual que não apresenta dificuldade.

Lima e Boccato (2009) avaliaram o desempenho terminológico dos descritores em CI do vocabulário controlado do SIBI/USP nos processos de indexação manual, automática e semiautomática e, conforme já constatado por Naraukawa, Gil Leiva e Fujita (2009), perceberam que o SISA não reconheceu os descritores compostos, por exemplo, **Ciência da Informação**, quando no texto surge também o termo *Informação*. Segundo Lima e Boccato (2009, p. 142), "isso ocorre porque *Informação* também é um descritor do Vocabulário e acaba sendo contado isoladamente. Consequentemente obtém um maior número de ocorrências".

Outros intervenientes identificados pelas autoras dizem respeito: à variação de termo do artigo científico em relação à linguagem documentária; ao uso no artigo científico exclusivamente de termos relacionados com o termo relevante "como em *oclusão dentária bilateral* ao invés de *oclusão dentária balanceada"* da LD DeCS; e ao fato de o SISA não ter atribuído que se encontrava apenas em uma parte da estrutura do artigo científico. É perceptível que esses inconvenientes que afetaram negativamente o desempenho do SISA podem ser dirimidos com a atualização e aperfeiçoamento da Linguagem Documentária utilizada. Além disso, o critério de ocorrência concomitante do termo em duas partes do documento pode ser revisto, buscando um melhor desempenho do *software*.

Evidenciando algumas limitações do *software* SISA, Lima e Boccato (2009, p. 143) ressaltam que:

> [...] ao processar apenas arquivos txt, o programa fica limitado ao que se encontra disponível nesse formato, o que também leva à necessidade de preparo do *corpus* a ser indexado. O aprimoramento do programa para indexar textos em outros formatos, mais especificamente o pdf, seria de grande valia para os sistemas de informação, visto que a maioria dos repositórios de textos completos se encontra nesse formato.

Ao compararem o desempenho terminológico das indexações manuais, automática e semiautomáticas, Lima e Boccato (2009, p. 143) consideram que a utilização do SISA, na indexação semiautomática, se mostra mais adequada, uma vez que permite ao indexador avaliar novos termos que surgem com os termos candidatos a descritores. Isso possibilita a expansão dos termos da área de estudo, assim, a indexação semiautomática permite atualizar

linguagens documentárias de domínios específicos, por meio do levantamento de termos candidatos a descritores documentais.

A indexação semiautomática vem suprir as limitações ainda presentes na indexação totalmente automática, uma vez que não exclui do processo o humano, o qual incorpora ao processo de representação temática a capacidade de identificação de conceitos implícitos, bem como a capacidade de analisar o contexto (de modo mais amplo) de determinada obra indexada, que ainda são habilidades presentes na indexação manual. A indexação semiautomática, que conjuga em um único processo a indexação automática e a participação do indexador humano, mostra-se como uma estratégia viável para dar conta do volume de informação, visto que alcança o processamento de grandes volumes de documentos, sem, com isso, eximir desse processo o indexador humano, o qual ainda é insubstituível pela máquina.

Após a aproximação com as propostas de indexação manual e semiautomática, dediquemo-nos agora aos estudos sobre a indexação automática e sobre os sintagmas nominais, os quais vêm se mostrando como alternativa promissora na prática da automação da indexação. Entenda-se, a partir da próxima seção, um pouco sobre as propostas de indexação automática e as estruturas linguísticas, que são utilizadas nesta pesquisa como unidade de representação da informação, bem como sua recuperação em coleções digitais em unidades de informação.

2.3 INDEXAÇÃO AUTOMÁTICA

Os procedimentos de representação temática da informação, sobretudo a indexação, quando associados a recursos computacionais, recebem distintas nomenclaturas, sendo possível encontrar, na literatura da área, os seguintes termos: indexação automática, indexação automatizada, indexação mecânica, indexação semiautomática, indexação auxiliada por máquina e indexação assistida por computador. Conforme pode ser visto, na revisão de literatura sobre a automatização da indexação realizada por Gil Leiva (1999; 2008), alguns desses termos são utilizados como sinônimos, outros são termos que representam conceitos distintos, sobre os quais falar-se-á a seguir.

Robredo (1986) compreende a indexação automática como sendo qualquer procedimento que possibilite a identificação e a seleção de termos que representem o conteúdo dos documentos, sem a intervenção direta do documentalista. Aqui é preciso distinguir a indexação automática da indexação manual que, embora realizada pela máquina, tem a intervenção humana. Vieira (1988) diferencia a indexação automática da indexação

manual, ressaltando que a automática identifica palavras ou expressões consideradas significativas com vistas a descrever o conteúdo de forma condensada por meio de *softwares* (programas) capacitados para isso. Artandi (1976) utiliza a expressão "indexação mecânica" para se referir à indexação realizada pela máquina.

Guimarães (2000) expõe a indexação realizada total ou parcialmente pela máquina em três concepções: *a primeira* envolve o uso de computador e de seus programas como suportes para armazenamento dos termos de indexação, obtidos pela análise conceitual; *a segunda*, caracterizada pelo uso de sistemas que analisam documentos de forma automática com a validação dos termos selecionados pelos programas informáticos por um profissional (indexação semiautomática); e *a terceira* envolve a indexação totalmente realizada pela máquina, sem a intervenção do indexador humano, chamada de indexação automática propriamente dita, consoante as definições de Robredo (1986), Vieira (1988), bem como outros autores.

Silva e Fujita (2004, p. 145, grifo nosso) ressaltam que a "indexação automatizada seria, portanto, aquela resultante do trabalho intelectual de um profissional para checagem do valor dos termos atribuídos a um documento por um programa de computador". Assim, para essas autoras, a indexação automatizada refere-se à indexação realizada pela máquina com validação de um profissional. Contudo, há autores que utilizam a expressão indexação semiautomática para se referir a esta atividade, o que indica que se estaria diante, então, na percepção dessas autoras, de termos sinônimos: indexação automatizada e indexação semiautomática.

Em termos de definição, conforme o Pequeno Dicionário Houaiss (2015, p. 106), o vocábulo "automatizado" vem de automatizar, que se refere a "prover de máquinas, para agilizar produção. Tornar-se maquinal ou inconsciente (pessoa, ação, reação etc.)", remetendo a algo que ficou automático; que passou a funcionar por um sistema de mecanização, logo pode-se compreender os termos "automático" e "automatizado[16]" como sinônimos, sendo mais viável, semanticamente falando, utilizar a expressão "semiautomática" para a indexação realizada pela máquina com a validação final de um profissional.

Isto posto, retomando a percepção de Guimarães (2000, p. 1), pode-se compreender a indexação associada aos recursos computacionais de três formas: indexação auxiliada ou

[16] É possível encontrar, em diversos trabalhos, a expressão "automatizada" como sinônima de "automática". Um exemplo disso pode ser percebido no artigo de Narukawa, Gil Leiva e Fujita (2009), os quais utilizam frequentemente a expressão automatizada. Logo, utilizamos a referida expressão neste trabalho como sinônima de "automática".

assistida por computador (ou por máquina), esta sendo inteiramente realizada pelo indexador humano, fazendo uso de programas de computador apenas para armazenamento de temos de indexação; a indexação semiautomática, que é realizada pela máquina, mas com validação final feita pelo indexador humano; e a última que é a indexação automática ou automatizada, realizada totalmente pela máquina, sem intervenção do indexador humano.

Após essas definições e distinções, é preciso que se entendam os motivos que impulsionaram o desenvolvimento de estudos de indexação automática. Vários pesquisadores evidenciaram distintos motivos e argumentos para credibilizar a indexação automática em detrimento da indexação manual, todavia alguns argumentos são comuns entre os pesquisadores para justificarem esses estudos, por exemplo, a morosidade da indexação manual, a subjetividade intrínseca à prática manual e o volume crescente de documentos digitais que não acompanha a capacidade de processamento manual. Sobre essa questão, Bandim e Corrêa (2019, p. 2) explicam que "no contexto das bases de dados científicas, a indexação automática tem sido adotada visando dar conta da indexação do volume crescente de artigos científicos e da necessidade de criação de índices para busca". Complementando, Borges (2010, p. 15), em relação aos estudos sobre a indexação automática, ressalta que "nota-se que seu surgimento se deu devido à necessidade de serem resolvidos problemas como a morosidade trazida pela indexação manual.

Por isso, a indexação automática é vista como uma alternativa para agilizar esse processo, através dos recursos oferecidos pela tecnologia". Ratificando essa perspectiva, Borges e Lima (2015, p. 50) esclarecem que, "embora a indexação automática possa não apresentar resultados totalmente satisfatórios, suas soluções podem contribuir para significativas melhoras no processo de indexação manual". Ainda nessa linha, Gil Leiva (1997, p. 60) ressalta que, "em suma, subjetividade, lentidão e custo são argumentos importantes contra a indexação intelectual, enquanto os defensores de sua automação alegam consistência, rapidez e exaustividade, que levam a maior produtividade e qualidade na indexação".

Encontra-se, em Silva e Fujita (2004), o percurso teórico e metodológico detalhado na prática da indexação manual e automática. Segundo essas autoras, as primeiras propostas de indexação automática baseavam-se nas palavras significativas dos títulos de documentos técnicos, os índices *Keyword-in-Context* (KWIC) e *Keyword-out-of-Context* (KWOC). Em Silva e Fujita (2004), encontram-se referências acerca da criação desse método para William Frederick Poole com surgimento em 1960. Este índice é mais conhecido como "índice

permutado por computador" e, nesse caso, é atribuído a Hans Peter Lunh, com criação em 1953 (Pinto Molina, 1993).

A história da indexação automática é baseada em métodos estatísticos que, em seus primeiros estudos, se mostraram superficiais e pouco rigorosos, abrindo, contudo, os estudos nessa vertente. Consoante Gil Leiva (1999, 2008), a automação da indexação se desenvolveu em três momentos históricos: *as propostas estatísticas*, seguidas de *propostas com ênfase em aspectos linguísticos*, e *os métodos mistos ou híbridos*, conjugando as duas propostas anteriores.

Os métodos estatísticos consistem em confrontar os documentos com listas de palavras vazias, eliminando as palavras que não serviriam como representativas dos conteúdos dos documentos, por exemplo, conjunções, preposições etc. Ademais, é atribuído peso a cada palavra com base na frequência de ocorrência de cada uma, assim, por exemplo, uma palavra que aparece vinte vezes tem um peso maior do que uma que aparece três vezes. Os métodos de base estatística estão presentes na maioria dos estudos voltados à automação da indexação e utilizam-se das palavras como recursos para representação temática dos documentos e recuperação deles. De acordo com Kuramoto (2002, não paginado), "No último século, a grande maioria dos modelos de recuperação de informação utilizou e utiliza a palavra como acesso à informação".

Segundo Narukawa, Gil Leiva e Fujita (2009, p. 103):

> A utilização do critério de frequência só foi possível através dos métodos estatísticos que foram os primeiros a surgir. Zipf em 1949 desenvolveu o princípio do mínimo esforço que se referia ao valor constante que tem a relação entre a frequência das palavras e a posição que essas ocupam na ordem frequencial. A partir dessas ideias, Hans Peter Luhn em 1957 sugeriu que a frequência das palavras em um texto tem relação com a utilidade que teriam na indexação. Para ele, a frequência com que as palavras aparecem no texto expressa quais são as palavras representativas do conteúdo do texto.

Contudo, os métodos eminentemente estatísticos apresentaram limitações, sobretudo no que se refere aos aspectos linguísticos. Dessa forma, o segundo momento dos estudos de automação da indexação dedicou-se a considerar os aspectos linguísticos do texto, buscando compreender a estrutura textual, as relações entre as palavras e os significados possíveis às estruturas linguísticas. Conforme Gil Leiva (2008) os primeiros analisadores linguísticos surgiram na década de 1960 para o processamento automático de informação.

Assim, nessa perspectiva, os *softwares* (analisadores linguísticos[17]), consoante ressaltado por Narukawa, Gil Leiva e Fujita (2009, p. 103), "se dedicam ao tratamento das palavras (analisador morfológico), ao tratamento das orações (analisador sintático) e ao tratamento das palavras e orações segundo o contexto em que se encontram para conhecer seu significado (analisadores semânticos)".

Após esses dois primeiros momentos, surgem as propostas que conjugam tanto os métodos estatísticos quanto os linguísticos, beneficiando-se do que as duas perspectivas podem oferecer às propostas de indexação automática. Conforme Guimarães (2000), esses métodos mistos ou híbridos apoiam-se nas propostas estatísticas, linguísticas e ainda incorporam tesauros como recursos destinados ao controle terminológico, evitando, assim, intervenientes ligados a fenômenos como a sinonímia, ambiguidade etc.

Vários são os critérios utilizados nas propostas de indexação automática, sempre buscando alcançar o mesmo propósito: identificar os termos descritores mais representativos dos conteúdos dos documentos, a fim de que estes sejam recuperados de forma eficiente. Esses critérios se desenvolveram conforme os avanços nas pesquisas voltadas à indexação automática, conforme já mencionados neste trabalho.

Borges e Lima (2015, p. 53-54), por meio da análise da literatura da área, desde a década de 1950 até o ano de 2008, identificaram 16 critérios utilizados para o desenvolvimento de *softwares* para indexação automática. Consoante as referidas autoras (2015, p. 54), "abrangeram-se somente os elementos essenciais para apoio no processo de escolha dos melhores critérios para o desenvolvimento de softwares de indexação automática". Os critérios são expostos no quadro seguinte:

[17]Segundo Othero e Menuzzi (2005, p. 49), um *parser*, no contexto da Linguística Computacional, "é um analisador automático (ou semiautomático) de sentenças [frases]. Esse tipo de programa é capaz de analisar uma sentença com base em uma gramática preestabelecida de determinada língua, verificando se as sentenças fazem parte ou não da língua, de acordo com o que autoriza a sua gramática. Um *parser* também analisa sintaticamente as sentenças decompondo-as em uma série de unidades menores".

Quadro 7 - Listagem de dezesseis critérios identificados na literatura

CRITÉRIO 1 - Formatação de frases-termo (Word phrase formation);
CRITÉRIO 2 - Fórmula de transição de Goffman;
CRITÉRIO 3 - Frequência absoluta de ocorrência da palavra no texto;
CRITÉRIO 4 - Frequência de co-ocorrência relativa de termos;
CRITÉRIO 5 - Frequência de co-ocorrência simples de termos;
CRITÉRIO 6 - Frequência relativa de ocorrência da palavra no texto;
CRITÉRIO 7 - Identificação de palavras (Comparação com uso de dicionário);
CRITÉRIO 8 - Identificação de radicais de palavras (Word stemming);
CRITÉRIO 9 - Lista de palavras proibidas (Stop-list/stop-words);
CRITÉRIO 10 - Palavras destacadas no texto;
CRITÉRIO 11 - Peso numérico;
CRITÉRIO 12 - Posição do termo no texto (Term weighting);
CRITÉRIO 13 - Primeira lei de Zipf;
CRITÉRIO 14 - Segunda lei de Zipf ou Lei de Zipf-Booth;
CRITÉRIO 15 - Tópico frasal;
CRITÉRIO 16 - Vocabulário semântico / Cabeçalhos conceituais / Tesauro.

Fonte: Borges e Lima (2015, p. 53-54).

Após a sistematização dos critérios mencionados no Quadro 7, as autoras iniciam a análise da combinação dos critérios utilizados nos textos, com vistas à interpretação dos resultados obtidos, quantificando os critérios mais utilizados nas pesquisas (amostragem) analisadas e verificando a utilização prática dos critérios mencionados. Do total de 16 critérios, 50% deles apresentaram uma taxa de utilização acima de 30% em relação ao número total de pesquisas analisadas.

Assim, os critérios mais utilizados foram: *Identificação de palavras (Comparação com uso de dicionário)*, *Formatação de frases-termo (Word phrase formation)*, *Posição do termo no texto (Term weighting)*, *Peso numérico*, *Identificação de radicais de palavras (Word stemming), Frequência absoluta de ocorrência da palavra no texto, Vocabulário semântico/vocabulário de cabeçalhos conceituais/Tesauro* e *Lista de palavras proibidas/Palavras proibidas (Stop-list / stop-words).*

Embora se tenham distintos critérios, sobretudo os que, cada vez mais, buscam extrair a semântica dos textos, alguns critérios eminentemente estatísticos, como frequência absoluta de ocorrência, continuam a ser utilizados nas pesquisas atuais, mormente por meio da combinação com outros critérios também estatísticos e critérios que consideram os aspectos linguísticos dos termos que compõem um determinado texto. Para os primeiros, pode-se citar como exemplo, frequência relativa de ocorrência, frequência de co-ocorrência relativa de termos, frequência de co-ocorrência simples de termos; para os segundos, o uso de tesauros, os quais apresentam grande potencial para o tratamento semântico do texto

Ao lado dos critérios eminentemente estatísticos e dos que consideram os aspectos linguísticos e semânticos dos documentos, há também as listas de palavras proibidas, que contribuem incisivamente para eliminação de termos desprovidos de semântica, por exemplo, as preposições, conjunções etc. Outros critérios bastante utilizados são: formatação de frases-termo (buscam formar novos termos por meio da junção de termos adjacentes e que carregam uma semântica mais específica, menos generalista); identificação de radicais de palavras (buscam realizar a seleção ou exclusão de termos que apresentem determinado radical); peso numérico (identifica a frequência absoluta do termo no texto e a frequência inversa na coleção); e, por fim, a posição do termo no texto (parte da ideia de que determinadas partes de um texto são potencialmente mais informativas do que outras, tendo, assim, mais chances de possuírem nessas partes termos que possuem mais chances de serem bons descritores).

As propostas de indexação, no tocante à fonte de onde são retirados os termos, são classificadas, consoante Lancaster (2004), de indexação por extração automática e indexação por atribuição automática. Aquela fazendo uso de termos que ocorrem no próprio documento, ou seja, utilizando-se a linguagem natural, e esta utilizando termos de alguma linguagem documentária.

Conforme Bruzinga, Maculan e Lima (2007, não paginado), os sistemas de indexação por extração automática realizam basicamente as seguintes tarefas: "(1) contar palavras num texto; (2) cotejá-las com uma lista de palavras proibidas; (3) eliminar palavras não significativas (artigos, preposições, conjunções etc.); (4) ordenar as palavras de acordo com sua frequência".

Já a indexação por atribuição automática envolve procedimentos mais complexos, uma vez que está implicado o controle terminológico dos termos utilizados para indexação. Neste procedimento, de acordo com Bruzinga, Maculan e Lima (2007, não paginado) "Deve-se desenvolver, para cada termo atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer nos documentos. Por exemplo, para o termo 'chuva ácida' incluir-se-iam as expressões 'precipitação ácida', 'poluição atmosférica', 'dióxido de enxofre' etc.".

No âmbito da indexação automática por atribuição, Silva e Corrêa (2020) discutem essa operação e realizam um estudo comparativo acerca dos sistemas de indexação automática por atribuição SISA (Sistema de Indexação Semiautomático) e MAUI. O SISA foi desenvolvido na Espanha, proposto inicialmente para a área da Biblioteconomia e Documentação, já o MAUI (Multi-purpose Automatic Topic Indexing), de origem neozelandesa.

Silva e Corrêa (2020, p. 22) verificam que:

> [...] os sistemas apresentam valores para as características em comum, porém o MAUI se mostra mais promissor por conta dos recursos adicionais de processamento de linguagem natural que implementa, e por apresentar a especificidade de treinamento de um modelo de indexação por meio de um algoritmo de aprendizado de máquina. O uso de aprendizado de máquina promete oferecer melhor eficácia na representação da informação por gerar modelo a partir da indexação manual, considerada a melhor forma de indexação por pesquisadores da área.

Estratégias de indexação automática que se aproximem da indexação manual se mostram bem mais promissoras, uma vez que a prática manual consegue lidar com fenômenos próprios da Linguagem Natural, bem como identificar elementos descritores dos conteúdos dos materiais informacionais. As propostas de indexação automática por atribuição, quando comparadas com a indexação automática por extração, apresentam a particularidade de possuírem maior controle do vocabulário e, por conseguinte, dos termos candidatos a descritores, uma vez que tomam como parâmetro um vocabulário de domínio específico, como um tesauro.

Embora as propostas de indexação automática ainda apresentam limitações quando comparadas à indexação manual, as pesquisas vêm apresentando resultados promissores. Consoante Lancaster (2004, p. 52), "vários programas de computador foram desenvolvidos para gerar, automaticamente, um conjunto de entradas de índices a partir de uma sequência de termos". Podem-se citar, por exemplo, os modelos Listagem Seletiva em Combinação - SLIC, o Preserved Context Index System - PRECIS, o KWIC, o KWOC e o Nested-phrase indexing system - NEPHIS. Dentre esses, o KWIC, para Robredo (1982, p. 238), pode ser considerado como a primeira proposta generalizada de indexação automática de documentos técnicos a partir das palavras significativas dos títulos desses documentos. O *Keyword in Context*– KWIC (Palavra-chave no contexto) constitui-se de um índice rotativo em que cada palavra-chave que aparece no título de um documento, tornando-se uma entrada do índice. Assim, para Lancaster (2004), as palavras que são significativas são destacadas, e as palavras restantes aparecem associadas ao termo de entrada, envolvendo-a. O critério para identificar as palavras que irão constituir o índice é denominado de processo 'reverso', visto que ele elimina as palavras desprovias de significado, por exemplo, preposições, conjunções etc., restando apenas as palavras que são significativas e que constituirão as entradas do índice.

Há vários estudos e experiências voltadas à indexação automática, sejam *softwares* voltados à indexação propriamente dita, sejam analisadores linguísticos, ou ferramentas de

análise de textos, etc. Nesse contexto, encontram-se variadas propostas, como o SISA[18], o PRECIS, o Automindex, o SitagMed, o SiRILiCO, o OGMA etc.

Segundo Vieira (1988), os estudos acerca da indexação automática, no contexto brasileiro, iniciam-se no final dos anos 60, com o desenvolvimento do KWIC (Keyword In Context) cujo intuito era construir os índices das bibliografias especializadas. Na década de 70, os estudos sobre a indexação automática centram-se na análise da frequência de palavras. Conforme Gil Leiva (1997), na década de 80, desenvolvem-se estudos que, além da abordagem estatística, basearam-se em aspectos linguísticos, como o estudo de Andreewski e Ruas (1983), que trata da adaptação da proposta francesa Systéme Syntaxique et Probabiliste d´Indexation et de Recherche d´Informaticos Textuelles (SPIRIT) para documentos em língua portuguesa.

Segundo Lapa e Correa (2014, p. 61), "o uso de referenciais linguísticos, mais exatamente de critérios sintático-semânticos, tal como a proposta de uso de sintagmas nominais como unidades de análise, estão presentes nos trabalhos de alguns autores brasileiros a partir da década de 90". Esses trabalhos iniciam com Kuramoto (1995), Souza (2006) e Borges, Maculan e Lima (2008).

Robredo (1991) propôs um sistema de geração de termos indexadores a partir da análise automática de títulos e resumos de textos, o Automindex. Esse *software*, em um primeiro momento, analisa o título e o resumo do documento, comparando as palavras que o compõem com as palavras de dois antidicionários: primeiro, com o dicionário de palavras invariáveis (preposições, advérbios, conjunções), as palavras que forem identificadas por este dicionário são desconsideradas, desprezadas; segundo, o texto analisado é comparado com o segundo antidicionário, o qual é constituído por palavras cujas raízes são consideradas como "não significativas à área de conhecimento em questão". Mais uma vez, as palavras que apresentam semelhança com esses radicais são eliminadas, em seguida as palavras restantes, após o cotejamento com esses dois antidicionários, são consideradas como possíveis descritores documentais.

Por fim, essas palavras são comparadas com um dicionário de palavras significativas. Se forem identificadas nesse dicionário, serão consideradas como descritores, já as que não forem identificadas são consideradas como candidatas a descritores para avaliação posterior, por meio da qual se decidirá se elas serão ou não incorporadas. Esses termos candidatos a descritores aparecem acompanhados de sua frequência de ocorrência na base de dados. Essa

[18] A subseção 2.3 INDEXAÇÃO SEMIAUTOMÁTICA expõe-se, de forma mais detalhada, o funcionamento deste sistema, o qual se configura como uma proposta de indexação semiautomática.

frequência de ocorrência se mostra um critério pertinente, uma vez que pode nos mostrar o quão um termo é geral (por ocorrer demais em vários documentos) ou é específico (por aparecer poucas vezes em outros documentos).

Com o desenvolvimento de pesquisas na área, sobretudo de mecanismos que permitem a análise mais aprofundada das estruturas linguísticas de documentos textuais, e não mais apenas da ocorrência de palavras isoladas, surgiram novas propostas de automação da indexação. As estratégias mais recentes se apoiam não nas palavras soltas, as quais podem adquirir vários ou nenhum significado, mas em estruturas sintagmáticas, ou seja, os sintagmas nominais que, por sua vez, possuem maior potencial semântico quando comparados com as palavras isoladas. A seção "**2.3.1 O uso dos Sintagmas Nominais na Indexação automática**" expõe-se, de forma detalhada, essas unidades sintagmáticas que compõem os textos.

Kuramoto (1995) propôs um modelo de indexação automática, de base linguística, centrada na identificação de sintagmas nominais que podem ser compreendidos como palavras ou grupos de palavras que fazem parte de sequências maiores na estrutura de um texto. Essas palavras se relacionam e mostram um grau de coesão entre elas.

Segundo Kuramoto (1997, p. 11):

> Considerando as limitações dos computadores e os problemas existentes no uso dos SRIs tradicionais, esta proposta poderá resolver os problemas colocados na introdução deste documento. O procedimento de indexação proposto consiste na extração dos sintagmas nominais e na sua indexação, como descritor, segundo uma estrutura em árvore. Em relação à interface de recuperação de informação, o enfoque proposto se baseia num procedimento de navegação na estrutura de sintagmas nominais.

Nesta proposta de Kuramoto (1997), há a possibilidade de o usuário construir uma expressão de busca por meio da navegação das estruturas dos sintagmas nominais até que seja encontrado o sintagma que represente a sua necessidade de informação. Após esse encontro, o usuário solicita ao sistema a busca, aqui o sistema faz a busca dos documentos de onde foram extraídos o sintagma nominal a que o usuário chegou.

Esta proposta diferencia-se substancialmente das estratégias que fazem uso de palavras isoladas, as quais já demonstraram limitações na representação e na busca da informação, tendo em vista fenômenos linguísticos como a polissemia, a sinonímia, a ambiguidade etc. presentes na língua natural. Nesse sentido, de acordo com Nascimento (2015, p. 19), algumas pesquisas vêm sendo desenvolvidas com o intuito de demonstrar a viabilidade do uso dos SN

como recursos potenciais para a recuperação de informações, por exemplo, as de "Kuramoto (1995; 2002), Souza (2005: 2006), Maia (2008), Corrêa *et al*. (2011), Lopes (2012), Silva (2014), Souza e Raghavan (2014) entre outros". As propostas de indexação automática devem ir além de critérios eminentemente estatísticos, ou seja, é necessário que se conjuguem critérios provenientes de métodos estatísticos, linguísticos, bem como de recursos que atuem no controle terminológico dos descritores, como já ressaltado por Guimarães (2000), os quais são chamados por ele de métodos mistos ou híbridos.

Nessa perspectiva, Bruzinga, Maculan e Lima (2007) evidenciam o papel da "Semântica" e da "Sintaxe" nas propostas de automação da indexação. Conforme as referidas autoras (2007, não paginado), "a semântica e a sintaxe têm papéis importantes na indexação automática, na medida em que permitem ao *software* identificar a estrutura lexical das frases e o significado dos termos que representam o conteúdo do documento".

De modo geral, a sintaxe se preocupa com a organização (relacionamento) entre as palavras que formam frases e textos. Desse modo, a sintaxe estuda as formas adequadas à linguagem formal de combinação e organização interna das frases de uma determinada língua. Analisa, para isso, os elementos constitutivos, por exemplo, sujeito, verbo, complemento etc. Já a semântica se preocupa com o significado[19] da frase construída, seguindo as recomendações da sintaxe. Entretanto essas duas áreas de estudo podem aparecer de maneira inversa nos enunciados. Veja-se o exemplo trazido por Bruzinga, Maculan e Lima (2007, não paginado): "a chuva gosta de cair sobre meus cabelos ruivos' = frase com sintaxe correta, porém sem semântica. 'Fingimos que fumus e vertemos' = frase com semântica, porém, sem sintaxe correta".

Isto posto, percebe-se o quão complexo é o desafio de capacitar as máquinas por meio dos *softwares* para realizar a análise de estruturas linguísticas, verificando se as mesmas estão adequadas, corretas e, por conseguinte, o sentido expresso por elas naquele contexto específico. É possível encontrar, em Bruzinga, Maculan e Lima (2007), uma proposta de indexação automática que faz uso de critérios sintático-semântico, por meio da qual a indexação automática é executada por meio de um *parser*, o Tropes, realizando uma análise léxico-semântica e morfossintática e da Taxonomia da área de Ciência da Informação.

[19] Para Ullmann (1964, p. 111), o significado "é um dos termos mais ambíguos e controversos da teoria da linguagem". Consoante esse autor, o significado da palavra "é a relação recíproca e reversível entre o som e o sentido" (1964, p. 117), ou seja, ao ouvir uma palavra, logo se pensará no objeto por ela nomeado e, assim, compreenderá o seu significado. Neste trabalho, compreende-se que o significado de uma palavra está intimamente ligado ao contexto de uso, o qual pode se modificar conforme as diferentes situações comunicacionais. Assim, o significado é móvel, ou seja, é suscetível de mudança.

Segundo as autoras,

> [...] para analisar o conteúdo utilizando critérios sintático-semânticos, o Tropes usa os recursos de uma gramática sintagmática e de um cenário padronizado determinado previamente à análise. [...]. Nessa nossa proposta, construiremos um cenário a partir de uma taxonomia da área de Ciência da Informação.

As propostas de indexação automática buscam dar conta do volume de informação que vem sendo produzido na atual conjuntura, embora ainda apresente limitações, tendo em vista a própria natureza na linguagem e seus fenômenos. Quaisquer que sejam as propostas de automação da indexação, é notória a necessidade de que os *softwares* sejam capacitados para analisarem um documento digital em seus aspectos linguísticos, a exemplo da semântica e da sintaxe. Entretanto, ainda que a indexação automática venha apresentando grandes avanços na área de representação temática da informação, há ainda muitas limitações quando se estuda a automação da indexação. Esses intervenientes mostram que a indexação manual ainda é a prática tida como referência no dia a dia das unidades de informação, bem como validada por meio das pesquisas científicas da área de CI.

É crescente o número de estudos que vêm sendo desenvolvidos e que se relacionam direta e indiretamente com a indexação automática. Embora a completa automação da indexação venha se mostrando promissora, sobretudo na realidade hodierna, essa prática ainda não está isenta de inconvenientes e limitações, especialmente quando comparada aos resultados alcançados com a indexação manual, a qual ainda é considerada como parâmetro para a avaliação da automática.

A esse respeito, Lapa e Correa (2014, p. 74) evidenciam que:

> Analisando o tipo de validação dos termos, percebemos a preferência pela aplicação da indexação semiautomática. O que pode ser justificado pelo fato de os processos totalmente automáticos ainda serem falhos e apresentarem limitações tecnológicas. Entretanto, a diferença em relação ao processo automático, na segunda posição, não é muito grande, podendo ser interpretada como um esforço no desenvolvimento de uma indexação automática de qualidade.

Como as propostas de indexação totalmente automáticas ainda apresentam intervenientes, a indexação manual (ou semiautomática) é frequentemente utilizada como parâmetro de referência na validação dos termos selecionados automaticamente. É nesse ponto que reside o caráter ímpar da indexação semiautomática, dado que faz uso da agilidade e objetividade da indexação automática, sem com isso eximir a parte humana do processo.

Tavares e Celerino (2018, p. 13), ao estudarem a relação entre a Bibliometria e a indexação automática, expõem o valor que as técnicas métricas têm para a automação da indexação, evidenciando o uso de instrumentos bibliométricos que contribuem diretamente para a escolha de descritores de documentos.

Os referidos autores, entretanto, ressaltam que:

> É importante destacar o valor que as leis de Zipf e de Ponto de Transição de Goffman possuem para a indexação automática. Já que ambas trabalham a ocorrência de palavras, diversos estudiosos aplicaram essas leis para a construção de dicionários que possibilitassem aos sistemas de indexação automática melhorar a sua qualidade e eficácia. Contudo, deve-se ressaltar que é necessário o desenvolvimento de mais estudos com base nas aplicações dessas leis bibliométricas, pois muitos dos trabalhos já realizados constatam algumas adversidades quanto a sua aplicação em sistemas diferentes.

Nesse contexto, Bandim e Correa (2019), ao estudarem a indexação automática por atribuição, evidenciam que tal indexação, com base no uso do *software* SISA com o Tesauro Brasileiro de Ciência da Informação (TBCI), atende, de forma satisfatória, quanto à qualidade na indexação automática. No entanto, apontam para as limitações que ainda estão presentes no processo automático de indexação, ao ressaltar que

> Durante a avaliação da qualidade na indexação automática por atribuição, foram encontrados fatores intervenientes que apontam para possíveis aprimoramentos nos algoritmos do SISA, visando uma melhor qualidade na indexação automática por atribuição com o uso de vocabulário controlado. (Bandim; Correa, 2019, p. 11).

A constatação desses autores ratifica a percepção de Lapa e Correa (2014) quando constatam que os processos totalmente automáticos ainda são falhos e apresentam limitações tecnológicas.

No tocante às limitações ainda presentes nos processos automáticos, Bandim e Correa (2019, p. 11) esclarecem que:

> Mesmo com um desempenho satisfatório do processo de indexação automática por atribuição proposto, entende-se que na construção de bases de dados científicas deva haver o uso conjunto dos termos advindos da indexação manual e da indexação automática. Isto porque se constituem em conjuntos não disjuntos de termos atribuídos com perspectivas complementares, refletindo respectivamente a intenção dos autores e o texto dos documentos.

Corroborando o exposto, apontam-se as considerações feitas por Araújo e Santos (2019). Essas autoras realizaram um estudo comparativo entre a indexação automática e a indexação semiautomática, possibilitada pelo *software* SISA, cujos resultados foram confrontados com a indexação manual. Araújo e Santos (2019, p. 2) apontaram que "Considerando-se que o sistema pode ser uma ferramenta de auxílio do indexador humano, sendo mais eficaz na ISA. Ademais, foram constatadas necessidades de correções, revisões terminológicas e atualizações no Vocabulário da USP". As referidas autoras apontam ainda que "em outras palavras, a IA fornece as "pistas", mas a construção de uma lista de descritores mais completa dependeu da intuição, humana, e não da automatização do processo" (Araújo e Santos, 2019, p. 15).

Os resultados alcançados por Araújo e Santos (2019) reforçam que, embora a indexação automática apresente avanços, a indexação manual ainda apresenta um diferencial na atribuição de termos descritores, sobretudo na seleção de termos subtendidos na leitura de artigos, mas que não se encontram explícitos no texto. Alguns exemplos elencados pelas autoras são: "investigação criminal - não havia no texto do artigo a palavra 'criminal'; e fontes de informação – conceito não descrito, mas implícito em um dos no artigo analisado". Araújo e Santos (2019, p. 17) apontam ainda que "talvez possamos afirmar que a IM possibilita que conceitos que não integram o texto possam ser indicados como descritores. E isso a IA, pela carência semântica, não é capaz de realizar ainda".

Ainda acerca das considerações a que chegaram, Araújo e Santos (2019, p. 17) ressaltam que:

> Apesar dos pontos que precisam ser aprimorados nessa ferramenta ainda em desenvolvimento, uma importante contribuição da IA para a indexação final, a ISA, foram os termos que, sem a visualização da lista de candidatos a termos do sistema, poderiam passar despercebidas numa indexação puramente manual. Foi o que ocorreu em 12 dos 34 artigos analisados. Em todos esses casos, o SISA acrescentou descritores à ISA, contribuindo com uma melhor cobertura temática da indexação.

A indexação automática tem muito a contribuir para as propostas semiautomáticas, uma vez que agregam o processamento rápido de grandes volumes de informação à capacidade singular que o humano tem de selecionar os termos que melhor descrevem o conteúdo de um documento, bem como de perceber descritores implícitos e não cobertos pela extração automática.

Acerca do caráter promissor da associação da indexação automática à manual, Gonçalvez (2020) propôs diretrizes para a elaboração e estruturação de índice de final de livro

por meio da indexação semiautomática. O referido autor, para execução da pesquisa mencionada, baseou-se em dois pressupostos, os quais são validados ao longo de seu trabalho. Tais pressupostos são: o conhecimento sobre o processo de indexação auxilia na elaboração de índice do tipo IFL; e o segundo refere-se à ideia de que

> [...] "a indexação semiautomática é mais eficiente na elaboração de índice porque os índices elaborados manualmente geram custos altos para a produção e nos índices elaborados de forma automática há a perda de semântica e de contexto". Esse relato é validado na revisão de literatura, na seção "2.3. Análise dos trabalhos correlatos" desta dissertação e, após analisar 16 (dezesseis) publicações que abordavam o tema construção de IFL, é possível inferir que construir IFL manualmente, é inviável, se considerado o alto número de produções de livros atualmente. E o uso da utilização da indexação automática não é recomendável, pois não lida com questões de polissemias e semânticas das palavras. Assim, conclui-se que a melhor alternativa é elaborar o IFL por meio da indexação semiautomática. Assim, os dois pressupostos que guiaram essa pesquisa foram confirmados.

A indexação semiautomática faz uso dos benefícios do processamento automático da informação ao mesmo tempo em que lida com fenômenos da linguagem que são facilmente compreendidos pela indexação manual. As limitações ainda presentes nas propostas totalmente automáticas apontam para a promissora possibilidade alcançada pela indexação automática.

Corroborando o já exposto por outros autores no decorrer deste trabalho, Ferreira e Correa (2023) propuseram e aplicaram um método para obtenção de indicadores temáticos sobre descritores representativos de temas, assuntos ou palavras-chave abordados em registros bibliográficos da área de Ciência da Informação.

Ao adotar o processo de indexação automática por atribuição, os autores identificaram os principais termos relevantes, permitindo segurança na delimitação de indicadores temáticos do *corpus* analisado. Entretanto, Ferreira e Correa (2023, p. 26) mencionam as limitações ainda presentes na indexação automática, ao cometer erros no estabelecimento automático de descritores, e evidenciarem que:

> Como limitações do método proposto, podem-se apontar: a dependência da atualização do tesauro para estudos que analisem registros publicados cinco anos após a publicação do mesmo, que pode não incluir termos consolidados mais recentemente; a dependência a um modelo de indexação automática com boa eficácia; e a possibilidade de o modelo de indexação automática cometer erros na atribuição automática de descritores aos registros bibliográficos.

Além das evidências identificadas nas pesquisas de abrangência nacional, estudos de âmbito internacional apontam para a viabilidade da indexação semiautomática como estratégia promissora. O SISA apresenta-se como um exemplo de proposta semiautomática em âmbito internacional. Proposto pelo professor Izidoro Gil-Leiva na Espanha, este sistema foi desenvolvido inicialmente para a área de Biblioteconomia e Documentação que, segundo Silva e Correa (2020, p. 3), "[...] indexa automaticamente seguindo um vocabulário controlado e regras predeterminadas de indexação com base na frequência e posição dos termos". Silva e Corrêa (2020), Araújo e Santos (2019), Simões (2021), Narukawa, Gil Leiva e Fujita (2009), Silva (2020) são alguns dos autores que desenvolveram estudos que direta e indiretamente estudaram o SISA.

O referido sistema é bastante utilizado na área da Ciência da Informação, e os experimentos envolvendo tal *software* apontam resultados satisfatórios. O sistema supracitado ainda se destaca, quando comparado a outras propostas, por realizar a indexação automática por atribuição.

A esse respeito, Moreiro González (2004) ressalta que a indexação automática não é equivalente à manual, contudo a indexação automática surge para dar conta do volume de informação, mesmo sabendo que ela não dará conta de aspectos que só a indexação manual dá. Complementando essa percepção, Borges e Lima (2015, p. 50) esclarecem que:

> Embora a indexação automática possa não apresentar resultados totalmente satisfatórios, suas soluções podem contribuir para significativas melhoras no processo de indexação manual. Soluções estas que almejam realizar automaticamente a extração inicial de termos (palavras ou expressões) do documento indexado, deixando para o profissional o trabalho de selecionar aqueles mais adequados para representar seu conteúdo.

Embora seja promissora, a indexação automática ainda não consegue executar a representação temática da informação tal qual o indexador humano. Devido às limitações próprias de sua natureza, a indexação automática apresenta inconvenientes. A percepção humana de realizar a representação temática ainda é um diferencial da indexação manual, a qual é tida, em muitos estudos, como parâmetro para a avaliação de propostas automáticas.

Vejam-se algumas vantagens e desvantagens elencadas por alguns pesquisadores da área sobre a indexação automática, esta, conforme já mencionada, sendo compreendida como totalmente realizada pela máquina capacitada por meio de *softwares* específicos para este fim. Nesse contexto específico, Lapa e Correa (2014, p. 60) compreendem que a indexação automática pode ser definida "[...] como um conjunto de operações, basicamente matemáticas,

linguísticas, de programação, destinadas a selecionar termos como elementos descritivos de um documento pelo processamento de seu conteúdo".

Expõem-se, no Quadro 8, algumas vantagens e desvantagens elencadas por distintos autores:

Quadro 8 – Vantagens e Desvantagens da Indexação Automática para Gil Leiva (2008, p. 320)

ARGUMENTOS A FAVOR DA AUTOMAÇÃO DA INDEXAÇÃO	ARGUMENTOS CONTRA A AUTOMAÇÃO DA INDEXAÇÃO
1. Lentidão, subjetividade e alto custo da indexação manual.	1. Incapacidade de os sistemas automáticos de indexação reconhecerem diferentes significados em diferentes contextos, relacionarem e selecionarem conceitos implícitos dos documentos;
2. Diminuição de erros com a automatização da indexação automatizada, repercutindo positivamente na recuperação da informação.	2. Capacidade de reconhecimento limitada à identificação de palavras e não de conceitos. (Deve-se buscar a captação de terminologias dos textos, porque esta cumpre a função representativa, cognitiva e comunicativa que apresentam os conceitos e, portanto, o conhecimento).
3. Maior precisão na indexação, possibilitando uma recuperação mais rica dos documentos.	3. Na maioria das ocasiões, a automatização da indexação restringe-se às áreas específicas do conhecimento;
	4. Impossibilidade, no estado atual da investigação, de conseguir indexação totalmente automática.

Fonte: Baseado em Gil Leiva (2008).

Se se pensar no argumento 1, levantado por Gil Leiva (2008) que aponta a subjetividade como um argumento a favor da indexação automática e, por conseguinte, contra a indexação manual, é importante ressaltar que os *softwares* de indexação automática em algum momento passam pela "subjetividade", pois são programados por seres humanos, ou seja, são capacitados para reconhecerem estruturas linguísticas definidas previamente por regras gramaticais de uma comunidade, de um grupo social.

Assim, é possível pensar em certa subjetividade também na prática automática, embora os autores que se refiram a esse conceito no confronto entre a indexação automática e a manual façam referência à capacidade de a máquina se comportar sempre da mesma maneira perante diferentes documentos em momentos distintos.

Ainda acerca dos argumentos a favor e contra a indexação automática, Ward (1996) evidencia as seguintes vantagens e desvantagens, como se pode observar no Quadro 9:

Quadro 9 – Vantagens e Desvantagens da indexação automática conforme Ward (1966)

ARGUMENTOS A FAVOR DA AUTOMAÇÃO DA INDEXAÇÃO	ARGUMENTOS CONTRA A AUTOMAÇÃO DA INDEXAÇÃO
1. Leitura instantânea de todo texto	1. Funciona somente em documentos separadamente
2. Diz-se que é mais coerente do que um indexador humano	2. Não consegue fazer relações entre os textos ou entre um texto e uma visão de mundo
3. Não é tendencioso	3. Fica amarrado ao vocabulário e à gramática usada no documento indexado
	4. Não consegue lidar com dados gráficos
	5. Não consegue lidar com línguas estrangeiras
	6. Não consegue avaliar textos
	7. Não consegue criar relações intertextuais
	8. Só consegue indexar o que está explícito, não consegue indexar o que está implícito
	9. Não é capaz de imitar o questionamento, a resposta humana a um texto, o que acrescenta valor à indexação
	10. Requer constante aprimoramento para manter-se em dia com os novos desenvolvimentos
	11. Não consegue catalogar ou classificar

Fonte: Ward (1966).

Os argumentos levantados por Ward (1996) e Gil Leiva (2008) se repetem e se complementam. Tais argumentos relacionam-se diretamente com as discussões acerca da viabilidade e do caráter promissor de propostas de indexação semiautomáticas, uma vez que conjugam os benefícios advindos das duas indexações, ora da prática automática, ora da prática manual.

Silva, Correa e Gil-Leiva (2020, p. 2) esclarecem que, na busca para contribuir com a disponibilização e recuperação da informação, "[...] a indexação automática se constitui numa ferramenta tecnológica para os profissionais indexadores, permitindo poupar tempo e trabalho na indexação, e possibilitando uma melhor uniformidade e homogeneidade com relação aos termos de indexação". Assim, a indexação automática proporciona mais agilidade no processamento temático da informação, auxiliando os indexadores humanos, sobretudo no que se refere à economia de tempo e ao processamento de grandes volumes de informação. Dedicam-se as seções seguintes à reflexão sobre o uso dos SN na prática da indexação automática.

2.3.1 O uso dos Sintagmas Nominais na Indexação automática

Os primeiros trabalhos voltados para a indexação automática baseada em frequência de ocorrência de palavras no texto surgiram na década de 50 com os trabalhos de Luhn (1957)

e Baxendale (1958). Estudos posteriores evidenciaram alguns inconvenientes presentes na indexação automática baseada nas palavras isoladas, uma vez que uma palavra isolada pode adquirir vários sentidos (fenômeno da polissemia), bem como a possibilidade de várias palavras diferentes apresentarem um mesmo sentido (fenômeno da sinonímia). Esses problemas relacionados à indexação automática baseada nas palavras isoladas surgem devido à perda de poder semântico que as palavras têm ao serem retiradas de seu contexto.

Isto posto, 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. Logo, percebe-se que os termos descritores de um documento devem ser retirados do documento sem perder sua semântica, ou seja, sem serem descontextualizados. Nessa linha de pensamento, surgiram estudos voltados para a utilização não das palavras isoladas de um documento, mas de unidades linguísticas providas de semântica específica, por exemplo, os Sintagmas Nominais.

Pinto (2000, p. 65), acerca da utilização de palavras isoladas, unitermos na indexação, ressalta que, quando se indexa utilizando essas unidades lexicais, está-se diante de uma desconstrução do discurso do autor do texto, na qual "as palavras tinham um sentido em função do contexto ditado por seu criador [...] Retiradas do seu contexto, tais palavras ou conceitos passam a significar apenas propriedades, portanto seu sentido vai mudar naturalmente". Em contrapartida, essa autora ressalta que na indexação baseada em SN ou frases, "[...] os índices serão constituídos por passagens do texto portadora de informação, neste caso pode-se ter uma representação mínima do conteúdo do documento [...]". Exemplificando o exposto pela referida autora, parece-nos lógico que o SN "recuperação da informação" sejam bem mais representativos de um documento que trate dessa temática do que o termo isolado "informação".

Dentre os precursores do estudo dos Sintagmas Nominais, pode-se citar Michel Le Guern (1991), o qual é responsável pelo desenvolvimento conceitual acerca desse recurso como unidade portadora de significado para a indexação e recuperação de informação. Nesse contexto, Kuramoto (1995) e Brito (1992) introduziram os estudos acerca do uso dos SN na indexação automática. Sendo Kuramoto (1999) orientado em seu doutorado pelo próprio Michel Le Guern (Nascimento, 2015). Para Kuramoto (2002, p. 6), "um sintagma nominal é a menor unidade do discurso portadora de informação". Um SN pode ser tanto uma palavra isolada como também um conjunto de palavras que possui semântica e sintaxe.

Nesse contexto, autores como Kuramoto (1995; 2002) e Souza (2005; 2006) observaram que nem todos os SN presentes no documento são passíveis de serem os seus

descritores, evidenciando assim a necessidade de estudos voltados para a reflexão desses problemas ligados à indexação automática por meio de SN. Isso pode ser visualizado ao se comparar os sintagmas: 1º *Recuperação da informação* e 2º *o presente trabalho*, é evidente que, apesar de possuírem os elementos estruturais que compõem um sintagma nominal, o segundo sintagma pouco contribui para a representação temática do conteúdo de um documento que trate, por exemplo, de organização da informação.

Nesse contexto, Souza e Raghavan (2014), Nascimento (2015), Corrêa e Bazílio (2017), Nascimento e Corrêa (2018, 2019), entre outros desenvolveram estudos relacionados diretamente com a temática em questão, ora voltando-se para os *softwares*, ora para os critérios de seleção dos SN, ou para a normalização dos SN extraídos. Nesse aspecto, Corrêa e Celerino (2019) desenvolveram estudos acerca da normalização dos SN, com vistas à canonização dos sintagmas, procurando aumentar o valor dos SN como descritores. Esses autores propuseram um método de normalização de SN em termos canônicos, por meio de critérios que controlam o vocabulário, diminuindo a dispersão existente nos SN. Isto posto, é notória a necessidade de mais estudos acerca da representação temática por meio de sintagmas nominais. Logo, faz-se necessário antes que se aprofunde no estudo acerca dos SN, evidenciando suas características e estruturas, as quais serão discutidas nas subseções seguintes.

O uso de sintagmas nominais, em propostas de indexação semiautomática e totalmente automática, está diretamente ligado à semântica contextualizada, ou seja, aos significados de cada sintagma conforme o contexto em que está inserido, havendo, assim, uma preocupação maior com semântica do texto. Nessa seara, a linguística e a semântica se mostram pertinentes para os estudos de indexação que usem sintagmas nominais. Conforme ressalta Trujillo (2012, p. 2), "há dois ramos importantes da linguística que tratam diretamente das palavras: a etimologia que é o estudo da origem das palavras, e a lexicologia que é o estudo do significado das palavras".

Lyons (1997, p. 16) aponta que "[...] hay vários tipos de semântica claramente diferentes, cada uno con un tema u orientación disciplinaria propia: de enfoque linguístico, filosófico, antropológico, psicológico, literário, etc.". Não é de interesse aqui exaurir os estudos acerca das concepções de semântica, mas de evidenciar o papel do sentido e do significado dos sintagmas para fins de representação e recuperação da informação. Logo, interessa-nos as contribuições da semântica e da sintaxe para a indexação semiautomática.

Para Katz (1982), a semântica dedica-se ao estudo do significado linguístico. Sacconi (2020, p. 407) complementa dizendo que a semântica "[...] é o estudo da significação das

palavras e das suas mudanças de significação, através do tempo ou em determinada época. Assim, a semântica pode ser sincrônica e diacrônica". A semântica sincrônica compreende a significação das palavras e a linguagem figurada.

Reside na semântica sincrônica a contribuição da semântica para os estudos de indexação e representação da informação, ou seja, os significados assumidos pelos signos linguísticos conforme os contextos em que estão inseridos, em nosso caso específico, os significados contextualizados dos sintagmas nominais extraídos dos textos. Os Sintagmas Nominais constituem unidades com maior densidade informacional e estão, quando comparadas às palavras isoladas de um documento, bem mais relacionados ao contexto semântico do documento (Souza, 2006). Le Guerne e Bouché (*apud* Kuramoto, 1999) apontam o sintagma nominal como a menor unidade de informação contida em um texto.

Diretamente ligado à semântica, está a sintaxe que se dedica ao estudo das relações entre as unidades lexicais presentes nos textos. A sintaxe é entendida como a parte da gramática "[...] que se preocupa com os padrões estruturais dos enunciados e com as relações recíprocas dos termos nas frases e das frases no discurso, enfim, com todas as relações que ocorrem entre as unidades linguísticas no eixo sintagmático" (Sautchuk, p. 35, 2004).

Dessa forma, a sintaxe trata das relações existentes entre os termos na construção de frases e sintagmas. Logo, à indexação semiautomática que faça uso de SN, torna-se essencial que o sistema seja capacitado para realizar análises sintáticas, obtendo assim, as estruturas semanticamente mais informativas.

Borges (2009) ainda ressalta a peculiaridade das sentenças que assumem significados convencionados. Um exemplo exposto pela referida autora é "bater as botas". Os significados convencionados evidenciam uma das possibilidades de dinamicidade da língua nos processos comunicativos. Para Rocha (2011), essas expressões idiomáticas são compreendidas como sintagmas representativos do caráter pragmático da língua formados por unidades lexicais, cuja significação não pode ser depreendida do significado dos elementos que as compõem. Ou seja, tais expressões devem ser analisadas no conjunto das unidades que as compõem.

Conforme Xatara (1995, p. 207), "uma expressão idiomática é um sintagma metafórico, cristalizado em um idioma pela tradição cultural, ou seja, consagrado pelo uso, pela frequência do emprego (tendo passado do individual para o social)". Essas expressões cristalizadas são bastante frequentes na linguagem coloquial (Rocha, 2011). Em textos científicos de determinadas áreas de conhecimento, tais expressões convencionadas não são frequentes, por exemplo, dificilmente encontram-se os sintagmas "lavar a roupa suja" com uma frequência significativa em artigos científicos que não tratem de estudos linguísticos.

A semântica e a sintaxe contribuem diretamente para as estratégias de indexação automática e semiautomática, permitindo que os *softwares* selecionem descritores documentais preservando os significados específicos dos contextos em que se encontram, ou seja, deixa-se de usar exclusivamente os métodos estatísticos na escolha de termos descritores em sistemas de recuperação da informação. Dessa forma, os estudos recentes de indexação automática e semiautomática conjugam métodos estatísticos e métodos que buscam se apoiar na semântica e na sintaxe de descritores documentais para fins de representação e recuperação de informação, alcançando resultados mais satisfatórios em termos de descritores com maior densidade informacional.

É possível fazer uso de vários critérios de seleção na indexação semiautomática com o intuito de selecionar os melhores descritores. Esses critérios vêm sendo bastante aplicados nos sistemas de indexação que usam as palavras isoladas como descritores documentais.

3 PRESSUPOSTOS TEÓRICOS ACERCA DOS SINTAGMAS NOMINAIS

Embora ressaltem-se aqui, em grande parte, as concepções adotadas por Perini (2010, 2016), outros autores são utilizados como forma de embasar o tema estudado, também como forma de evidenciar outros olhares para o referido tema, são eles: Lemle (1984), Othero (2008), Sautchuk (2010), Fante e Othero (2015), Souza e Silva & Koch (2009), Prim (2010), Bezerra (2015), Pestana (2017).

Estudar a estrutura interna dos sintagmas que compõem o Português Brasileiro contribui para o aprendizado dos aspectos sintáticos de análise tradicional (normativa) que envolvem os enunciados linguísticos usados nos processos comunicativos.

Acerca dos estudos sintáticos na hodiernidade, Vieira (2020, p. 16) esclarece que:

> A análise sintática de orientação tradicional – ou "sintaxe normativa tradicional", nos termos de Azeredo (2015) – centra-se: na decomposição do período em orações; na identificação e etiquetagem das funções dos constituintes da oração; e, posteriormente, no estabelecimento de regras de concordância, regência e colocação entre os constituintes etiquetados.

Destarte, os estudos sintáticos, no contexto da gramática normativa, baseiam-se na análise da oração e das funções que cada constituinte exerce dentro delas (sujeito, predicado, complementos etc.). Compreender o funcionamento das estruturas sintagmáticas que se encontram de forma intermediária entre os níveis das palavras e das frases contribui para o aprendizado do português de forma mais clara, compreendendo, assim, as estruturas internas que envolvem a formação dos sintagmas para depois compreender as funções (sujeito, complemento etc.) que esses sintagmas exercem.

Fante e Othero (2015) evidenciam claramente as definições equívocas de alguns gramáticos ao definirem, por exemplo, a classe de palavras "pronomes". É possível encontrar em gramáticas do português definições como o pronome substitui ou acompanha o substantivo, quando, na verdade, substituem sintagmas formados por mais de uma palavra, dentre as quais se encontra o substantivo. No exemplo: a indexação semiautomática se mostra promissora. Ao usamos o pronome "ela", é possível perceber que o pronome "ela" não substitui apenas o substantivo "indexação", mas todo o constituinte "a indexação semiautomática". Compreender o papel que os sintagmas exercem nas orações torna-se fundamental para o aprendizado adequado e mais claro do português, facilitando, assim, a compreensão das funções sintáticas presentes nas orações.

Além dos estudos voltados para o aprendizado do português, a compreensão dos sintagmas, sobretudo dos sintagmas nominais, os quais apresentam diferentes potenciais, quando comparados com outros sintagmas, se mostram pertinentes em outros domínios, como as propostas de indexação automática e semiautomática que se desenvolvem dentro da Organização da Informação, que, por sua vez, se encontram dentro do escopo de estudo da Ciência da Informação. No contexto da organização e recuperação da informação, área a que se relaciona esta pesquisa, evidencia-se nos parágrafos seguintes uma descrição detalhada dos sintagmas nominais, os quais se mostram recursos promissores na indicação de conteúdos de documentos em sistemas de recuperação de informação.

Antes de adentrar na aplicabilidade dos Sintagmas Nominais na indexação automática, é necessário que fundamente, por meio de uma revisão bibliográfica, esses constituintes presentes nos enunciados linguísticos do Português Brasileiro. Para tanto, expõem-se, no decorrer desta subseção, as percepções dos principais autores acerca dos Sintagmas nominais, evidenciando a concepção de cada autor no tocante à proposta de reescrita do Sintagma Nominal.

As respostas às seguintes questões: O que é o sintagma nominal? Quais as características de um Sintagma Nominal? e, por fim, como são formados(constituídos) esses elementos?, são expostas e discutidas no decorrer desta subseção.

Ao analisarmos as sentenças do português brasileiro, especialmente, rememorando os ensinamentos advindos da educação básica, percebe-se, nas explicações dos professores, a existência de dois elementos: as palavras e as frases, as quais, ao serem vinculadas por meio de diferentes estratégias, formam os textos. Contudo, entre o nível estrutural da palavra e da frase, há um nível intermediário, pouco conhecido nos estudos durante a educação básica, contudo com grande potencial semântico e funcional: o nível sintagmático.

O termo “sintagma” foi cunhado por Ferdinand de Saussure (1997) para se referir à combinação de formas mínimas em unidades linguísticas superiores. Esse relacionamento entre essas unidades mínimas da língua configuram as relações sintagmáticas. Para Dubois (1978, p. 558), as relações sintagmáticas compreendem “[...] toda relação existente entre duas ou mais unidades que aparecem efetivamente na cadeia da fala”. São três os níveis dos sintagmas: as relações entre fonemas, as relações entre os morfemas e as relações entre as palavras, este último sendo chamado de nível sintático.

Veja-se o exemplo (1), o qual evidencia as relações sintagmáticas (no nível sintático) entre as palavras, as quais apresentam certo grau de coesão.

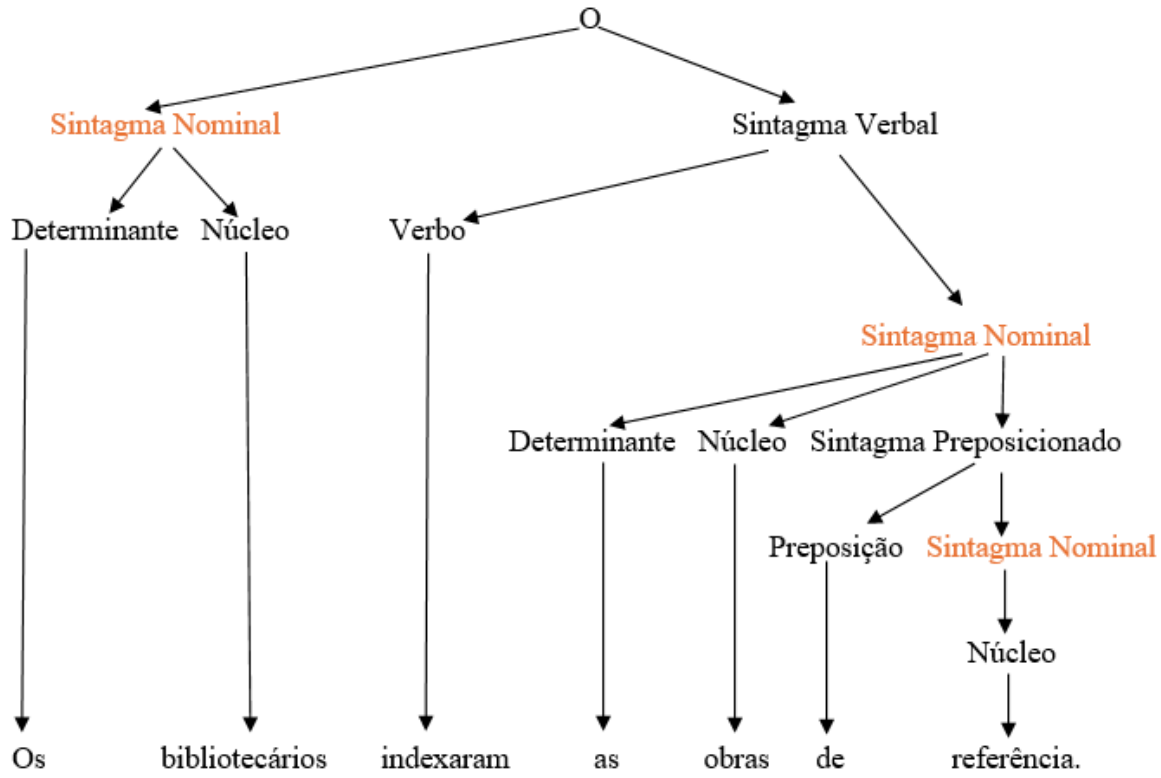
(1) Os bibliotecários indexaram as obras de referência.

Ao analisar isoladamente as palavras da sentença (1), identificam-se: “os”, artigo definido no plural, “bibliotecários” substantivo masculino no plural, “indexaram” verbo, “as” artigo definido no plural, “obras” substantivo feminino no plural, “de” preposição e “referência” substantivo feminino no singular. A nível de frase, pode-se falar que a sentença (1) é uma frase verbal com sua semântica específica na ordem SVO (sujeito + verbo + complemento).

O nível sintagmático diz respeito às estruturas (formadas por uma ou mais palavras) que se encontram dentro da oração e que mantêm entre si uma relação de dependência e de ordem (Perini, 2001). Essas estruturas chamadas de sintagmas possuem um núcleo e, em torno dela, giram os outros elementos, ou seja, outras palavras ou vocábulos. Perini (2001, p. 70) compreende esses grupos de palavras que formam sintagmas como “grupos de unidades que fazem parte de sequências maiores, mas que mostram certo grau de coesão entre eles”. Esse “certo grau de coesão”, mencionado por Perini, diz respeito ao relacionamento significativo e com sentido que existe entre o núcleo e os termos a ele associados. A classificação do sintagma dependerá de seu núcleo, o qual poderá ser um: verbo, sendo um sintagma verbal, ou um nome, sendo um sintagma nominal, ou adjetivo, sendo um sintagma adjetival, um advérbio, sendo um sintagma adverbial, ou ainda podendo ser constituído por uma preposição e um substantivo, sendo um sintagma preposicional. E dentro de um sintagma também pode existir outros sintagmas.

Analisando a Figura 4, observe-se a sentença, não no nível da palavra (morfologia), embora o esquema arbório dê conta desta parte, mas no nível dos sintagmas (sintaxe), que é nosso interesse:

Figura 4 – Esquema arbóreo



Fonte: Elaborado pelo autor (2024)

Parece pertinente que se atente para os enunciados sem esse olhar reducionista, dando ênfase apenas no nível da palavra e da frase, conforme se aprende na educação básica. Cruz (2014, p. 402, grifo nosso) evidencia bem as unidades de nossa língua ao dizer que:

> As unidades linguísticas, em suas possibilidades de combinação, seguem uma hierarquia em que morfemas formam palavras, as quais criam sintagmas, que dão origem a frases/orações, as quais formam o texto, como no esquema: morfema → palavra → sintagma → frase/oração → texto. **Dessa forma, são os sintagmas, e não as palavras, os constituintes imediatos das frases/orações**.).

No nível sintagmático, as orações da língua portuguesa apresentam no mínimo dois constituintes, conforme aprendemos nas aulas de Português: sujeito e predicado, entretanto, aqui deixam-se de lado essas nomenclaturas e passa-se a analisar como constituintes sintagmáticos, ou seja, sintagmas. Essas nomenclaturas provenientes da educação básica conduzem muitas vezes a interpretações equivocadas e a um estudo cansativo e que prioriza a memorização de regras.

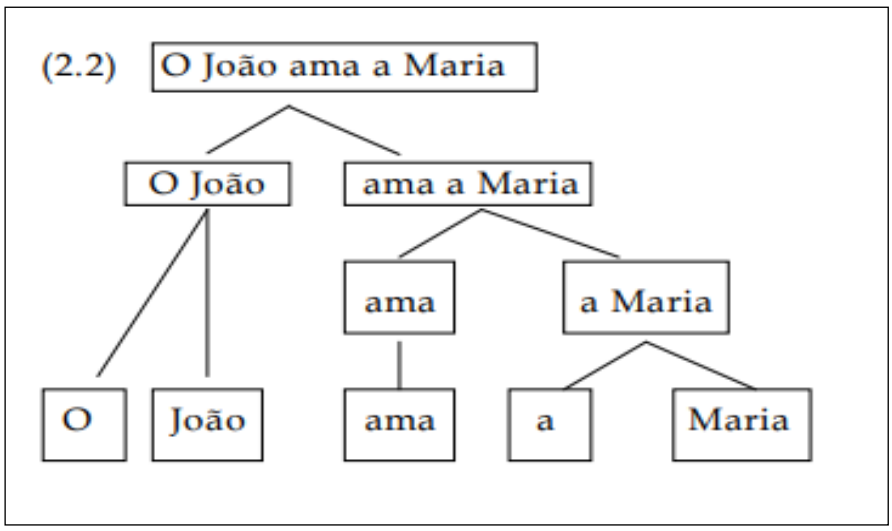
O estudo da gramática do Português Brasileiro com base no reconhecimento das estruturas sintagmáticas facilita e torna mais simples o ensino de gramática no Brasil,

conforme afirmam Fante e Othero (2015, p. 1) em estudo acerca da descrição do sintagma nominal em Português.

Esses dois constituintes básicos de toda oração são: o sintagma nominal (exerce funções típicas de substantivo, ou seja, sujeito, objeto e complemento de preposição) e o sintagma verbal (exerce função de predicado), conforme pode ser visto na sentença (1), ainda que, algumas vezes, o sintagma nominal inicial não apareça explícito na oração, que chamamos nas aulas de Português de "sujeito oculto" e "sujeito indeterminado". Ainda analisando a referida sentença, tem-se o sintagma nominal: "os bibliotecários", cujo núcleo é um nome (um substantivo), tem-se um sintagma verbal "indexaram as obras de referência", entretanto, verificam-se, dentro do sintagma verbal, mais três sintagmas, "as obras de referência" sintagma nominal e, dentro deste, "de referência" sintagma preposicional e, ainda, dentro deste, "referência" sintagma nominal.

Assim fica perceptível, a presença desses constituintes nas orações funciona como unidades intermediárias mais significativas que as palavras isoladas e que juntas constituem a frase. Othero (2008, p. 32) ressalta que "ao efetuarmos análises das sentenças em português, estaremos pressupondo que elas apresentem uma determinada organização sintática estrutural". Ou seja, não é algo aleatório, como um aglomerado linear de palavras. Othero (2008, p. 33) deixa clara essa percepção de organização sintática estrutural conforme o exemplo abaixo retirado de sua tese:

Figura 5 - Modelo de análise de sentença do Português retirado de Othero (2008, p. 32)



Fonte: Othero (2008, p. 32).

Ao se ler a oração “O João ama a Maria”, já se faz uma divisão preliminar quase que intuitivamente, como pode ser visto: “O João” e “ama a Maria”, e assim sucessivamente, evidenciando a existência de uma organização interna lógica entre a palavra e a frase, ou seja, entre as unidades sintagmáticas.

Segundo Silva e Corrêa (2015, p. 7), há duas classificações para os sintagmas, os essenciais e os facultativos. Para esses autores:

> Os essenciais são tidos como elementos básicos de uma oração, o sintagma verbal (SV), cujo núcleo é um verbo ou uma locução verbal, e o SN (sintagma nominal) que tem como núcleo o substantivo ou palavra substantivada. Esses dois tipos de sintagmas são compostos por outros sintagmas como o sintagma adjetival (SA), o qual tem como núcleo um adjetivo ou locução adjetiva, o sintagma preposicional (SP) que é composto por um núcleo chamado preposição e o sintagma adverbial (SAdv) cujo núcleo é um advérbio ou uma locução adverbial.

Sautchuk (2010), no tocante à classificação dos sintagmas, ressalta que a classificação deve partir da base de cada sintagma para sua classificação, se de natureza substantiva, adjetiva, verbal ou adverbial de seus núcleos ou da presença de preposição, encabeçando sua composição. Seguem alguns exemplos das diferentes categorias de sintagmas que encontramos no Português Brasileiro (exemplos de Othero (2008)):

- [A **Maria**] adora chocolates. Sintagma nominal.
- João é um professor [**competente**]. Sintagma adjetival.
- João nasceu [**em** Belo Horizonte]. Sintagma preposicional.
- O João [**certamente**] escreveu bastante. Sintagma adverbial.
- A Maria [**morreu**]. Sintagma verbal.

Há várias regras de formação de cada constituinte sintagmático, seja verbal, nominal, adjetival etc. Contudo, o Sintagma Nominal (SN) merece uma atenção maior devido ao fato de este sintagma ter um potencial diferenciado analogamente aos outros sintagmas, uma vez que os SN fazem referência a uma entidade do mundo real ou imaginário, quando os outros SN não o fazem. Para Liberato (1997), o SN compreende a parte de um enunciado linguístico que tem a capacidade de representar conceitos ou referentes.

Dar-se-á ênfase, a partir de agora, ao entendimento mais detalhado acerca das descrições, características e regras de formação especificamente dos sintagmas nominais, os quais constituem uma categoria a parte dentre os vários tipos de sintagmas, conforme ressalta Perini (2010), ao evidenciar propriedades específicas desses sintagmas. Segundo Perini (2010,

p. 251), o Sintagma Nominal é “um constituinte composto de uma ou mais palavras, que apresenta certas propriedades”. As propriedades a que se refere o autor são: a) o SN pode ocorrer nas funções de sujeito, objeto ou complemento de preposição; e b) semanticamente, o SN pode se referir a uma entidade do mundo (real ou imaginário). Essa entidade pode ser entendida como um objeto específico (por exemplo, minha caneta), uma classe geral (os seres humanos) ou uma abstração (a inteligência).

Um SN pode ser tanto uma palavra isolada como também um conjunto de palavras que possui semântica e sintaxe. Para Perini (2010), o núcleo do SN é a sua base referencial, ou seja, o nome que tem a capacidade referenciar algo ou alguém no mundo real ou imaginário. Esse núcleo pode aparecer sozinho ou acompanhado de outros termos, os quais são chamados por este autor de *limitadores.* Esses limitadores são elementos que podem vir antepostos ou pospostos ao núcleo, este sendo, no caso dos SN, o nome.

Souza e Silva & Koch (2009) ressaltam que o núcleo do SN pode ser representado por um nome ou pronome substantivo (aquele que substitui o substantivo), podendo aparecer acompanhado ou não de outros termos. Abaixo seguem dois exemplos de sintagmas nominais constituídos apenas de um núcleo ou acompanhados de outros elementos, os quais são chamados por Perini (2010) de limitadores (pré-nucleares e pós-nucleares) e por Souza e Silva & Koch de determinantes (antes do núcleo) e modificadores (após o núcleo).

[**Pedro**] realizou a atividade.

[Meus três **alunos** estudiosos] fizeram a atividade.

No primeiro exemplo acima, percebe-se o primeiro sintagma nominal entre colchetes formado apenas por um núcleo, no caso, “Pedro”. Já na segunda oração, tem-se o núcleo “alunos” acompanhado por três elementos, os quais, com base na nomenclatura utilizada por Perini (2010) são: Pré-nucleares = [meus]: possessivo sintético, [três]: numeral e pós-nuclear: [estudiosos]: modificador.

Sejam quais forem os elementos, eles giram em torno do núcleo e concordam nominalmente com ele. Os sintagmas, de modo geral, são classificados em *simples,* quando têm apenas um sintagma e *complexos,* quando têm outros sintagmas embutidos em um sintagma nominal.

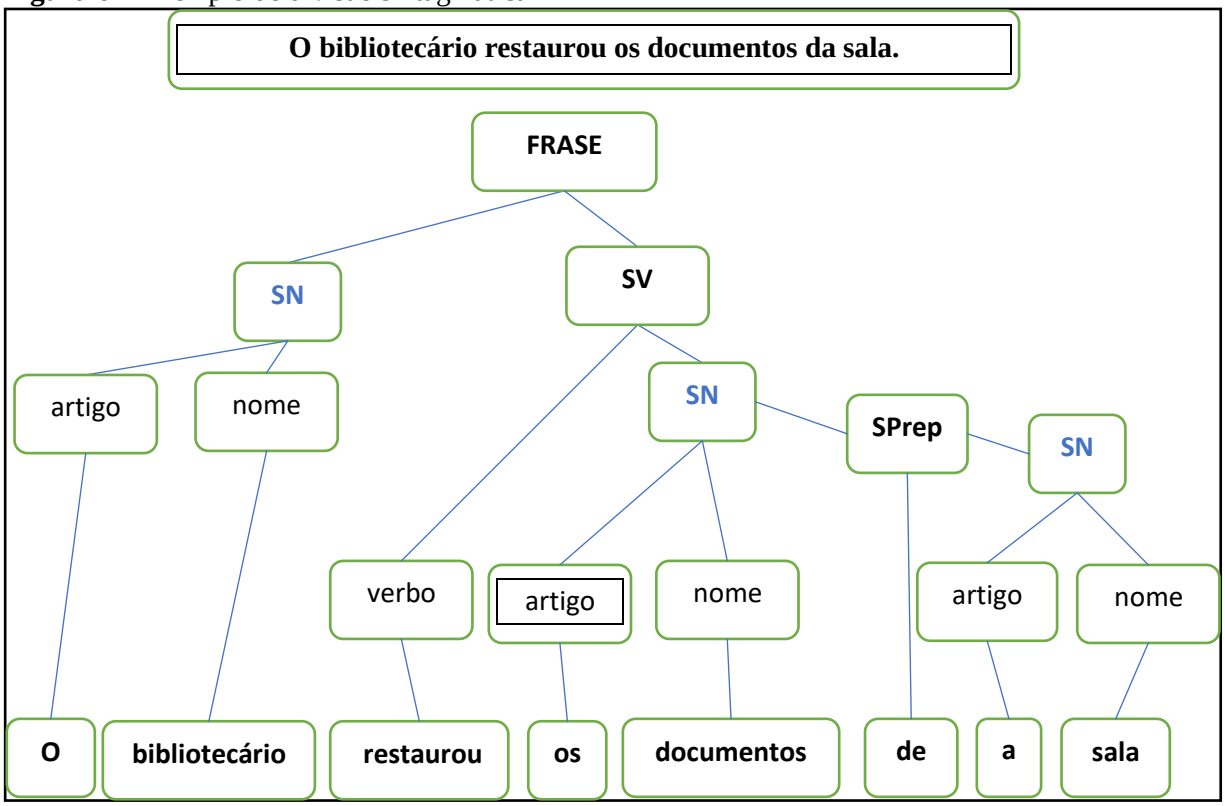
Conforme Pinheiro (2009, p. 33, grifo nosso):

> Os Sintagmas Nominais podem ter diversas configurações em termos sintáticos. Assim, para simplificar, segue um exemplo de Sintagma Nominal: • O estudo da economia da informação. Trata-se de um Sintagma Nominal complexo, pois dois outros sintagmas nominais encontram-se dentro dele: • **A economia da informação**; • **A informação**.

Os sintagmas nominais possuem a capacidade de assumir diferentes funções na sintaxe, como, sujeito, objeto direto, objeto indireto etc. Assim, podem-se ter SN aninhados dentro de outros sintagmas, integrando sintagmas verbais, bem como os sintagmas preposicionados.

Veja-se, na Figura 6, o indicador sintagmático para a sentença (1) mencionada no início dessa seção, evidenciando os sintagmas aninhados exercendo diferentes funções sintáticas:

Figura 6 – Exemplo de divisão sintagmática



Fonte: Elaborado pelo autor (2023).

É possível notar que os SN podem aparecer de forma recorrente na oração, aninhados dentro de outros sintagmas, fazendo parte de sintagmas verbais, bem como de complemento de preposição (dentro de sintagmas preposicionados), por exemplo, o SN “a sala” presente no sintagma preposicional (SPrep) “da sala”. Essa recorrência de SN pode ser vista também no SN “As características do ambiente do mundo dos negócios” extraído de Kuramoto (1995): SN1 “os negócios”, SN2 “o mundo dos negócios”, SN3 “o ambiente do mundo dos negócios” e SN4 “As características do ambiente do mundo dos negócios”. Por meio dessas relações de

encadeamento dos SN, Souza (2005) menciona a possibilidade de classificação do nível do SN com base na quantidade de outros sintagmas nele embutidos, assim, no exemplo de Kuramoto (1995), o sintagma “as características do ambiente do mundo dos negócios” seria classificado como de nível 4, visto que ele engloba três outros sintagmas.

Segue-se uma descrição detalhada dos elementos que compõem o Sintagma Nominal (SN), bem como sua estrutura e regras de formação. Esse detalhamento se faz necessário, uma vez que esta pesquisa irá desenvolver metodologia baseada em sintagmas nominais. Isto posto, é imperativo refletir de forma mais incisiva acerca dessas unidades portadoras de significado.

Embora, no decorrer desta subseção, utilizem-se as percepções de distintos autores, a próxima subseção apoia-se majoritariamente na concepção, definição e terminologia de Perini (2010) por mostrar uma descrição mais detalhada e que se aproxima da compreensão que os autores deste trabalho têm acerca desta temática. Além disso, a escolha por esse autor se deu também em razão da posição de destaque que ocupa na literatura especializada. Todavia, sempre que pertinente, são expostos os pensamentos e concepções de outros autores acerca da temática em questão.

3.1 DESCRIÇÃO DETALHADA DO SINTAGMA NOMINAL

Não obstante tenha-se, em Perini (2010 e 2016) e em Fante e Othero (2015), uma rica descrição dos SN, a presente subseção associa as percepções desses autores, tornando, assim, uma descrição mais detalhada de todos os elementos que podem constituir os SN. Além disso, traz também reflexões sobre algumas nuances que circundam alguns constituintes dos SN.

Como já mencionado, para Perini (2016), o Sintagma Nominal, quando análogo aos outros sintagmas, verbais, adjetivais etc., possui características bem peculiares. Dentre elas, a ideia de que SN é um constituinte que possui um **potencial funcional** e **referencial**, **funcional** no sentido de exercer determinadas funções, por exemplo, sujeito, objeto e complemento da preposição (sintagma proposicional), **referencial** no sentido de ser capaz de fazer referência a entidades do mundo real ou imaginário, diferentemente das outras categorias de sintagmas.

Para essa última categorização, Perini (2016, p. 356) ressalta ainda que:

> [...] o SN difere de outros tipos de sintagmas, como o sintagma verbal, que nunca se refere a uma entidade, mas a um evento (*trabalha no banco*), um

> estado (*está triste*) ou uma qualidade (*dá muito trabalho*); ou o sintagma adverbial, que denota um momento no tempo (*hoje de manhã*), ou um lugar (aqui), ou uma maneira (*com muito cuidado*) etc.

No que se refere à conceituação, o Sintagma nominal pode ser compreendido, conforme Fante e Othero (2015, p. 16), baseados em Perini (2010), Souza e Silva & Koch (2009) e Lemle (1984), como "todo agrupamento que tem por núcleo uma palavra de ordem nominal, de extensão bastante variada, e que exerce as funções de sujeito e complemente do verbo em uma sentença".

No tocante à estrutura do SN, aprendemos em sintaxe, especialmente no tópico "termos essenciais da oração", conforme as principais gramáticas, que os termos que acompanham, determinam, caracterizam o substantivo são chamados todos de "adjuntos adnominais", entretanto Perini (2010, p. 93) ressalta que esta visão é muito simplista e que esses termos, chamados genericamente de adjuntos, exercem diferentes funções sintáticas. Essas diferentes funções são exercidas por termos que se encontram ora antepostos, ora pospostos ao núcleo.

Perini (2016, p. 253, grifo do autor) compreende o SN "como uma unidade constituída internamente por um **centro de referência** e **limitadores**". Os limitadores, como o próprio nome indica, limitam o sentido do centro de referência, ou seja, do núcleo. Para este autor, "o **centro de referência** e **os limitadores** são funções semânticas, que correspondem a determinadas funções sintáticas: o centro de referência corresponde ao núcleo do SN, e os limitadores têm diversas funções, como modificador, determinante, predeterminante etc." (Perini, 2016, p. 357).

O núcleo do SN, conforme Fante e Othero (2015, p. 16), "pode ser constituído por um substantivo ou por um pronome substantivo". Ainda segundo esses autores, as palavras que podem constituir o núcleo de um SN são:

> - Substantivos, como por exemplo: menino, menina, moça, moço, João, filho, animais, bicicleta, cachorro, felicidade, amor etc.
> - Pronomes substantivos: pronomes pessoais, demonstrativos, indefinidos, interrogativos e de tratamento, como por exemplo: eu, tu, ele(s), ela(s) [...]

A identificação do centro de referência (ou do núcleo) de um sintagma parece ser uma tarefa fácil e que ocorre quase que intuitivamente em na mente ao ler os sintagmas, por exemplo. Veja-se o exemplo a seguir:

[Aqueles **bibliotecários** especialistas] catalogaram as obras.

Como pode ser visto nesse exemplo, percebe-se logo que se está falando dos seres "bibliotecários", que é o núcleo do sintagma. E os outros termos: aqueles e especialistas estão limitando o sentido do núcleo "bibliotecários", ou seja, não foi qualquer bibliotecário que fez a atividade de catalogação, mas aqueles especialistas. Perini (2016, p. 359) ressalta que "cada nominal da língua tem um potencial referencial: uns podem, outros não podem evocar uma entidade do mundo real ou imaginário (isto é, uma **coisa**)" (grifo do autor). O autor utiliza os seguintes traços para marcar o potencial referencial (de indicar coisas reais ou imaginárias) dos nominais: [+R] (de "referencial") para indicar aqueles nominais que possuem potencial de indicar uma entidade do mundo real ou imaginário e [-R] àqueles que não possuem esse potencial, já os que podem designar propriedades possuem o seguinte traço [+Q] (de "qualificativo") e os que não possuem o potencial de designar propriedades são marcados com [-Q] (Perini, 2016).

Entretanto existem nominais que podem ter as duas propriedades, ou seja, ora pode fazer referência a uma coisa, ora a uma propriedade, qualidade, como ocorre com o nome "velho" que, dependendo do contexto utilizado, pode se referir a um ser humano ou a uma característica de um ser. Têm-se, assim, para o nome "velho", os dois traços: [+R, +Q].

O sistema não é perfeito, o que, por sua vez, pode causar redundância, como no exemplo abaixo, retirado de Perini (2016, p. 361, 362):

[Um **velho palhaço**]

O exemplo provoca dupla interpretação, pois tanto o termo velho quanto palhaço possuem os traços [+R, +Q], podendo ser compreendido como "um palhaço idoso" ou "um velho que se comporta como palhaço". Já o sintagma abaixo não apresenta essa redundância:

[Um palhaço velho]

Como pode ser visto nesse último exemplo, já não se tem redundância, pois o termo "velho" não pode ser núcleo, funcionando como um modificador, introduzindo uma qualificação para o termo "palhaço". Isso mostra que fatores sintáticos, de posição e relacionamento, interferem na identificação do núcleo de um SN.

É possível ainda que se tenham SN sem os núcleos explícitos, os quais ocorrem nas seguintes situações: em contextos anafóricos e com quantificadores, os quais podem ser visualizados por meio dos exemplos abaixo, respectivamente:

Aquele bibliotecário possui [um **carro** vermelho] e [um azul].

omissão do núcleo ‘carro’ expresso no sintagma anterior

[Muitos] realizaram a indexação das obras de referência.

omissão do núcleo, o qual geralmente representa um ser humano em construções desse tipo.

Adentrem-se agora nas funções semânticas denominadas de “limitadores”, as quais podem ocorrer antes ou após o núcleo. Detenham-se agora nos que se encontram à esquerda do núcleo, chamados de pré-nucleares. Para todo o embasamento teórico acerca dos elementos constituintes dos SN, embora se tenham utilizado vários autores, evidenciam-se as contribuições de Perini, mormente sua obra mais recente, do ano de 2016.

No tocante aos elementos pré-nucleares, verificou-se, nos estudos de Perini (2005), que os pré-nucleares podem ser pré-núcleo externo e pré-núcleo interno[20], vindos após os quantificadores, contudo não se encontra essa distinção na obra de Perini (2016) que faz referência aos pré-nucleares, consoante é mostrado abaixo:

Quadro 10 - Elementos pré-nucleares conforme Perini

Predeterminantes:	são os elementos que ocorrem no início do SN, antes do determinante, como *ambos* e *todos.*
Determinantes:	São os elementos que aparecem depois do predeterminante e, na ausência dele, aparecem como primeiro elemento do SN. São eles: a, um, esse, aquele, algum, nenhum, cada, que, qual.
Quantificadores:	Quantos, tantos, poucos, muitos, vários, qualquer, certos, meio.
Possessivos Sintéticos:	Meu, seu, nosso.
Numeral:	Todos os numerais ordinais e cardinais.

Fonte: Perini (2010, p. 260).

Conforme Perini (2016, p. 365), a ordem básica desses elementos pré-nucleares é:

Predeterminante – determinante – quantificador /possessivo sintético / numeral

[20] Prim (2010) evidencia em sua dissertação os elementos pré-nucleares internos e externos, baseando-se em Perini (2007), não obstante deixa claro que essa distinção não é conclusiva por Perini.

Os elementos quantificadores, possessivo sintético e numeral apresentam certa possibilidade de mudança de posição, por isso estão expressos entre barras. Apesar de apresentarem certa mobilidade, não são totalmente livres, pois a mudança de posição em alguns casos acarreta mudança de sentido (Perini, 2016).

Citaram-se acima alguns exemplos de elementos pré-nucleares. A esses exemplos associam-se também suas flexões em gênero e número, por exemplo, quantos, quantas, seu, seus, suas etc. Embora esses sejam os elementos que podem aparecer antepostos ao núcleo de um SN, nem sempre se tem a presença de todos. Também, ainda que geralmente se apresentem na ordem exposta no Quadro 10, alguns desses elementos apresentam certa mobilidade dento do sintagma. Essa mobilidade diz respeito à capacidade de mudança de posição do elemento (limitador) sem que se perca o sentido original, conforme ver-se-á no decorrer deste trabalho. O núcleo do SN, conforme Souza e Silva & Koch (2009), pode aparecer acompanhado desses elementos, chamados, de maneira geral, por Perini (2010) de "limitadores", bem como também pode vir sozinho. Os limitadores, como já ressaltado, na acepção de Perini são elementos que aparecem à esquerda ou à direita do núcleo. Os que se encontram à esquerda são chamados de "Pré-nucleares".

Vejam-se, a partir daqui, esses elementos de forma mais detalhada, acompanhados de exemplos de orações em Português Brasileiro – PB cujas ocorrências podem ser visualizadas.

Predeterminantes: **ambos** e **todos.**

[Ambos os bibliotecários] realizaram a catalogação.

Tem-se, nesse exemplo, o sintagma nominal constituído por: Predeterminante [ambos] + determinante [os] + nome [bibliotecários]. No contexto da Gramática Tradicional, o predeterminante [ambos] expressa quantidade. A esse respeito, Bezerra (2015, p. 214) ressalta que, "[...] quando antepostos a um substantivo, empregam-se "ambos" e "ambas" seguidos geralmente de um artigo definido e mais raramente de um pronome possessivo ou demonstrativo". Esse autor informa ainda que as ocorrências "'ambos' e 'ambas' são consideradas numerais duais, pois sempre se referem a um par de coisas ou pessoas. Logo, 'ambos os dois' e 'ambas as duas' são expressões pleonásticas e devem ser evitadas na linguagem formal.

Ainda sobre o uso desse predeterminante [ambos], Fante e Othero (2015, p. 18) ressaltam que, por suas características próprias, esse elemento não aceita outro quantificador

ou numeral no mesmo sintagma nominal: ora por redundância, conforme também ressaltado por Bezerra (2015), já que o numeral "ambos" representa um par de dois, ora por produzir uma sentença contraditória, ilógica, como no exemplo que se segue, no qual o asterisco (*) indica uma sentença agramatical[21], ou seja, inadequada:

*[**Ambas** os cinco bibliotecários] realizaram a catalogação.

Logo, o predeterminante "ambos" não pode ocorrer junto a outro quantificador nem a numeral. Além do exposto, Fante e Othero (2015, p. 18) ressaltam que esse predeterminante não aceita outra posição, ou seja, ele não apresenta a possibilidade de poder aparecer em outra posição, vindo sempre anteposto ao determinante.

No tocante ao outro predeterminante elencado por Perini (2010), o elemento "todos"[22], que é análogo a "ambos", apresenta ideia de quantidade também e também não aparece junto a outro quantificador, a exemplo de "vários" e "poucos". Entretanto, aceita vir junto a outros numerais, conforme poderá ser visto nos exemplos abaixo:

*[**Todas** as várias bibliotecárias] realizaram a catalogação.

É nítida a agramaticalidade do exemplo anterior, ficando a sentença ilógica, o que não ocorre no exemplo que segue:

[Todas as seis bibliotecárias**]** realizaram a catalogação.

Além do exposto, outro ponto que diferencia o predeterminante "todos" do "ambos" é o fato de que aquele pode mudar de posição dentro do sintagma, o que não ocorre com o "ambos". Veja-se:

[**Todas** as bibliotecárias] realizaram a catalogação.

[As bibliotecárias **todas**] realizaram a catalogação.

Perini (2016, p. 368) ressalta ainda que o "todos" pode aparecer deslocado, após o verbo, nesse caso, não sendo predeterminante, mas um modificador, conforme o exemplo: [Os macacos] fugiram **todos** do zoológico. Percebe-se aqui que esse entendimento talvez não se aplique a toda construção, por exemplo, na construção [as bibliotecárias] realizaram **todas** a catalogação, tornando-se agramatical, uma vez que deveríamos pluralizar os outros elementos.

[21] Partindo da noção da Gramática como um instrumento linguístico, descritivo e normalizador de determinada língua, pode-se falar em construções da língua classificando-as em Gramaticais ou Agramaticais. As frases são consideradas agramaticais quando violam pelo menos uma regra da gramática (Eliseu, 2008), entretanto a frase que não apresenta nenhuma anormalidade ou disparidade em confronto com as regras gramaticais é considerada como gramatical (Kenedy, 2008).

[22] É pertinente ressaltar que a respeito do elemento "todos", Perini (2005, p. 108) diz que "A função de *todos* se denominará predeterminante (PDet), um termo que não pertence à estrutura do SN ". Percebemos que Perini modifica a sua proposta, contida em seu livro de 1989. Mais recentemente, não encontramos em Perini (2016) nenhuma remissiva dos elementos "todos" como elemento externo ao SN, pelo contrário, evidencia-o como primeiro elemento do SN, sem associá-lo a 'elementos externos' ao SN.

Por fim, conforme mencionado por Bezerra (2015), percebe-se ainda que o [todos] pode aparecer acompanhado ou não de determinante, diferentemente de [ambos] que, se aparecer sem determinante, torna-se agramatical no PB, como nos exemplos abaixo:

[**Ambas** as bibliotecárias] realizaram a catalogação.

*[Ambas bibliotecárias] realizaram a catalogação.

[Todas as bibliotecárias] realizaram a catalogação.

[Todas bibliotecárias] realizaram a catalogação.

Ainda sobre o predeterminante "todos", é pertinente ressaltar que se está se referindo exclusivamente ao masculino/plural e ao feminino/plural, ou seja, "todos" e "todas", o que não ocorre com a forma no singular, "todo(a)", a qual, para a GT, apresenta algumas peculiaridades, conforme ressalta Bezerra (2015, p. 235), ao dizer que hodiernamente é empregado sem a presença do artigo com o sentido de "qualquer", diferindo dos gramáticos mais antigos que não aceitam essa acepção, admitindo apenas o emprego desse termo com o sentido de "totalidade".

À possibilidade de o "todos" e sua flexão para o feminino não ter posição fixa, consoante os exemplos anteriores, apresentando certa mobilidade dentro da estrutura do SN, Pabst (2014, p. 32) e Fante e Othero (2015, p. 20) dão o nome de **quantificador flutuante.**

Determinantes: São os elementos que aparecem depois do predeterminante e, na ausência dele, aparecem como primeiro elemento do SN. São eles: **(artigos definidos e indefinidos) o, a, os, as, um, uma, umas, (pronomes demonstrativos) esse, esses, essa, essas, aquele, aqueles, aquela, aquelas, (pronomes indefinidos) algum, alguma, alguns, algumas, nenhum, nenhum, nenhuns, nenhumas, cada** e **(pronomes relativos) que, qual, quais.**

Artigo definidos e indefinidos (o, a, os, as, um, uma, umas)

Após os predeterminantes, temos **os determinantes**, conforme visto logo acima, os quais assumem a primeira posição, na ausência dos predeterminantes (ambos(as) e todos(as)), conforme os exemplos abaixo:

[As **bibliotecárias**] realizaram as visitas dirigidas.

Pronomes demonstrativos (esse, esses, essa, essas, aquele, aqueles, aquela, aquelas)

[Aqueles **documentos**] foram restaurados.

Percebe-se a posição fixa desses determinantes no exemplo. Não é possível utilizar esses pronomes demonstrativos pospostos ao núcleo, uma vez que acarretaria agramaticalidade, como pode ser visto no exemplo abaixo:

*[**Documentos** aqueles] foram restaurados.

Pronomes indefinidos (algum, alguma, alguns, algumas, nenhum, nenhum, nenhuns, nenhumas, cada)

[Nenhum **bibliotecário**] ocupou a vaga.

Ao analisar o exemplo anterior e o que se segue, verificamos que o pronome indefinido [nenhum] é passível de mudança de posição, sem que seja alterada a semântica do sintagma e que permaneça gramatical, ou seja, lógica, consoante o exemplo abaixo:

[**Bibliotecário** nenhum] ocupou a vaga.

Nesse acima, a mudança de posição do determinante "nenhum" (pronome indefinido) não acarreta mudança de sentido quando análogo ao exemplo [Nenhum bibliotecário], o que, por sua vez, não ocorre, por exemplo, com o determinante "algum", o qual assume distintos significados de acordo com a mudança de posição. Vejam-se os exemplos que seguem retirados de Pestana (2017, p. 433):

[Algum **amigo**] te traiu?

[**Amigo** algum] me traiu.

No exemplo [algum amigo], percebe-se o "sentido genérico, impreciso", já no exemplo [Amigo algum] verifica-se um sentido negativo, equivalendo a nenhum, embora nas duas ocorrências os termos pertençam à mesma classe gramatical, pronome indefinido. Logo, conclui-se que alguns pronomes indefinidos aceitam a mudança de posição sem alteração de sentido, já outros, ao mudarem de posição, acarretam alteração semântica. Em suma, esse duplo comportamento dependerá do termo, no caso, do pronome indefinido.

Pronomes relativos (que, qual, quais)

[Que **documento** restaurado]

Como pôde ser visto nesses exemplos, os determinantes (artigos, pronomes demonstrativos, pronomes indefinidos e pronomes relativos) aparecem como primeiros elementos na ausência dos predeterminantes "todos" e "ambos". Note-se também que esse segundo elemento do SN, ou seja, o determinante, só pode representar uma ocorrência, sendo

ilógico aparecer, por exemplo, em um SN dois determinantes (um pronome demonstrativo + um pronome indefinido).

Na sequência de formação do SN, após o **predeterminante** e o **determinante**, têm-se os **Quantificadores**.

Quantificadores: (pronomes indefinidos) quantos, tantos, poucos, muitos, vários, qualquer, certos, meio.

[**Quantos** bibliotecários] foram convocados?

[**Os poucos** bibliotecários] foram convocados.

Nos exemplos acima, percebe-se que os quantificadores ocorrem sempre antes do núcleo, uma vez que, após o núcleo, resultariam em construções agramaticais, como no exemplo abaixo:

[Os bibliotecários poucos] foram convocados.

Fante e Othero (2015, p. 22) ressaltam que os "quantificadores podem ocupar a primeira, a segunda ou a terceira posição do SN", consoante os exemplos abaixo:

[Aqueles **poucos** bibliotecários] - quantificador na segunda posição do SN.

[**Muitos** bibliotecários] - quantificador na primeira posição do SN.

[As nossas **várias** estratégias] - quantificador na terceira posição do SN.

Percebe-se, por meio dos exemplos supracitados, certa mobilidade dos quantificadores. Vejam-se os próximos elementos que podem constituir um SN: os pronomes possessivos sintéticos.

Possessivos sintéticos: meu, minha, meus, minhas, seu, suas, nosso, nossos, nossa, nossas

Conforme ressalta Perini (2016, p. 365), os "possessivos sintéticos se opõem aos possessivos analíticos, que ocorrem sempre depois do núcleo: são eles *dele, dela, deles, delas*.". Perini (2016, p. 367) ressalta que "Os possessivos sintéticos, quando pospostos e sem artigo, têm significado genérico, ou seja, referem-se a uma classe geral, não a um ou mais indivíduos". Vejam-se os exemplos abaixo:

[Aluno **meu**] sempre alcança a aprovação.

[Meu aluno] sempre alcança a aprovação.

É possível perceber o sentido genérico do pronome possessivo do exemplo [Aluno meu], referindo-se a qualquer aluno. Em contrapartida, no exemplo [Meu aluno], percebe-se um sentido direcionado, tratando-se de um aluno específico. Entretanto, ao associar esses possessivos sintéticos aos pronomes demonstrativos (esse, essa, esses, essas, aquele, aqueles, aquela, aquelas, estes, estas), a mudança de posição do possessivo sintético não acarretará em mudança de sentido como se percebe nos exemplos a seguir. Vejam-se:

[Aquelas **minhas** funcionárias]
[Aquelas funcionárias **minhas**]

Percebem-se, nos exemplos acima, que associados aos pronomes demonstrativos, os pronomes possessivos sintéticos podem mudar de posição sem alterar o sentido conforme já ressaltado por Perini (2016) e Fante e Othero (2015).

Ainda a respeito do uso dos possessivos sintéticos, Perini (2016) diz que o determinante "o" só ocorre junto ao possessivo sintético quando este se encontra anteposto ao núcleo, uma vez que posposto resultaria em um constituinte agramatical, conforme os exemplos abaixo:

[O **meu** amigo]
*[O amigo meu]

Percebe-se que diariamente não se profere sentenças como o exemplo [O amigo meu], contudo, em termos de semântica, talvez não haja mudança de sentido, visto que a ideia de possessão, ou seja, de o amigo ser "seu, meu" permanece. Já Fante e Othero (2015) evidenciam que o uso dos artigos indefinidos ou pronomes indefinidos como determinantes de um SN obriga o pronome possessivo a assumir apenas a posição final do SN. Vejam-se:

[Uns colegas **meus**]
*[Uns **meus** colegas]

Assim, conclui-se que o "o" (artigo definido), como determinante de um SN, obriga o pronome possessivo a ficar anteposto ao núcleo, já a utilização de artigos indefinidos (um, uns, uma, umas) ou pronomes indefinidos, como determinantes do SN, obriga o pronome possessivo a ficar posposto ao núcleo do SN.

No tocante à ordem dos possessivos sintéticos e quantificadores, há de se observar que, ao associarmos os pronomes possessivos aos quantificadores, os quais foram vistos há pouco, aqueles assumem apenas a primeira ou segunda posição. Vejam-se os exemplos:

[M**inhas** poucas aulas],
[As **tuas** várias atitudes].

Assim, percebe-se que o pronome possessivo, ao associar-se a um quantificador, não pode assumir a terceira posição, senão resultaria em uma sentença agramatical, conforme o exemplo a seguir:

*[As poucas **minhas** aulas]

Numerais (cardinais e ordinais): um, dois, três, quatro etc. e primeiro, segundo terceiro etc.

Em termos de posição dos numerais, Fante e Othero (2015, p. 22) ressaltam que eles podem assumir a primeira, a segunda ou a terceira posição do SN, conforme os exemplos a seguir:

[**Três** candidatas] foram aprovadas.
[As **três** candidatas] foram aprovadas.
[As minhas **três** candidatas] foram aprovadas.

É pertinente que se ressalte que esses elementos (numerais cardinais) não são utilizados com os Determinantes (Pronomes indefinidos **algum, alguma, alguns, algumas, nenhum, nenhum, nenhuns, nenhumas, cada**), visto que, juntos, tornam a sentença agramatical. Veja-se:

*[**Algumas três** candidatas]

Já com outros determinantes, essa junção é permitida. Vejam-se exemplos com os determinantes (**artigos** e **pronomes demonstrativos**):

[**Aquelas três** candidatas]
[**As três** candidatas]

Perini (2016) ressalta que os numerais cardinais, quando antepostos ao núcleo, assumem o sentido de quantidade, todavia, quando estão pospostos, indicam ordem. Vejam-se os exemplos:

[Os **cinco** livros] foram restaurados.

[O livro **cinco**] foi restaurado.

Percebe-se, nos exemplos acima, a mudança de sentido ocasionada pela mudança de posição do numeral cardinal, conforme ressaltado por Perini (2016). No exemplo [O livro cinco], é nítido o sentido de ordem, sendo interpretado como: "o livro quinto". O referido autor evidencia ainda que algumas vezes o numeral cardinal pode ser utilizado anteposto ao núcleo e apresentar ainda esse sentido de ordem, conforme o exemplo abaixo retirado da obra do referido autor (2016, p. 367):

[O **quarenta e três** aniversário**]**

Segue-se com os elementos que podem constituir o SN, agora, os chamados "Modificadores", os quais podem aparecer antepostos ou pospostos ao núcleo.

Modificadores[23] (pré e pós-nucleares)

Enquanto os elementos que ocorrem antepostos aos núcleos são limitados, os que ocorrem após o núcleo constituem uma classe aberta de número indefinido e com formação bem variada (Perini, 2016). Além dessa característica, esses modificadores possuem uma particularidade em termos de semântica, havendo aqueles que só aparecem antepostos ao núcleo, aqueles que podem aparecer antepostos ou pospostos ao núcleo sem que essa alteração de posição resulte em mudança de sentido, aqueles que só aparecem pospostos ao núcleo e, por fim, aqueles que aparecem antepostos ou pospostos ao núcleo com alteração de sentido conforme a sua posição.

[23] Em Perini (2005), encontra-se uma referência aos modificadores, como sendo divididos em 'Modificadores internos' e 'Modificadores externos', respectivamente, (ModI) e (ModE). Segundo o autor, um sintagma que exemplifica essas duas funções é: [Um ataque cardíaco fulminante], no qual "cardíaco" é o ModI e fulminante" é o ModE. O argumento exposto pelo referido autor para diferenciar essas duas funções mencionadas é que "... (a) *cardíaco tem função diferente de fulminante, pois cardíaco não pode ocorrer nem em último lugar, nem antes de ataque, ao passo que fulminante pode ocorrer em ambas essas posições; (b) ataque e cardíaco têm igualmente funções diferentes, pois só podem ocorrer na ordem ataque cardíaco, e não o inverso; se devessem a mesma função, ambas as ordens deveriam ser possíveis; (c) ataque e fulminante também têm funções diferentes, pois ataque, mas não fulminante, pode ocorrer logo antes de cardíaco (Perini, 2005, p. 101). Perini (2005, p. 103) ressalta ainda que "A diferença de funções entre cardíaco e fulminante é confirmada pelo seguinte fato: só fulminante, e não cardíaco, é que pode ser separado do resto do SN por algum sinal de pontuação: (31) a. Um ataque cardíaco, fulminante b. Um ataque, fulminante".* No entanto, neste trabalho não se adentras nessa diferenciação, uma vez que se pauta mais em Perini (2016), o qual já não faz referência a essas nomenclaturas.

Modificadores (Adjetivos) que se encontram exclusivamente antepostos ao núcleo (Pré-nucleares)

Após os elementos pré-nucleares expostos anteriormente, há ainda a possibilidade de ter-se anteposto ao núcleo do SN mais um limitador, o "modificador", ou seja, **o sintagma adjetival – SA (adjetivo)**, lembrando também que esses modificadores podem aparecer pospostos ao núcleo. Antes de adentrar de forma mais detalhada nos modificadores que podem aparecer tanto anteposto como posposto ao núcleo, dedique-se aos que se apresentam em posição fixa anteposto ao núcleo.

Acerca desses modificadores, itens que se apresentam sempre antepostos ao núcleo, Perini (2016, p. 368) ressalta que:

> [...] embora sejam semanticamente semelhantes aos elementos pós-nucleares (isto é, exprimem qualidade ou propriedade) aparecem sempre antes do núcleo, aparentemente em virtude de uma marca inidiossincrática, sem correlato semântico. Por exemplo: *mero, pretenso, reles, suposto, parco*.

O exposto por Perini (2016), pode ser visualizado nos exemplos abaixo:

[Um **mero** vendedor de livros] me atendeu.

Percebemos que o modificador "mero" só pode se apresentar anteposto ao núcleo "vendedor", uma vez que, se mudar a posição, a sentença ficaria sem lógica, ou seja, com o sentido comprometido. Veja-se:

*[Um vendedor **mero**] me atendeu.

Alguns desses itens variam para concordar com o seu núcleo, outros permanecem invariáveis. O exemplo anterior mostra um item que se flexiona, já nos exemplos que se seguem, identifica-se um exemplo dos invariáveis, variando apenas o determinante "um/uma", artigo indefinido:

(2) [Um **baita** cargo]

(3) [Uma **baita** festa]

Acerca do uso desses itens de posição fixa, em posição exclusivamente pré-nuclear, Prim (2010), em sua dissertação voltada para a sintaxe de adjetivos nas posições pré- e pós-nominal, evidencia, de forma clara, esses elementos de posição fixa (pré-nuclear), consoante o Quadro 11:

Quadro 11 - Adjetivos restritos à posição pré-nominal em Português Brasileiro

ADJETIVO	EXEMPLO
Baita	Uma baita fila
Bastante*	Bastante gente; os bastantes exemplos
Bel	Um bel viver
Big	Um big problema
Cardinais	Os três filhos de Maria; dois meninos
Ledo	Um ledo engano
Meio	Aquele meio limão; estava a meio caminho
Mero	Um mero detalhe
Mesmo	O mesmo problema anterior
Pretenso	Os pretensos candidatos
Primeiro, segundo etc.	O primeiro filho
Prisco	(As) Priscas eras
Puta	Uma puta festa
Reles	Uma reles observação
Senhor	Uma senhora bronca
Sumo	Suma importância; o sumo sacerdote
Suposto	O suposto atentado terrorista
Último	A última vez

Fonte: Quadro 1: Prim (2010, p. não paginado), adaptado de Negrão, Müller e Nunes-Pemberton (2002) (em pesquisa no *corpus* mínimo da gramática do português falado) e Perini (2007).

Prim (2010)[24] ressalta que o adjetivo "bastante", assim como o adjetivo *reles, "*não é mais encontrado no campo pós-nominal (salvo quando modificado, no caso de reles: *uma piada muito reles*). Parece que a posição pós-nominal, no caso do adjetivo *reles*, ficou para seu suposto equivalente informal: *um homem relo*.".

Perini (2016, p. 369) evidencia que os modificadores que ocorrem antepostos ao núcleo, conforme os vistos no Quadro 11 (Prim, 2010), além de se apresentarem somente na posição pré-nuclear, são constituídos por apenas uma palavra, já os modificadores que se encontram após o núcleo possuem a possibilidade de serem compostos por uma ou mais palavras, incluindo, também, até orações introduzidas por pronomes relativos.

Modificadores (adjetivos) que podem ocorrer antepostos ou pospostos ao núcleo (Pré- ou pós-nucleares) – sem mudança de sentido.

[24] A fonte consultada não é paginada.

Apresentam-se aqui os modificadores que possuem a possibilidade de mudança de posição, podendo aparecer antes ou após o núcleo, sem que essa mudança acarrete alteração de sentido. Vejam-se os exemplos abaixo:

[Uma **excelente** profissional]

[Uma profissional **excelente**]

[O **principal** motivo]

[O motivo **principal**]

Percebe-se, nos exemplos acima, os quais foram retirados de Prim (2010), que a mudança de posição dos modificadores (adjetivos) não acarreta mudança de sentido nos sintagmas.

Modificadores (adjetivos) que só podem ocorrer pospostos ao núcleo (pós-nucleares)

Perini (2016, p. 370) evidencia que esses nominais

> [...] são de dois tipos: primeiro, há nominais invariáveis em gênero e número, muitos dos quais designam cores, como *laranja, rosa, gelo, grená*, aos quais podemos acrescentar *alerta.* Os nominais que designam cores e variam em gênero e número podem ocorrer antes e depois do núcleo: *um vestido branco, o branco vestido da noiva.*

Àqueles nominais indicativos de cor, que ficam invariáveis, mencionados por Perini (2016), podem-se acrescentar também os seguintes substantivos indicativos de cores, que ficam invariáveis: salmão e creme. Assim, temos: laranja, rosa, gelo, grená, alerta, salmão e creme. Vejam-se alguns exemplos de sintagmas com nominais (modificadores) que não variam:

[Aquelas camisas **rosa**]

[As calças **laranja**]

[Blusas **salmão**]

Enfatizando apenas os que podem ocorrer pospostos ao núcleo, Fante e Othero (2015, p. 23), corroborando a posição de Perini (2016), citam como exemplos de nominais que só podem ocorrer pospostos ao núcleo: rosa, verde, gelo, laranja etc. Os referidos autores enfatizam ainda que, além desses nominais que indicam cores, exclusivamente pós-nucleares, há outros nominais variáveis que não indicam cores e que também são encontrados apenas

após o núcleo, por exemplo: "*ruim, comum*[25]*, esnobe, macho, fêmea*" (Fante e Othero, 2015; Perini, 2016), consoante os exemplos abaixo:

[Uma biblioteca **comum**]

[Uma pessoa **esnobe**]

[Cobra **macho**]

[Cobras **fêmea**s]

[A invasão **japonesa**] Perini (2010, p. 265).

[O carnaval **brasileiro**] –Fante e Othero (2015, p. 23).

Obviamente, se se colocar os nominais destacados nos exemplos acima expostos, ter-se-ão sentenças com a semântica comprometida, ou seja, ilógicas. Vejam-se:

(4) *[O **brasileiro** carnaval]

(5) *[**Fêmeas** cobras]

Em relação aos exemplos [A invasão **japonesa**] e [O carnaval **brasileiro**], identifica-se, respectivamente, que o modificador [japonesa] expressa a ideia de agente da ação no caso da invasão, já o modificador [brasileiro] expressa ideia de proveniência ou origem. Esses sentidos promovidos pelos modificadores pospostos são chamados por Perini (2016) de "Papéis temáticos", os quais são, em outras palavras, os diferentes sentidos que os modificadores, com seus potenciais qualificativos, podem expressar.

No tocante a esses *papéis temáticos*, Perini (2016, p. 371) elenca oito papéis temáticos desempenhados por modificadores que se encontram exclusivamente após os núcleos, a saber: "Agente, Paciente, Posse, Autor, Proveniência ou origem, Classificação, Comportamento estereotipado, Qualificação extensional".

O referido autor (2016, p. 371) ressalta ainda que "isso não quer dizer que sempre que um modificador estiver posposto deve ter esse papel temático; quer dizer apenas que esse papel temático não pode ser expresso por um modificador anteposto". Têm-se os seguintes exemplos de sintagmas constituídos por modificadores exercendo os referidos papéis temáticos:

Quadro 12 – Papéis temáticos de modificadores pós-nucleares

Papel temático	***Sintagma Nominal***
AGENTE	*A invasão **japonesa** (Retirado de Perini)*
PACIENTE	*Preservação **ambiental** (Retirado de Perini)*
POSSE	*O palácio **presidencial** (Retirado de Perini)*
AUTOR	*As sonatas **mozartianas** (Retirado de Perini)*

[25] Perini (2016, p. 370) evidencia que "*Comum* aparece antes do núcleo na expressão idiomática *de comum acordo.*".

PROVENIÊNCIA OU ORIGEM	*A música* ***paraibana***
CLASSIFICAÇÃO	*Biblioteca* ***escolar***
COMPORTAMENTO ESTEREOTIPADO	*Uma professora* ***gata***
QUALIFICAÇÃO EXTENSIONAL	*Meu tio* ***bibliotecário***

Fonte: Perini (2016, p. 371-372).

No tocante a esses papéis temáticos elencados por Perini (2016), o autor menciona que o papel temático "Qualificação extensional" apresenta algumas incertezas, as quais carecem de mais pesquisas.

Modificadores (adjetivos) que podem ocorrer antepostos ou pospostos ao núcleo (Pré- ou pós-nucleares) – com mudança de sentido.

Sobre os nominais (adjetivos) que podem ocorrer antes e após o núcleo, acarretando mudança de sentido, Perini (2016, p. 370) ressalta que:

> A maioria dos nominais pode ocorrer tanto antes quanto depois do núcleo – nesse caso imediatamente antes do núcleo, ou seja, depois dos eventuais predeterminantes, determinantes, quantificadores etc. Em alguns casos há diferença nítida de significado.

Vejamos os exemplos a seguir retirados de Perini (2016) e Fante e Othero (2015, p. 23):

[Uma secretária **simples**] - Representa humildade, simplicidade.

[Uma **simples** secretária] - Sentido de "apenas", "somente".

[Um **grande** homem] - Grandeza, bom caráter.

[Um homem **grande**] – Estatura, um homem alto.

Fante e Othero (2015, p. 23), a esse respeito, ressaltam que, quando esses Sintagmas Adjetivais aparecem antepostos ao núcleo do SN, eles apresentam um sentido abstrato (com valor afetivo, subjetivo), no entanto, quando se encontram pospostos ao núcleo do SN, apresentam um sentido mais objetivo, restritivo e especificativo. É possível verificar essa percepção desses autores com os dois últimos exemplos expostos acima, por exemplo.

Corroborando o entendimento exposto de Fante e Othero (2015), Pestana (2017, p. 340) destaca que:

> Vale dizer ainda que a maioria dos adjetivos antepostos ao substantivo, senão todos, [...] têm valor subjetivo, normalmente modalizadores. Podemos dizer que eles têm um valor que beira muitas vezes a conotação. Os que vêm pospostos têm normalmente valor objetivo, denotativo; são frequentemente descritivos.

No tocante a esses nominais, há alguns que apresentam mudança radical de sentido ao mudarem de posição. Perini (2016, p. 375) evidencia que esse fenômeno ocorre com os seguintes itens: "*simples, pobre, verdadeiro, antigo, certo, semelhante, caro*".

Vejam-se os exemplos a seguir:

Foi [um **verdadeiro** furacão], Perini (2016, p. 375) – representa algo que de fato não é verdade e apenas se aproxima do que seja um furacão

[Um documento **verdadeiro**] – representa um documento exato, autêntico, fiel.

Já em relação aos nominais "velho, novo e antigo", Perini (2016) ressalta que eles podem apresentar os mesmos sentidos, estando antepostos ou pospostos ao núcleo, com o diferencial de que há a possibilidade de um outro significado disponível, evidenciando uma ambiguidade. Vejam-se os exemplos abaixo:

[Um **novo** celular]

[Um celular **novo**]

No primeiro exemplo, o sentido expresso pelo nominal "novo" é de "outro aparelho", um novo aparelho adquirido, já no segundo, tem-se o nominal "novo" expressando o sentido de um celular nunca utilizado, mas ao mesmo tempo pode expressar o mesmo sentido do exemplo [um novo celular]. É evidente que, convencionalmente, esses nominais nessas posições vêm sendo utilizados com sentidos diferentes: o primeiro com o sentido de "diferente, outro celular"; e o segundo com o sentido de um celular sem uso, ou seja, novo.

A esses exemplos, podem-se acrescentar outros com nominais que também apresentam mudança de significado ao mudarem de posição. Veja-se o Quadro 13, composto de três colunas, respectivamente: nominal (adjetivo), sintagma e sentido expresso. Ressalta-se que exemplos do Quadro 13 não têm a intenção de exaurir todos os nominais (adjetivos), que podem aparecer antes e depois dos núcleos dos SN, uma vez que não é esse o foco deste trabalho, mas apenas o de elencar exemplos encontrados nos discursos e falas do Português

Brasileiro, nos quais são encontrados nominais que acarretam mudança de sentido ao mudarem de posição.

Quadro 13 – Relação de nominais (adjetivos) que mudam de sentido ao serem modificados de posição em relação ao núcleo do SN

único	**[Uma** única **vaga]**	**apenas uma vaga, só uma vaga.**
	[Uma vaga **única]**	especial, ou seja, uma vaga valiosa.
falso	[Um bibliotecário **falso**]	um bibliotecário sem confiança.
	[Um **falso** bibliotecário]	um bibliotecário impostor.
fino	[Um **fino** livro]	grossura
	[Um livro **fino**]	qualidade, sofisticação.
mau	[Um **mau** profissional]	péssimo
	[Um profissional **mau**]	ruim, cruel
alto	[Um **alto** funcionário]	posição, nível
	[Um funcionário **alto**]	estatura
velho	[Um amigo **velho**]	idoso
	[Um **velho** amigo]	antigo
belo	[Um **belo** dia]	indeterminado, um dia qualquer
	[Um dia **belo**]	bonito
senhor	[Um **senhor** bibliotecário]	excelente, ímpar
	[Um bibliotecário **senhor**]	idoso
nobre	[Um **nobre** funcionário]	digna
	[Um funcionário **nobre**]	importante, aristocrata, de família importante
bravo	[Um **bravo** detetive]	corajoso
	[Um detetive **bravo**]	chateado, irritado

Fonte: Elaborado pelo autor (2024)

Há nominais que, além de mudarem de sentido, mudam também de classe gramatical, como pode ser visto nos exemplos abaixo:

[**Qualquer** bibliotecário] poderá se candidatar.

[Um bibliotecário **qualquer**].

Percebe-se, no primeiro exemplo, que o termo "qualquer" possui o sentido de indefinição, indeterminação e, morfologicamente, é um "pronome indefinido"; já, no segundo exemplo, o mesmo termo apresenta o sentido pejorativo, negativo, classificando-se, morfologicamente, como adjetivo.

Além dos nominais (modificadores - adjetivos), que podem ocorrer antepostos ou pospostos ao núcleo do SN, têm-se outros elementos que podem aparecer à direita do núcleo do SN, os chamados "Sintagmas Preposicionais", aos quais é dada uma atenção a seguir.

Sintagmas Preposicionais: esses sintagmas preposicionais aparecem à direta do núcleo nominal ou verbal

Os sintagmas preposicionais são formados por “uma preposição + um nome (SN)”, os quais caracterizam o primeiro SN ou completa um verbo, ou seja, é um SN caracterizando outro SN ou complementa um verbo, que se encontra à esquerda. Nesta pesquisa, o interesse é no SN que caracteriza o SN, eis por que a ênfase se limitará a esta condição. Observe-se o exemplo abaixo:

[A biblioteca **da escola**]

Com base nesse exemplo, percebe-se que “da escola” é formado por uma preposição, no caso, uma contração da preposição “de” + o artigo definido “a”= da, seguido do nome (SN) “escola”. É nítido que “da escola” caracteriza o SN [A biblioteca]. Não se está a falar de qualquer biblioteca, mas a da escola, aquela em especial. Esse sintagma pode representar várias circunstâncias, de posse, de proveniência, de qualidade etc.

A este respeito, Perini (2016, p. 377) utiliza a expressão “Modificadores expandidos”. Conforme esse autor, “Quando um modificador é expandido, isto é, contém um elemento adverbial ou um sintagma preposicionado, é colocado depois do núcleo, como no exemplo [62] Um vestido lindo de morrer”.

Passa-se agora para o último elemento que pode constituir a estrutura de um SN, chamados de “orações relativas”.

Orações relativas que funcionam como adjetivos para o núcleo do SN.

Além dos elementos explícitos anteriormente, pode-se ter ainda, na estrutura do SN, uma “sentença”, ou seja, uma oração relativa. Essa oração relativa é uma oração que se liga ao núcleo do SN por meio de um pronome relativo (que, o qual a qual, cujo, cuja etc.) e que tem valor de um adjetivo, ou seja, qualifica, caracteriza o termo antecedente, no caso, o SN. Veja-se o exemplo abaixo:

[O bibliotecário **que realiza a catalogação**]

Percebe-se, nesse exemplo, a presença de uma oração relativa, uma sentença que serve para qualificar o SN “O bibliotecário”, restringindo o sentido do núcleo “bibliotecário”. Não se está falando de qualquer bibliotecário, mas daquele em especial “que realiza a catalogação”. Essas orações relativas sempre são introduzidas por um pronome relativo. Para a Gramática Tradicional (GT) são chamadas de orações subordinadas adjetivas e podem ser classificadas em duas categorias: restritiva e explicativa. A esse respeito, Bezerra (2015, p. 478) ressalta que “as orações subordinadas adjetivas são aquelas que possuem o valor de um

adjetivo; exercem a função sintática de um adjunto adnominal de um substantivo ou pronome antecedente". Ou seja, essas orações exercem a função de um adjetivo para o SN.

As orações relativas (adjetivas) podem ser classificadas em: "restritivas" ou "explicativas". Como os próprios nomes indicam, as chamadas restritivas restringem o sentido do nome a que se referem, limitam o sentido; já as explicativas têm o intuito de dar mais informações sobre o nome, no caso, o núcleo do SN, exprimindo o sentido geral do SN. Estas últimas, as explicativas, apresentam ainda mais um diferencial quando comparadas com as restritivas, elas são sempre isoladas por vírgulas.

Veja-se essa diferença nos exemplos abaixo:

[A pessoa **que estuda muito**] alcança a aprovação. (Oração relativa restritiva)

[Minha mãe, **que é bibliotecária**,] estuda muito. (Oração relativa explicativa)

Observando esses exemplos, verifica-se que, no primeiro, está-se falando de uma pessoa específica, não foi qualquer pessoa que que alcançou aprovação, mas o que estudou muito. Percebe-se que essa oração relativa restritiva faz jus ao nome, restringindo, limitando o sentido do núcleo do SN "pessoa". Já, no segundo exemplo, identifica-se que a oração relativa explicativa exprime o sentido geral do SN "Minha mãe". Além do fato de essa oração vir isolada por vírgula, ela pode facilmente ser retirada sem prejuízo para o sentido. A oração relativa explicativa sempre se refere a um ser único, individualizado, como no exemplo, a palavra "mãe", a qual obviamente se refere a um ser individualizado.

Com base nos estudos de Perini (2010, 2016) e Fante e Othero (2015), bem como nos de outros autores que, de forma mais discreta, contribuíram para a reflexão aqui proposta, puderam-se compreender as várias possibilidades de formação do SN, bem como algumas inconsistências, as quais mostraram que os estudos acerca dos SN ainda precisam de mais aprofundamento. As reflexões até aqui expostas contribuem diretamente para o desenvolvimento desta pesquisa, a qual, conforme exposto na seção seguinte, denominada Percurso Metodológico, analisa o comportamento da indexação semiautomática por meio da extração de sintagmas nominais em Língua Portuguesa. O entendimento da constituição e do funcionamento dos SN fornece embasamento teórico para que se realize estudos práticos de indexação semiautomática por extração de sintagmas nominais.

4 PERCURSO METODOLÓGICO

Nesta seção, apresenta-se inicialmente a caracterização da pesquisa, no tocante a sua classificação quanto aos meios utilizados, bem como ao que se refere à forma de abordagem do problema em estudo. Em seguida, detalha-se o *corpus* utilizado para realizar a pesquisa experimental. Por fim, expõem-se as etapas propostas para a realização do referido experimento, o qual envolveu a indexação semiautomática de um conjunto de artigos científicos da área de Organização e Representação do Conhecimento.

4.1 CARACTERIZAÇÃO DA PESQUISA

O interesse pelas temáticas estudadas pelo Grupo de Trabalho 2 (GT 2), bem como a relevância do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) para a Ciência da Informação foram fatores que subsidiaram a escolha do *corpus* utilizado nesta pesquisa. Assim, o GT 2 "Organização e Representação do Conhecimento" do ENANCIB constitui o universo da pesquisa, o *corpus* constituiu-se pelos trabalhos apresentados no GT 2 - Edição 2022. A recentidade foi o critério que guiou a escolha dos trabalhos dessa edição, uma vez que era a edição disponível durante o momento de realização da pesquisa empírica. A referida edição abrange um conjunto de 24 trabalhos apresentados no GT 2.

Por fim, expõem-se, com base nos resultados alcançados, diretrizes para a indexação semiautomática, por meio de sintagmas nominais, para a área de Representação Temática da Informação. O estudo exploratório ou a pesquisa bibliográfica, consoante ressalta Michel (2009, p. 40), consiste na fase inicial da pesquisa, que busca "[...] o levantamento bibliográfico sobre o tema, com o propósito de identificar informações e subsídios para definição dos objetivos, determinação do problema e definição dos tópicos do referencial teórico".

Isto posto, o presente estudo, associando a pesquisa bibliográfica à pesquisa empírica (realização de experimento), propõe-se a gerar novos conhecimentos acerca da indexação semiautomática por meio de SN, refletindo sobre as possibilidades oferecidas pela indexação automática, bem como os inconvenientes presentes nessa prática, propondo, assim, diretrizes para uma indexação semiautomática por meio da seleção de sintagmas nominais relevantes de artigos da área de CI.

Com a pesquisa bibliográfica, foi possível se aprofundar com mais propriedade nos estudos da indexação automática e semiautomática por meio de SN fornecendo ao

pesquisador um embasamento teórico necessário para a compreensão das etapas que compõem a indexação manual e automática, os inconvenientes presentes nessas atividades, ora ligados aos *softwares*, ora aos elementos próprios da língua, propondo, assim, caminhos e diretrizes para a indexação semiautomática baseada na seleção dos SN mais relevantes que funcionem como descritores documentais.

Para a execução da pesquisa bibliográfica, foram realizadas buscas por trabalhos científicos no Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–CAPES e na Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (BRAPCI). A escolha por esta base se deu por esta ser uma base especializada na produção científica da Ciência da Informação, sendo, assim, direcionada aos propósitos deste trabalho, uma vez que guarda relação direta com a Ciência da Informação. Além das fontes mencionadas, realizaram-se pesquisas na Biblioteca Digital de Teses e Dissertações – BDTD, identificando dissertações e teses que se relacionaram direta e indiretamente com o propósito deste trabalho.

Com o intuito de abranger o maior número de trabalhos relacionados à temática em estudo, buscaram-se também trabalhos indexados na *Web of Science*, verificando, assim, trabalhos na literatura nacional e internacional. Considerando a especificidade do tema tratado, sobretudo ao refinar ainda mais o objeto de estudo, a indexação semiautomática, não foi definido recorte temporal dos materiais que compuseram o referencial teórico, sendo, assim, o mais exaustivo possível.

Para a fundamentação teórica desta pesquisa, bem como à aproximação com a temática estudada, buscaram-se as produções científicas existentes que se relacionavam diretamente com objeto de estudo desta tese. Dessa forma, consoante as diferentes interfaces de buscas, utilizaram-se, como filtros frequentes, os campos “título”, “autor” e “assunto”.

Foram usados os seguintes termos de busca para recuperação dos textos: “representação temática da informação”, “indexação manual”, “indexação automática”, “indexação semiautomática”, “sintagmas nominais”, “automatic indexing”, “semiautomatic indexing” e “noun phrases”, esta última sempre associada às expressões “automatic indexing” ou “semiautomatic indexing”. Alguns termos de busca foram utilizados de forma associada a outros termos, com o intuito de precisar ainda mais os documentos recuperados. Embora alguns trabalhos recuperados não contribuíssem para a presente pesquisa, a avaliação feita pela leitura do resumo e das partes significativas dos textos procurou refinar as buscas com vistas a coletar trabalhos que se relacionassem diretamente com o propósito desta tese.

A pesquisa empírica, por sua vez, se caracteriza pela observação e experimentação dos fenômenos. “É a pesquisa que busca respostas e soluções através da observação e da prática dos fenômenos, que embasam suas conclusões” (Michel, 2009, p. 42). A pesquisa empírica é aquela que manipula dados, fatos concretos, mensurando os resultados. Por meio desse processo, realizou-se um experimento que envolveu a indexação semiautomática de um conjunto de documentos (*corpus* da pesquisa), com o intuito de refletir sobre a viabilidade e o desempenho da indexação semiautomática, propondo, dessa forma, diretrizes à indexação semiautomática baseada em sintagmas nominais para a área de Representação Temática da Informação.

No tocante à forma de abordagem do problema, esta pesquisa classifica-se como quantiqualitativa, uma vez que busca e utiliza dados quantitativos e qualitativos acerca do processo de indexação manual, automática e semiautomática. O seu caráter quantitativo fica perceptível quando se trabalha com a quantificação de SN extraídos do texto do documento, com a quantificação de SN recorrentes e que possivelmente são representativos de cada documento e, por conseguinte, com a quantificação das inferências obtidas.

Já o caráter qualitativo evidencia-se no momento em que se reflete acerca dos dados quantitativos obtidos, proporcionando estabelecer relações de variáveis, permitindo, quando possível, uma melhor compreensão dos fenômenos e, também, o surgimento de possíveis soluções para problemas encontrados.

Para Michel (2009, p. 36), a pesquisa qualitativa considera que:

> [...] há uma relação dinâmica, particular, contextual e temporal entre o pesquisador e o objeto de estudo. Por isso, carece de uma interpretação dos fenômenos à luz do contexto, do tempo, dos fatos. O ambiente de vida real é a fonte direta para obtenção dos dados, e a capacidade do pesquisador de interpretar essa realidade, com isenção e lógica, baseando-se em teoria existente, é fundamental para dar significado às respostas.

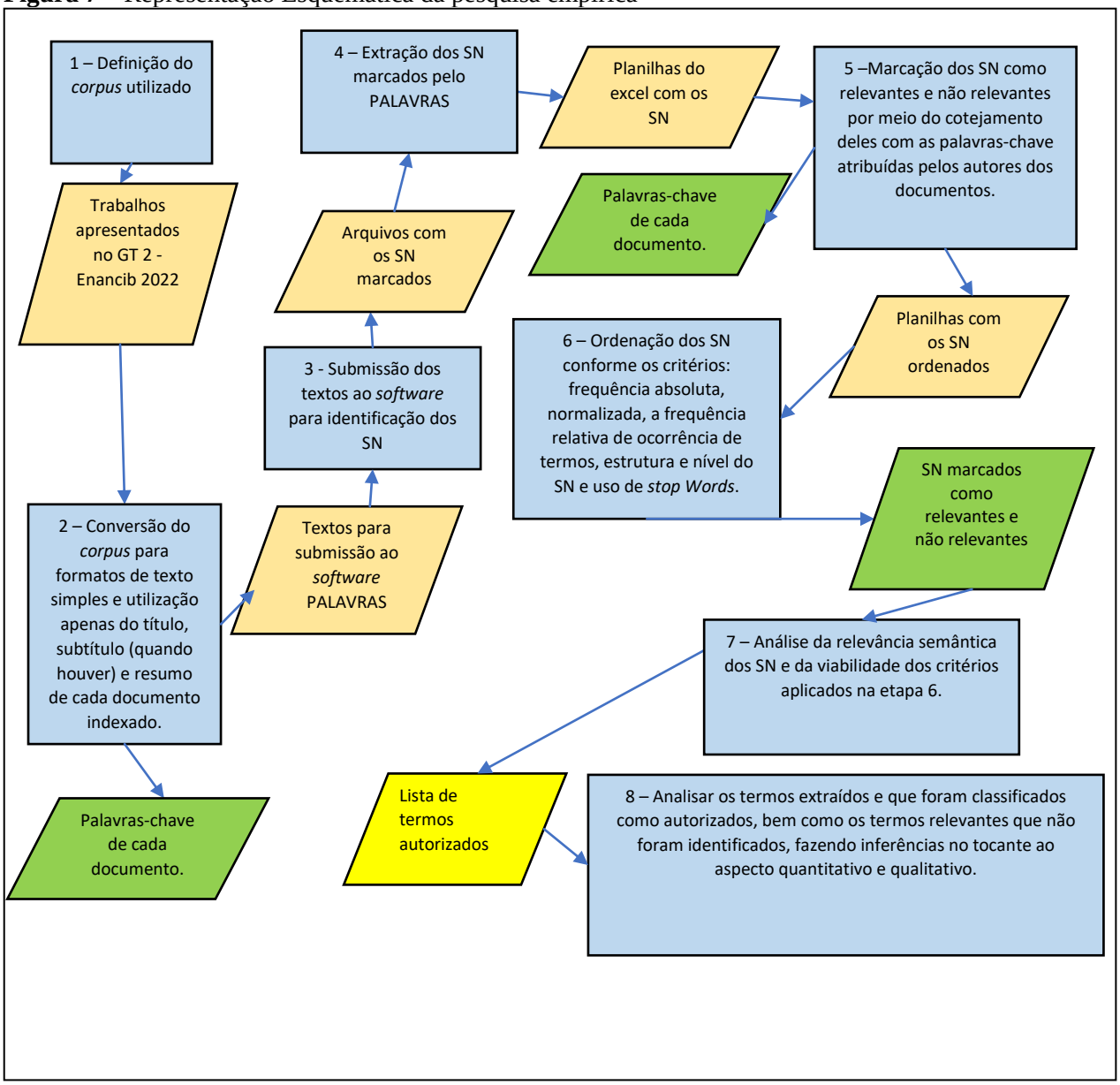
A abordagem qualitativa permite ao pesquisador fazer inferências com a argumentação lógica, fundamentada no contexto particular vivenciado (contexto, tempo e fatos), permitindo ao pesquisador a participação, a compreensão e a interpretação. Nesse sentido, para além dos aspectos quantitativos, esta tese, por meio da abordagem qualitativa, reflete, a partir das interpretações que foram feitas com base do contexto particular vivenciado, acerca das possibilidades da indexação semiautomática da informação sobre representação temática da informação, bem como do que foi evidenciado a partir do experimento realizado, que nos

permitiu ver e compreender o potencial semântico dos sintagmas nominais como unidades portadoras de informação e como elementos descritores de conteúdos informacionais.

4.2 SÍNTESE DAS ETAPAS QUE CONSTITUEM A PARTE EMPÍRICA DA PESQUISA

Expõem-se de forma sintetizada as etapas que compuseram a pesquisa empírica deste trabalho. Como forma de permitir melhor visualização de todo o experimento, na Figura 7, detalham-se essas etapas gerais, evidenciando por meio de fluxograma cada ação e resultado das referidas etapas.

Figura 7 – Representação Esquemática da pesquisa empírica



Fonte: Elaborado pelo autor (2024).

Considerando todo o arcabouço teórico tratado nesta pesquisa sobre a indexação manual, automática e semiautomática, é possível visualizar esta parte empírica como sendo uma simulação da indexação semiautomática por meio da extração de sintagmas nominais. Assim, com base na Figura 7, as Etapas 1, 2, 3 correspondem às etapas manuais de organização do documento a ser indexado e submetido ao *software* para identificação dos SN. Na sequência, as Etapas 4, 5 e 6 e seus produtos (A Forma do Paralelogramo) representam a parte automática da indexação semiautomática. As Etapas 7 e 8 são realizadas manualmente: naquela o indexador faz o refinamento dos termos selecionados automaticamente; nesta, acontece a representação de uma síntese do resultado da indexação semiautomática, que contribuirá para a reflexão acerca do desempenho da indexação semiautomática e da construção de diretrizes que guiem propostas de indexação semiautomática de forma mais eficiente.

A realização deste estudo em um contexto temático específico se justifica, uma vez que a escolha de um domínio particular é necessária, tendo em vista as características específicas de cada campo. No domínio jurídico, por exemplo, é frequente a ocorrência de numerais, para identificação das leis, artigos etc., entretanto em outros domínios essa característica já não é tão marcante. Essas particularidades de cada área de conhecimento influenciam diretamente no desempenho de propostas de indexação automática e semiautomática.

Entretanto, propostas de indexação semiautomática para os diferentes domínios podem ser alcançadas por meio da associação de diretrizes gerais às diretrizes particulares de cada área do conhecimento. Por exemplo, a utilização de numerais como elementos descritores tendem a apresentar potencial representativo no domínio do Direito, uma vez que as referências às leis são feitas por meio de numerais cardinais, todavia em outros domínios essa heurística pode não ser tão útil.

Além de instruções de caráter geral, diferentes instrumentos podem ser utilizados em metodologias de indexação semiautomática, como tesauros de domínios específicos. Tais instrumentos permitem maior controle na seleção de elementos descritores, possibilitando, assim, uma linguagem uniforme na entrada e na saída de sistemas de recuperação de informação.

4.3 DETALHAMENTO DAS ETAPAS DA PESQUISA EMPÍRICA

Ao lado do aporte teórico da pesquisa bibliográfica, realizou-se a pesquisa empírica. Assim, o estudo em questão, por meio do referido experimento, realizou a indexação semiautomática de 24 trabalhos apresentados no GT 2 – Organização e Representação do Conhecimento, no Enancib edição 2022, representando a totalidade de artigos dessa edição. Conforme já ressaltado no início da seção "Procedimentos metodológicos", a escolha por esta edição se deu por ser a mais recente até a execução empírica desta pesquisa. Considerando os fundamentos da representação temática da informação acerca da leitura documentária na indexação, sobretudo, com base na recomendação da norma NBR 12676, definiu-se como fonte para realização deste experimento o título, o subtítulo, quando existente, e o resumo de cada trabalho apresentado no GT 2 do referido Enancib. Além da NBR 12676, autores, como o UNISIST (1981), Fujita (2003), Lancaster (2004), Silva e Fujita (2004) evidenciam a relevância da leitura técnica do documento no processo de indexação, bem como a leitura de partes consideradas relevantes. Com base no exposto e tendo em vista o potencial semântico das distintas partes que constituem os textos do *corpus* desta pesquisa, realizou-se o experimento com o título, subtítulo e resumo de cada documento.

Conforme exposto no arcabouço teórico que integra este trabalho, a indexação semiautomática pressupõe a identificação de descritores documentais de forma automática e, em seguida, a seleção pelo indexador humano. Dessa forma, após a definição do *corpus*, a indexação semiautomática foi realizada com base nos elementos: títulos, subtítulo, quando existente, e resumo de cada documento, submetendo-os ao *software* PALAVRAS para a identificação dos SN que os constituem. A escolha pelo *parser* PALAVRAS (Bick, 2000) se deu por ser um analisador que, além de apresentar bom desempenho, consoante algumas pesquisas, vem sendo frequentemente utilizado por diversos pesquisadores.

Como resultado dessa etapa automática, alcançou-se uma lista de SN. Com esse resultado, iniciou-se a etapa manual da indexação, que consistiu na análise detalhada dos SN identificados automaticamente e na aplicação de critérios de seleção dos SN com potencial para funcionarem como descritores documentais de cada documento.

Antes da análise detalhada dos critérios de seleção dos SN candidatos a descritores documentais, realizou-se a comparação de cada sintagma com as palavras-chave dos autores, com o intuito de identificar os SN que eram semelhantes às palavras-chave ou que continham em sua estrutura as palavra-chave e marcados como relevantes. Esse cotejamento foi necessário para, na aplicação dos critérios, observar se o critério apresentava bom

desempenho em eliminar sintagmas que não eram marcados como relevantes. Feito isso, aplicou-se cada critério por vez a cada sintagma nominal com o intuito de levantar as taxas de revocação e precisão de cada critério de seleção.

Por fim, construíram-se diretrizes para a indexação semiautomática baseada em Sintagmas Nominais, as quais, além de servirem para o domínio estudado, podem ser adaptadas a outras áreas do conhecimento.

4.3.1 Etapa 1 – Definição do *corpus* utilizado

A **etapa 1** consistiu na definição do *corpus* submetido à indexação semiautomática. Foram utilizados os títulos, subtítulos (quando havia) e os resumos de um total de 24 trabalhos apresentados no GT 2 Organização e Representação do Conhecimento - ENANCIB 2022. Para a definição desse *corpus*, apoiou-se em Lancaster (2004), na NBR 12.676 (1992), em Lima e Boccato (2009) e em Correa e Celerino (2017; 2019), conforme justificativa exposta.

Quanto à seleção das partes utilizadas neste experimento (títulos, subtítulos e resumos), guiou-se por alguns dos elementos informativos considerados no processo de indexação manual. Do mesmo modo que o bibliotecário na indexação manual, deve-se ater às partes mais importantes do documento, tendo em vista os curtos espaços de tempo e a demanda de material a ser indexado.

Na análise de assunto da indexação manual, Lancaster (2004, p. 24) recomenda que o indexador faça “um misto de ler e ‘passar os olhos’ pelo texto”. O indexador deve priorizar as partes mais importantes do documento, as quais, conforme Lancaster (2004, p. 24), são “autor, título, resumo, sinopse e conclusões”.

A NBR 12.676 (1992), no tocante aos documentos impressos frequentes em bibliotecas e unidades de informação, como monografias, relatórios, periódicos etc., ressalta a importância de que nenhuma informação seja negligenciada, orientando que se considerem especialmente: título e subtítulo; resumo (se houver); sumário; introdução; ilustrações (diagramas, tabelas e seus títulos explicativos); palavras ou grupos de palavras em destaque; e referências. A referida norma ressalta que, no exame de documentos não impressos, tais como multimeios, utilizam-se procedimentos diferentes. Para essas situações, a norma esclarece que “A indexação, então, é geralmente feita a partir do título e/ou da sinopse”. Considerando as partes recomendadas pela NBR 12.676, bem como por Lancaster (2004), fez-se uso de três dos elementos sugeridos.

Nos estudos voltados às indexações automáticas e semiautomáticas, é bastante comum os pesquisadores utilizarem como fontes para a indexação os títulos, subtítulos e os resumos. Lima e Boccato (2009), em trabalho que objetivou avaliar o desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do Sistema Integrado de Bibliotecas/Universidade de São Paulo - SIBI/USP, nas atividades de indexação manual, automática e semiautomática, utilizou 70 resumos de teses e dissertações em seu *corpus*. Correa e Celerino (2019), em trabalho voltado à normalização de Sintagmas Nominais na indexação automática, também fizeram uso de resumos em seu *corpus*, além dos títulos de cada trabalho.

Sobre o uso dos títulos e resumos, como fontes para a indexação automática por sintagmas nominais, Celerino e Correa (2017) concluíram que

> [...] a indexação automática por sintagmas nominais do título e resumo dos artigos científicos apresentou bons resultados quanto ao nível de revocação das palavras-chave e que com o tratamento de casos especiais, como a utilização de termos estrangeiros e de caracteres especiais, os resultados podem ser melhorados.

Ademais, o trabalho de mestrado do autor desta tese também fez uso de tais fontes para a indexação automática, as quais, com aplicação de critérios específicos, mostraram-se satisfatórias. Os referidos trabalhos apoiaram a escolha dos títulos, subtítulos e resumos como *corpus* desta tese.

4.3.2 Etapa 2 – Conversão do *corpus* para formatos de texto simples e utilização apenas do título, subtítulo (quando houver) e resumo de cada documento indexado

Esta etapa envolveu a preparação do material a ser submetido ao *software*, tanto no que se refere à seleção das partes que seriam submetidas, como na conversão dos arquivos em formato legível pelo *software*. Converteram-se os textos originalmente em formato pdf para texto simples e, em seguida, separaram-se em planilhas o título, o subtítulo (quando havia), o resumo e as palavras-chave de cada resumo. Dando prosseguimento, após conversão dos textos (em formato pdf) para texto simples e a separação de cada elemento (título, subtítulo e resumo), realizou-se a submissão de cada texto ao *software* PALAVRAS, conforme descrito na subseção seguinte.

4.3.3 Etapa 3 - Submissão dos textos ao *software* para identificação dos SN

Essa etapa consistiu em utilizar o *software* para a identificação de SN. Há diversas ferramentas automáticas de identificação de SN, como o *Parser* PALAVRAS, o LX-*Parser* e o OGMA. Segundo Celerino e Corrêa (2017, p. 12), "essas ferramentas podem ser classificadas como: ferramentas de etiquetagem, ferramentas de identificação de sintagmas nominais, ferramentas de extração de sintagmas nominais, e ferramentas de seleção de sintagmas nominais".

Silva e Corrêa (2015) realizaram uma análise comparativa entre as ferramentas: *parser* PALAVRAS, o OGMA e o LX-*Parser*, no que se refere à identificação/extração de SN. O PALAVRAS ainda apresenta melhor performance pelo número menor de erros, bem como pela possibilidade de submeter um texto completo à análise do programa, ação não permitida pelo LX-*Parser*.

O *parser* PALAVRAS já vem sendo utilizado por diversos pesquisadores, como Vieira *et al*. (2000), Miorelli (2001), Souza (2005), Santos (2005), Arcoverde (2007), Maia (2008), Lopes, (2012), Silva (2014), Martins (2014), Souza e Raghavan (2014), Nascimento (2015), Celerino e Corrêa (2017), entre outros, o que demonstra, por sua vez, viabilidade no que diz respeito ao seu desempenho.

Isto posto, optou-se por utilizar o referido analisador. Conforme Nascimento (2015, p. 95),

> O PALAVRAS foi desenvolvido pela Southern University of Denmark. Esse software pode ser acessado e utilizado livremente via internet, apesar de possuir algumas limitações quando utilizado online. O PALAVRAS recebe o documento em formato textual e analisa-o, levando em consideração o caráter morfológico, sintático e semântico do texto, onde cada palavra é etiquetada em uma classe gramatical, e as orações são marcadas de acordo com as várias possibilidades de classificação sintática e semântica. Esse software analisa as orações e suas classificações, tendo em vista alcançar uma classificação exata dos elementos que compõem o texto, evitando assim ao máximo a existência de ambiguidades. Depois de submetido o texto ao software, o mesmo retorna-o com os SNs marcados, bem como com todas as palavras componentes do texto com suas classificações. A saída do texto analisado pode ser visualizada de várias formas, inclusive na forma de visualização arbórea.

O resultado da etapa três alcançou uma lista de SN identificados automaticamente dos documentos submetidos ao *parser* PALAVRAS. Esse resultado foi utilizado na etapa quatro, detalhada a seguir.

4.3.4 Etapa 4 – Extração dos SN marcados pelo PALAVRAS

Esta etapa foi completamente manual, uma vez que o PALAVRAS apenas marcava os SN identificados. Dessa forma, conforme os SN eram marcados pelo *software*, realizou-se a extração e inserção de cada SN identificado em planilhas do *excel* para cada documento submetido.

Figura 8 – Exemplo de marcação dos SN pelo PALAVRAS para extração manual

```
|-H:prp("em" &lt;sam-&gt; &lt;right&gt;)      em
|-D:g(np)
  |-D:pron(det "o" &lt;-sam&gt; &lt;artd&gt; DET F S) a
  |-H:n("confecção" F S)      confecção
  |-D:g(pp)
    |-H:prp("de" &lt;np-close&gt;)    de
    |-D:g(np)
      |-H:n("instrumento" M P)        instrumentos
      |-D:g(pp)
      | |-H:prp("de" &lt;np-close&gt;)        de
      | |-D:g(np)
      | | |-H:n("recuperação" F S §AG)        recuperação
      | | |-D:g(pp)
      | |   |-H:prp("de" &lt;sam-&gt; &lt;np-close&gt;)       de
      | |   |-D:g(np)
      | |     |-D:pron(det "o" &lt;-sam&gt; &lt;artd&gt; DET F S)     a
      | |     |-H:n("informação" F S §TP)     informação
      | |-D:pron(det "o" &lt;-sam&gt; &lt;artd&gt; DET F S)   a
```

Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

Com base nas marcações acima, têm-se os seguintes SN: “a confecção de instrumentos de recuperação da informação”, “instrumentos de recuperação da informação”, “recuperação da informação” e “a informação”. No âmbito de todo o experimento, esta etapa demandou mais tempo, uma vez que a extração de cada sintagma nominal marcado pelo PALAVRAS foi realizada de forma manual.

Durante esse procedimento, foram registradas algumas limitações do referido *software*, relacionadas ora à não identificação de SN, ora à identificação errônea. Tais situações estão diretamente vinculadas à etapa inicial de classificação morfológica das palavras, posto que a identificação adequada de SN depende da correta classificação morfológica.

4.3.5 Etapa 5 – Marcação dos SN como relevantes e não relevantes por meio do cotejamento deles com as palavras-chave atribuídas pelos autores dos documentos

Neste momento, procedeu-se a sinalização dos SN que eram semelhantes ou que continham, em suas estruturas, as palavras-chave atribuídas por cada autor do documento. Esta etapa se fez necessária exclusivamente no contexto desse experimento, considerando, assim, tais palavras como referências para a marcação dos SN considerados relevantes ou não.

Nesta etapa foi feita a comparação dos SN considerados relevantes com as palavras-chave atribuídas pelos autores como indexadores. Esse cotejamento permitiu identificar a viabilidade do uso dos SN como descritores documentais, bem como a eficácia dos critérios utilizados na seleção de melhores SN candidatos a descritores documentais.

A escolha por usar as palavras-chave como parâmetro (referência) se deu com o intuito de permitir a análise dos SN considerados relevantes. No âmbito da indexação e recuperação de informações, o uso das palavras-chave é frequente, seja para sinalizar os documentos, representando-os, seja para permitir a recuperação no momento da busca. É comum encontrar em pesquisas sobre indexação automática e semiautomática o uso das palavras-chave como termos que servem de parâmetro para comparar termos de indexação extraídos automaticamente.

Bandin e Correa (2018, p. 65), acerca das palavras-chave, esclarecem que:

> As pesquisas sobre a importância e características das palavras-chave têm abrangido vários aspectos (MIGUÉIS et al., 2013), como: o da eficiência na recuperação da informação; o uso para a extração automática de termos a partir de diferentes metodologias e algoritmos; o uso por parte dos autores e editores; a sua utilização nas etiquetagens (metatags); e a comparação com os títulos, resumos e textos integrais.

São várias as aplicações com as palavras-chave no âmbito da representação temática da informação e da recuperação de informações em unidades e sistemas de recuperação de informações. Conforme Lu *et al*. (2019, p. 415), o uso da palavra-chave ultrapassa a etapa de busca, passando a ter aplicação na indexação, recuperação de informações, bibliometria, marcação social, extração de palavras-chave, bem como no desenvolvimento de tesauros e outros Sistemas de Organização do Conhecimento (SOC).

Fujita (2020, p. 3), em ensaio acerca do uso da Linguagem natural e da Linguagem Controlada na indexação e recuperação de informações, ressalta que:

> [...] a linguagem natural, com todas as suas características presentes que impedem a precisão e revocação na recuperação, sempre foi o recurso natural economicamente disponível e, ao que tudo indica, continua sendo, assim como a indexação e o uso de vocabulários controlados por bases de dados politicamente situadas. Neste cenário, a palavra-chave é, sem dúvida, protagonista desse debate assim como as palavras do título de uma publicação.

As palavras-chave continuam exercendo papel fundamental na indexação e recuperação de informações, seja em ambiente analógico, seja em ambiente digital. Lu *et al*. (2019) e Miguéis *et al.* (2013) ratificam o papel ímpar das palavras-chave que, para esses autores, são escolhidas pelos autores dos documentos. Tais palavras são multifuncionais, dado que têm sido utilizadas em diferentes contextos (indexação, construção de tesauros, estudos bibliométricos etc.).

Embora não seja o interesse aqui discutir os intervenientes das palavras-chave nem realizar estudos comparativos entre a Linguagem Natural (LN) e a Linguagem Controlada (LC), as unidades linguísticas extraídas da LN apresentam inconvenientes, às limitações próprias da linguagem natural. Estudos, como os de Borst (2012), Fujita (2020), Lopes (2002), Silva e Lima (2015) e Tartarotti (2019), recomendam o uso concomitante da Linguagem Natural e da Linguagem Controlada, bem como o uso de estratégias para normalizar e padronizar a atribuição de palavras-chave por autores/indexadores.

As palavras-chave, atribuídas aos documentos, exercem papel ímpar no âmbito da representação e recuperação da informação (indexação, bibliometria, construção de tesauros etc.), não obstante é preciso que se dê atenção à padronização no momento de atribuição dessas palavras como forma de combater a dispersão terminológica.

Fujita e Tartarotti (2020, p. 333), acerca do papel das palavras-chave atribuídas pelo autor, ressaltam que a

> [...] relevância de palavras-chave para tais aplicações reside no fato de que é o próprio autor quem garante a representatividade ‘chave’ de seus textos. Isso é compreensível, porque o autor produziu o texto, é o especialista do tema tratado e tem domínio do conteúdo e do vocabulário utilizado. Portanto, é ideal que ele próprio atribua palavras-chave que representem o texto por ele produzido.

O autor, na condição de escritor, tem a capacidade de expressar o conteúdo tratado em seu trabalho, pois ele conhece de perto o assunto estudado. No entanto, é preciso lembrar que normalmente esse autor, no momento de atribuição das palavras-chave, não dispõe de auxílio de um profissional. Ademais, os autores, na condição de indexadores, não atentam às outras

finalidades de uso dessas unidades linguísticas no âmbito da representação e recuperação de informações.

No âmbito das pesquisas sobre indexação automática e semiautomática, frequentemente as palavras-chave são utilizadas ora como parte do processo, ora como fonte para comparação com descritores identificados automaticamente: como em Bandim e Correa (2019), propondo e avaliando um processo de indexação automática por atribuição na representação de artigos escritos em português; Souza e Raghavan (2014), apresentando uma abordagem para extrair frases-chave de textos, com base na semântica intrínseca do texto; Silva e Correa (2019), propondo percurso metodológico para a construção de *corpus* de referência em CI; dentre outros.

Com base no exposto, usaram-se as palavras-chave dos autores como parâmetro complementar à análise manual de relevância dos SN extraídos e selecionados como descritores, para avaliar os descritores identificados e selecionados automaticamente. À avaliação da indexação semiautomática, realizou-se a avaliação intrínseca quantitativa e a avaliação extrínseca mediante interconsistência, as quais são explicitadas no Quadro 4 – Formas de avaliação da indexação desta tese. A avaliação intrínseca quantitativa mede o grau de consistência da indexação automática com a indexação intelectual dos autores (palavras-chave dos artigos). Ou seja, na análise de consistência, há o cotejamento entre os SN considerados descritores e as palavras-chave atribuídas pelos autores de cada artigo.

Adicionalmente, realizou-se a avaliação extrínseca, com o cálculo das métricas de Revocação e Precisão. O índice de revocação é alcançado por meio da relação entre os termos relevantes atribuídos e o total de termos relevantes existentes para cada documento. O índice de precisão é alcançado com base na relação entre os termos relevantes atribuídos e o total de termos atribuídos para cada documento.

4.3.6 Etapa 6 - Ordenação dos SN conforme os critérios: a frequência absoluta, a frequência normalizada, a frequência relativa de ocorrência de termos, a estrutura (nível) dos sintagmas e o uso de *stop words*

Normalmente, após a seleção automática ou semiautomática de possíveis descritores documentais, aplicam-se critérios com o intuito de refinar os candidatos a descritores (palavras-chave, expressões, termos, sintagmas nominais etc.). Os estudos acerca da indexação automática, como alternativa a intervenientes presentes na indexação manual, se desenvolveram a partir da década de 50.

A aplicação de critérios na seleção automática de descritores documentais[26] buscou aumentar o desempenho e a eficácia das propostas automáticas. Borges (2009), em sua dissertação, dedicou-se ao estudo dos critérios utilizados na indexação automática. Borges e Lima (2015, p. 53-54), em artigo científico, identificaram 16 critérios utilizados para o desenvolvimento de *softwares* para indexação automática. Tais critérios estão expostos no Quadro 6 desta tese. Após uma análise detalhada e minuciosa, Borges e Lima (2015, p. 66) propõem os seguintes critérios considerados como relevantes para o desenvolvimento de *softwares* de indexação automática.

> 1. Formatação de frases-termo (Word phrase formation);
> 2. Frequência absoluta de ocorrência da palavra no texto;
> 3. Identificação de palavras (Comparação com uso de dicionário); 4. Identificação de radicais de palavras (Word stemming);
> 5. Lista de palavras proibidas / Palavras proibidas (Stop-list / stop-words);
> 6. Peso numérico;
> 7. Posição do termo no texto (Term weighting);
> 8. Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro.

A frequência absoluta de ocorrência da palavra no texto é um critério bastante frequente nas pesquisas acerca da indexação automática, bem como a lista de palavras proibidas, por exemplo. Alguns desses critérios foram estudados no contexto específico da seleção de Sintagmas Nominais por Nascimento e Correa (2018, p. 191). No âmbito da seleção de sintagmas nominais, os referidos autores identificaram como mais eficazes e promissores na seleção de sintagmas nominais "os critérios de eliminação de sintagmas nominais considerados *stop words* ou contendo pronomes no núcleo, e os critérios de seleção por posição de ocorrência, nível do sintagma nominal, inverso da frequência nos documentos e frequência de ocorrência no documento".

Considerando o exposto, utilizaram-se, nesta etapa do experimento, cinco critérios para a seleção dos sintagmas nominais: **a frequência de ocorrência, a frequência normalizada, a frequência relativa de ocorrência de termos, a estrutura (nível) dos sintagmas e o uso de *stoplist*** com o intuito de avaliar os critérios para o estabelecimento de SN com maior potencial descritivo. A utilização desses critérios apoia-se em leis e princípios

[26] Adotou-se nesta tese a nomenclatura "Descritor(es) documental(is)" para se referir a termos de indexação representativos do documento do qual foram retirados. Em Gil Leiva (1999), é possível encontrar tal nomenclatura, sendo os termos propostos na indexação automática armazenados como descritores do documento. Em Corrêa e Bazílio (2017) também encontra-se o uso da expressão "descritores documentais". De acordo com Cintra (1983, p. ???, grifo nosso), "o processo de indexação consiste na tradução de um documento em termos documentários, isto é, em **descritores**, cabeçalhos de assunto, termos-chave, que têm por função expressar o conteúdo do documento".

da bibliometria, que contribui para a reflexão acerca da relação entre a frequência de ocorrência de termos e a representatividade temática desses termos para os documentos dos quais foram extraídos. Associado aos critérios estatísticos, consideraram-se também aspectos semânticos e sintáticos dos SN, por meio da análise da estrutura e do nível de cada SN.

Tavares e Celerino (2018, p. 11), acerca dos métodos estatísticos, apontam que:

> A maioria dos softwares de indexação automática mensura a ocorrência de palavras que estão presentes no documento, e isso está diretamente ligado a algumas leis e princípios da bibliometria, como as Leis Bibliométricas de Zipf e Goffman, que são voltadas ao estudo da frequência de palavras no texto, e, de acordo com Guedes (1994), já são aplicadas no processo de indexação automática.

Ainda conforme os referidos autores, na indexação automática

> [...] é comum a utilização dessas listas ordenadas de termos, pois, como constatado por Zipf, os termos mais utilizados identificam sobre qual assunto o documento aborda, possibilitando, assim, a atribuição desses termos ao documento durante o processo de indexação (Tavares; Celerino, 2018, p. 12).

A frequência inversa de ocorrência de termos seleciona termos ou expressões que ocorrem num documento com mais frequência do que sua taxa de ocorrência na base de dados como um todo. Borges (2009, p. 60), acerca desse critério de seleção, esclarece que "a frequência com que uma palavra ocorre na base de dados como um todo é ainda mais importante que a frequência com que uma palavra ocorre num documento". Ou seja, os termos, em nosso caso os SN, que são melhores descritores, são aqueles que são imprevisíveis e raros numa coleção.

Quanto à estrutura dos SN, conforme exposto na seção "**3 PRESSUPOSTOS TEÓRICOS ACERCA DOS SINTAGMAS NOMOINAIS**", os sintagmas nominais podem aparecer aninhados em outros sintagmas maiores (exemplificado na Figura 04). Kuramoto (1995), em sua pesquisa, apresentou uma classificação de níveis dos SN. Conforme o autor, os SN classificam-se conforme a quantidade de outros SN contidos em um determinado SN. Assim, o SN que não contém outro SN em sua estrutura é considerado um SN de nível 1, por conseguinte, o SN que possui um SN de nível um em sua estrutura é considerado como SN de nível 2 e assim por diante.

Com base em Kuramoto (1995), no sintagma "As características do ambiente do mundo dos negócios" há três outros sintagmas encadeados: SN1: Os negócios; SN2: O mundo

dos negócios; SN3: O ambiente do mundo dos negócios. Souza (2005, p. 23) verifica que "através dessas relações de encadeamento, podemos classificar o nível dos sintagmas nominais pela quantidade de outros sintagmas que esses englobam, sendo que, no exemplo citado, o sintagma nominal original é de nível 4".

Em Kuramoto (1999), é possível perceber a relação da estrutura sintática dos SN com a sua relevância como descritor. Conforme Souza (2005), Kuramoto verificou que 50% dos SN únicos verificados são de nível 1ª, ou seja, possuem a estrutura simples (D + N), sendo N uma estrutura sintática genericamente considerada como um substantivo ou nome próprio e D um determinante (artigo, pronome ou numeral), composto usualmente pelas estruturas, como, determinantes: o, a, os, as, dois, três, quatro [...], esse, essa, aquele, aquela [...]. Souza (2005, p. 110) corrobora essa constatação de Kuramoto, afirmando "[...] que a estrutura sintática dos SN está relacionada à sua relevância como descritores". Sousa (2005, p. 110) aponta que "Podemos notar que essas estruturas (D + N) sempre constituem SN de nível 1ª [...] e não diferem muito em termos de densidade informacional das palavras-chave, que se diferenciam desses SN apenas pela ausência dos determinantes". Ainda conforme Souza (2005, p. 111), no contexto de sua pesquisa, verifica-se que "a densidade informacional do SN cresce com seu nível (ao menos até os de terceiro nível) [...]. A menor densidade informacional ocorre entre os SN de estrutura (D+ N)".

Considerar a estrutura e o nível do SN é fundamental na seleção de SN mais relevantes em termos de potencial informativo. Um bom exemplo que evidencia a diferença entre a densidade informacional de sintagmas nominais é o caso dos sintagmas "Sistemas de Organização do Conhecimento" e "o conhecimento", enquanto este é mais genérico, mais amplo, aquele possui mais precisão com sentido mais específico.

Com base no exposto, optou-se por considerar o nível dos SN como critério na definição dos considerados relevantes. Por fim, a utilização de *stoplist*, ou seja, de uma lista de SN menos relevantes, mas bastante frequentes em textos científicos, contribui diretamente à seleção de SN com alto poder informativo, uma vez que elimina alguns SN com pouco valor informativo, enxugando a lista de SN candidatos a descritores documentais.

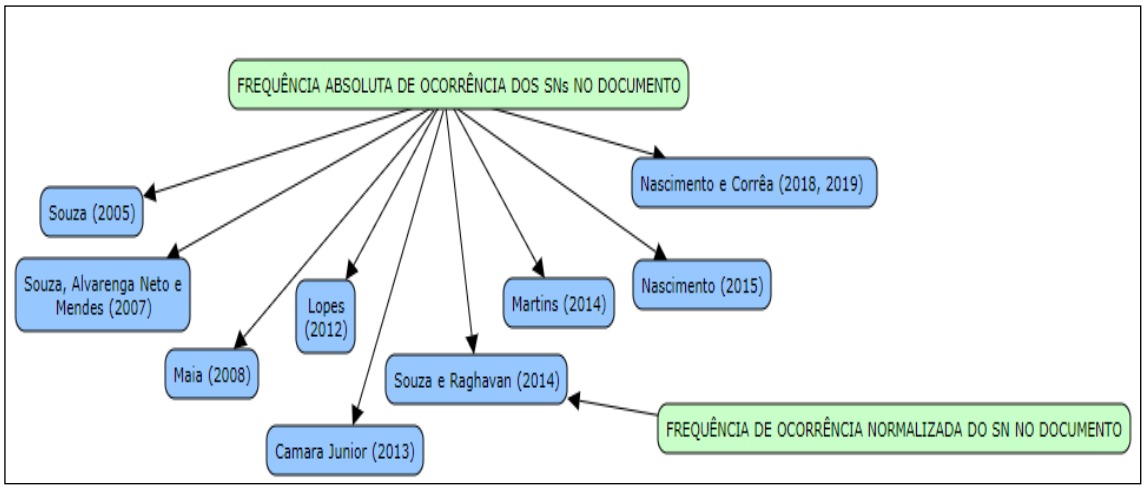
Os critérios utilizados nesta pesquisa foram aplicados em diversas pesquisas, algumas voltadas à indexação de assuntos, outras à classificação de textos etc. Esses critérios vêm apresentando resultados satisfatórios no processo de indexação automática de documentos para fins de representação e recuperação posteriormente.

A frequência de ocorrência de palavras acompanha os primeiros estudos acerca da indexação automática, mostrando-se como um critério de seleção de termos descritores. No

final da década de 1950, desenvolveram-se estratégias de escolha de termos descritores com o intuito de construir índices a partir de textos, sobretudo a partir de palavras que ocorriam nos títulos dos documentos. O *Keyword in Context* – KWIC (Palavra-chave no Contexto) – e o *Keyword out of Context* – KWOC (Palavra-chave fora do Contexto) são exemplos dessas estratégias. Para Lancaster (2004), a indexação automática tradicionalmente começa com o uso de unidades léxicas individuais, ou seja, com palavras isoladas presentes nos documentos. De modo geral, conforme ressalta Lancaster (2004), o programa computacional é preparado para extrair os termos a partir dos mesmos princípios utilizados por seres humanos. São eles: frequência da palavra dentro do texto, posição da palavra no texto (no título, nas legendas, no resumo etc.) e seu contexto.

A Figura 21 apresenta um mapa mental de pesquisas que utilizaram os critérios: Frequência absoluta de ocorrência do SN e Frequência normalizada, seja em seus experimentos, suas propostas, seja em discussões teóricas.

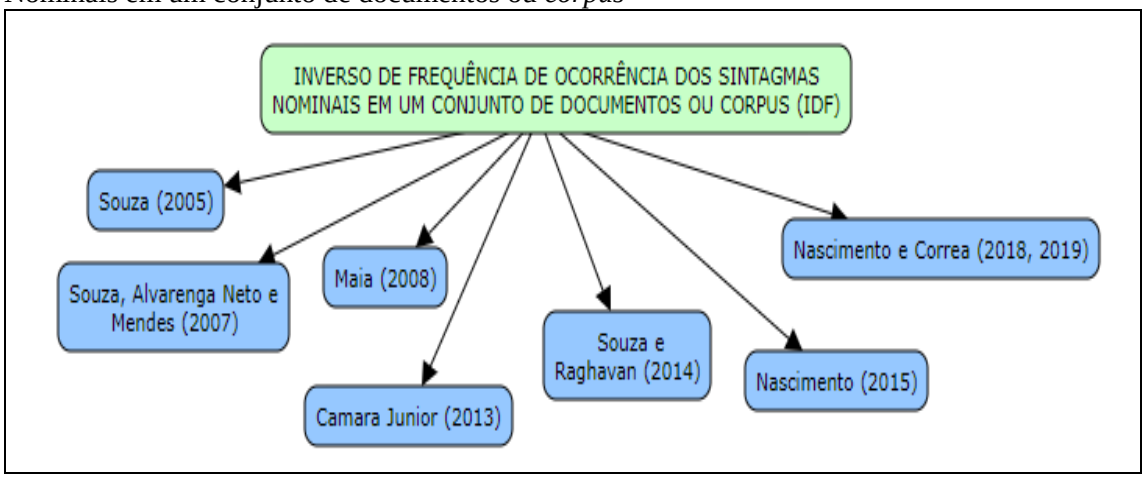
Figura 9 – Mapa mental de uso do critério "Frequência absoluta de ocorrência dos SN no documento" e "Frequência normalizada"



Fonte: Desenvolvido pelo autor (2024)

Ao lado dos critérios que analisam a frequência do SN do documento indexado, é preciso observar a frequência do termo na coleção, uma vez que um termo comum em diversos documentos tende a ser genérico e com pouco poder discriminatório. O critério "Inverso de Frequência de Ocorrência dos Sintagmas Nominais em um conjunto de documentos ou *corpus*" observa justamente isso, a popularidade de um sintagma, o que o torna muito genérico. Na figura 22, verificam-se pesquisas que fizeram uso do referido critério.

Figura 10 – Mapa mental de uso do critério "Inverso de Frequência de Ocorrência dos Sintagmas Nominais em um conjunto de documentos ou *corpus*"

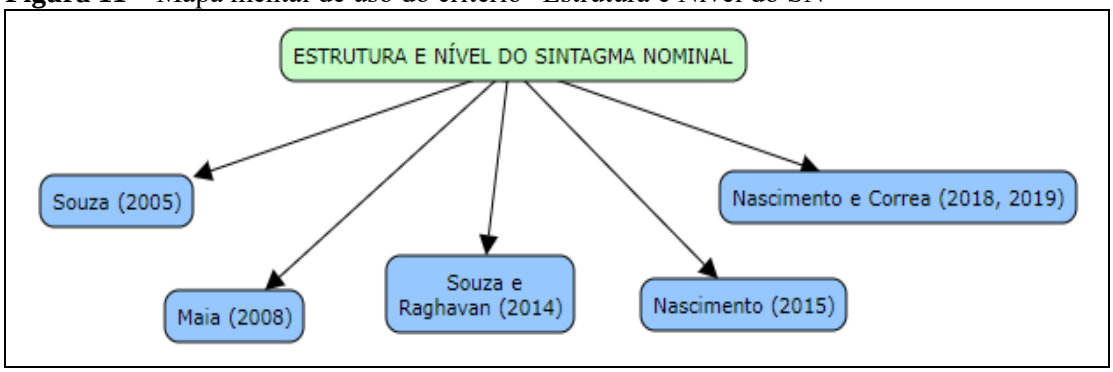


Fonte: Desenvolvido pelo autor (2024)

Estratégias exclusivamente estatísticas apresentam limitações quando se observam os aspectos semânticos do texto. Uma palavra isolada pode ter vários ou nenhum significado quando retirada de seu contexto, logo, os estudos acerca da indexação automática passaram a utilizar critérios de seleção semânticos concomitantemente com os métodos estatísticos. Os métodos que buscam analisar a semântica dos textos mostram-se promissores, uma vez que as relações semânticas são fundamentais para a construção de conceitos, bem como para escolha de termos representativos de significado. Nesse contexto, os estudos da linguística, mais especificamente da semântica e da sintaxe, contribuem incisivamente para o desenvolvimento de sistemas de indexação semiautomática que extrapolem métodos estatísticos.

Ao lado de critérios eminentemente estatísticos, há de se considerar critérios que analisem a semântica presente no texto. Nesse contexto, um critério que vem apresentando bons resultados nas pesquisas que envolvem sintagmas nominais é observar "A estrutura e nível do sintagma nominal". Na Figura 23, é possível visualizar as pesquisas que fizeram uso do critério "estrutura e nível do SN" como determinante para a seleção de SN candidatos a descritores documentais.

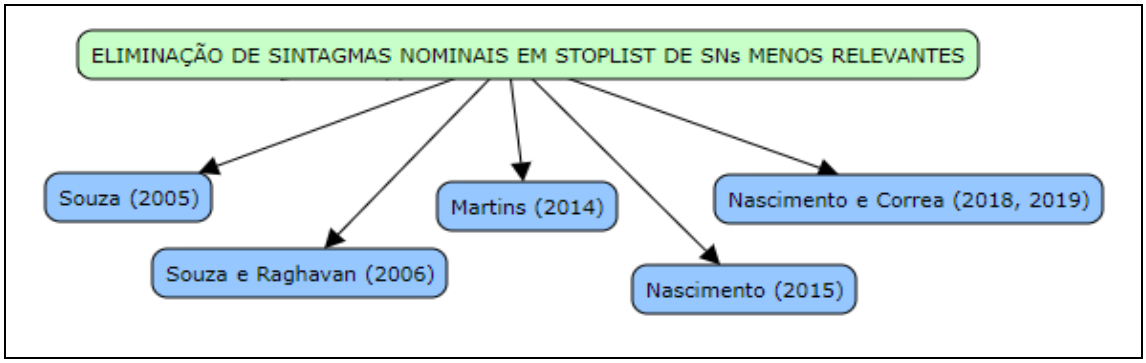
Figura 11 – Mapa mental de uso do critério "Estrutura e Nível do SN"



Fonte: Desenvolvido pelo autor (2024)

Nem todo sintagma nominal presente em um texto é representativo desse documento em termos de conteúdo, sobretudo, com termos que geralmente são comuns em textos acadêmicos, ou seja, sintagmas que constam em artigos científicos, dissertações, teses, devido à própria natureza dos documentos. Um critério que tem apresentado bom desempenho é a construção de "Listas de *Stop-words*", ou seja, palavras, sintagmas que embora sejam comuns em textos acadêmicos pouco contribuem para a representação temática dos documentos nos quais estão inseridos. Na Figura 12, expõem-se os autores que fizeram uso desses critérios no contexto de Sintagmas Nominais. No estudo de Martins (2014), fizeram parte da *stop-words*, sintagmas como "resumo, abstract, conclusões, referências, página etc.". Alguns exemplos de *stop words*, que constaram na *stop list* de Souza (2005), são "ciência, cientista(s), conhecimento(s), curso(s), documento(s) etc." É importante ressaltar que cada *stop list* deve ser considerada no contexto de cada *corpus* trabalhado ou base utilizada. Expõe-se abaixo, na afigura 24, pesquisas que fizeram uso do referido critério no contexto da seleção de Sintagmas Nominais para representação temática de documentos.

Figura 12 - Mapa mental de uso do critério "Eliminação de sintagmas nominais em *stop list* de SN menos relevantes"



Fonte: Desenvolvido pelo autor (2024)

Após a aplicação dos critérios acima expostos com o intuito de se chegar aos Sintagmas Nominais de cada documento com maior potencial representativo de cada documento, buscou-se analisar a relevância de cada SN de cada documento, bem como o desempenho dos critérios expostos nesta subseção.

4.3.7 Etapa 7 – Análise da relevância semântica dos SN e da viabilidade dos critérios aplicados na Etapa 6

Neste momento, já com os SN ordenados com base nos critérios mencionadas na subseção anterior, fez-se a análise dos SN melhor classificados. Esta análise empírica procurou verificar o comportamento dos critérios utilizados para a ordenação dos SN no tocante à seleção dos SN com maior potencial descritivo de cada documento.

4.3.8 Etapa 8 – Análise dos termos extraídos e que foram classificados como autorizados, bem como os termos relevantes que não foram identificados, fazendo inferências no tocante aos aspectos quantitativos e qualitativos dos resultados alcançados

Nesta etapa, realizou-se a análise dos SN extraídos e selecionados após aplicação dos critérios já expostos na Etapa 6, bem como os termos relevantes que não foram identificados e o resultado da comparação com as palavras-chave dos autores com o intuito de contribuir para a construção de diretrizes para a indexação semiautomática por meio de SN.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS EXPERIMENTAIS

Nesta seção, faz-se uma discussão sobre algumas etapas do experimento, as quais merecem destaque, sobretudo pelo grau de complexidade e por influenciarem, de forma mais incisiva, nos resultados alcançados pelo experimento como um todo. Dando prosseguimento, discute-se o resultado geral do experimento na seleção de SN considerados descritores documentais para fins de representação e recuperação da informação.

Análise das Etapa 3 e 4 - identificação e extração dos SN

O *software* PALAVRAS foi utilizado para a identificação das estruturas consideradas Sintagmas Nominais. Embora as Etapas 3 e 4 tenham sido realizadas de modos distintos (um manual, outra automática), foi no momento da extração (manual) que se identificaram algumas limitações do PALAVRAS, as quais são expostas nos parágrafos seguintes.

Na Figura 13, evidencia-se um exemplo de marcação que o PALAVRAS realiza na análise de textos.

Figura 13 – Modelo de análise feita pelo PALAVRAS na identificação de SN

Trecho analisado: Título e subtítulo do Artigo 05 que constituiu o *corpus* do experimento.

Sistemas de Organização do Conhecimento e a teoria da desclassificação: diálogos possíveis

```
SOURCE: Running text
1. SISTEMAS DE ORGANIZAÇÃO DE O CONHECIMENTO E A TEORIA DE A DESCLASSIFICAÇÃO:
A1
|-PRED:par
|-CJT:g(np)
| |-H:n("sistema" &lt;*&gt; M P)        SISTEMAS
| |-D:g(pp)
|   |-H:prp("de" &lt;*&gt; &lt;np-close&gt;)    DE
|   |-D:g(np)
|     |-H:n("organização" &lt;*&gt; F S §AG)    ORGANIZAÇÃO
|     |-D:g(pp)
|       |-H:prp("de" &lt;*&gt; &lt;sam-&gt; &lt;np-close&gt;)   DE
|       |-D:g(np)
|         |-D:pron(det "o" &lt;*&gt; &lt;-sam&gt; &lt;artd&gt; DET M S) O
|         |-H:n("conhecimento" &lt;*&gt; M S §TH)       CONHECIMENTO
|-CJT:g(np)
  |-D:gr*(pp)
  | |-CO:conj("e" &lt;*&gt;)    E
  |-D:pron(det "o" &lt;*&gt; &lt;artd&gt; DET F S)      A
  |-H:n("teoria" &lt;*&gt; F S) TEORIA
  |--D:g(pp)
    |-H:prp("de" &lt;*&gt; &lt;sam-&gt; &lt;np-close&gt;)       DE
    |-D:g(np)
      |-D:pron(det "o" &lt;*&gt; &lt;-sam&gt; &lt;artd&gt; DET F S)     A
      |-H:n("desclassificação" &lt;*&gt; F S §AG)       DESCLASSIFICAÇÃO
      |-:
```

Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

Como pode ser visto na Figura 13, observam-se os termos marcados com símbolos que identificam as classes morfossintáticas de cada termo. Com base no trecho acima analisado, tem-se o lexema “sistemas”, marcado pelo *software* como nome (substantivo), utilizando o símbolo “n” que indica um “nome”, o M e o P entre parênteses indicam, respectivamente, o gênero e o número, ou seja, masculino e plural, além de mostrar a forma canônica de cada palavra, no caso da palavra tida como exemplo, “sistema”, no singular. As marcações “np” sinalizam Sintagmas Nominais.

Todo o processo de submissão dos documentos ao PALAVRAS iniciou-se primeiramente com a inserção do título do ART. 1 ao *software*. Depois de extraídos os SN do título, inseriu-se o resumo e foram extraídos os SN do resumo. Esse procedimento foi feito em todo o *corpus* da pesquisa, ART. 1, ART. 2, ART 3 até ART. 24.

Com base na Figura 13, é possível perceber os seguintes sintagmas nominais identificados pelo *software*: “sistemas de organização do conhecimento”, “organização do conhecimento”, “o conhecimento”, “a teoria da desclassificação” e “a desclassificação”. Embora tenha sido feita de modo automática a identificação dos SN, no momento da extração retomava-se frequentemente à identificação, sobretudo quando se identificavam erros por parte do *software*. Tais erros são registrados mais adiante.

No âmbito de todo o experimento, a etapa de extração dos SN demandou mais tempo, uma vez que foi realizada, de forma manual, a extração de cada sintagma nominal marcado pelo PALAVRAS.

Durante esse procedimento, foram registradas algumas limitações do referido *software*, relacionadas ora à não identificação de SN, ora à identificação equivocada. Tais situações estão diretamente vinculadas à etapa inicial de classificação morfológica das palavras, posto que a identificação adequada de SN depende da correta categorização das palavras.

A Figura 14 mostra um exemplo de equívoco na classificação morfológica do termo “aportes” na análise de trecho do resumo do ART. 1.

Figura 14 – Equívocos de etiquetagem (classificação morfológica) do PALAVRAS

Trecho analisado: Parte do resumo do ART. 1

Para isso, utiliza-se dos aportes teóricos fornecidos pela Análise do Discurso, especificamente a de origem francesa, tendo como precursor Michel Pêcheux [...]

```
|-P:vp*
| |-VD:v(fin "entender" &lt;fmc&gt; &lt;*&gt; PR 3S IND VFIN §COG)      Entende-
|-S:pron(pers "se" M/F 3S/P ACC)        se
|-CJT:cl(fcl)
| |-A:g(pp)
|   |-H:prp("para" &lt;*&gt; &lt;left&gt;)      Para
|   |-D:pron(indp "isso" &lt;dem&gt; M S)       isso
|-Od:g(np)
| |-P:v(fin "utilizar" &lt;fmc&gt; PR 3S IND VFIN §AG)  utiliza-
| |-Od:pron(pers "se" &lt;refl&gt; M/F 3S ACC)  se
| |-Op:prp("de" &lt;sam-&gt; &lt;right&gt;)     de
|-CJT:x
  |-fCs:gr-(np)
  | |-D:pron(det "o" &lt;-sam&gt; &lt;artd&gt; DET M P) os
  |-P:v(fin "aportar" &lt;fmc&gt; PR 2S SUBJ VFIN §AG)  aportes
  |--fCs:g(np)
    |-H:n("teórico" M P)        teóricos
    |-D:cl(icl)
    | |-P:vp*
    | | |-VD:v(pcp "fornecer" M P §AG)  fornecidos
    | |-A:g(pp)
    |   |-H:prp("por" &lt;sam-&gt; &lt;right&gt;)       por
    |   |-D:g(np)
```

Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

Verifica-se que a estrutura “os aportes teóricos”, foi analisada de forma equivocada, uma vez que “aportes” foi etiquetado como verbo, o termo “teóricos”, por sua vez, foi marcado como nome (substantivo), entretanto os referidos termos constituem um sintagma nominal, em que “teóricos” caracteriza o nome “aportes”.

Outro inconveniente encontrado na análise do *software* diz respeito à identificação de estruturas que são de fato sintagmas nominais, ou seja, possuem as regras básicas de formação de sintagmas nominais, mas são estruturas vazias de conteúdo, ou seja, possuem pouca densidade informacional, como “desse modo”, “dessa forma”, “nesse sentido” etc. Essas locuções conclusivas e consecutivas em nada contribuem para representação de conteúdos

documentais, embora sejam estruturas verdadeiramente sintagmáticas. Para isso, recorre-se a listas de *stop-words*, no caso, lista de sintagmas vazios em termos de conteúdo.

A maior parte dos problemas relacionados à identificação de SN, por parte do PALAVRAS, diz respeito à etiquetagem (categorização) das palavras.

Um exemplo que não interferiu na identificação de SN, mas evidencia mais claramente isso, encontra-se na Figura 15:

Figura 15 – Equívoco de etiquetagem (classificação morfológica) realizada pelo PALAVRAS

Trecho analisado:

A metodologia se orientou na revisão de literatura enquanto estudo bibliográfico capaz de [...].

```
A1
SOURCE: Running text
7. A metodologia se orientou na Revisão de literatura enquanto estudo bibliográfico capaz de sinteti
A1
UTT:cl(fcl)
.
|-PRED:par
|-CJT:g(np)
| |-H:prop("SOCIOTERMINOLOGIA" &lt;*&gt; M/F S forg)    SOCIOTERMINOLOGIA
|-CO:conj("e" &lt;*&gt;)        E
|-S:g(np)
| |-D:pron(det "o" &lt;*&gt; &lt;artd&gt; DET F S)      A
| |-H:n("metodologia" F S)      metodologia
|-Od:pron(pers "se" M/F 3S ACC) se
|-P:v(fin "orientar" &lt;fmc&gt; PS 3S IND VFIN §SP)    orientou
|-A:g(pp)
  |-H:prp("em" &lt;sam-&gt; &lt;right&gt;)      em
  |-D:g(np)
    |-D:pron(det "o" &lt;-sam&gt; &lt;artd&gt; DET F S) a
    |-H:n("revisão" &lt;*&gt; F S)      Revisão
    |-D:g(pp)
      |-H:prp("de" &lt;np-close&gt;)    de
      |-D:g(np)
        |-H:n("literatura" F S) literatura
        |-D:cl(acl)
          |-H:adv("enquanto" &lt;rel&gt; &lt;np-close&gt;)      enquanto
          |-SUB&lt;:g(np)
            |-H:n("estudo" M S §TP)     estudo
            |-D:adj("bibliográfico" M S)        bibliográfico
            |-D:adjp
              |-H:adj("capaz" M S §ACT) capaz
              |-D:g(pp)
              | |-H:prp("de")   de
              | |-D:cl(icl)
              |   |-P:v(inf "sintetizar" §SP)   sintetizar
```

Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

O termo “enquanto” foi marcado como “advérbio”, entretanto tal termo é convencionalmente conjunção temporal. Para os gramáticos Celso Cunha e Maria H. de M. Neves, tal conjunção pode assumir a função de conjunção proporcional. Ainda acerca do

“enquanto”, Domingos P. Cegalla aponta que equivale a “ao passo que, mas”, exercendo papel de conjunção adversativa.

Verificou-se também a ocorrência de sintagmas nominais iniciando com “conjunção”, ou seja, um elemento que não é importante para o SN.

Observe o exemplo abaixo:

Figura 16 – Sintagmas iniciando com conjunção

Trecho analisado: Inovação da informação e linguagem jurídicas e sua aplicação na representação da informação jurídica.

```
SOURCE: Running text
1. INOVAÇÃO DE A INFORMAÇÃO E LINGUAGEM JURÍDICAS E SUA APLICAÇÃO EM A REPRESENTAÇÃO DE A INFORMAÇÃO JURÍDICA
A1
PRED:par
|-CJT:g(np)
| |-H:n("inovação" &lt;*&gt; F S §AG)   INOVAÇÃO
| |-D:g(pp)
|   |-H:prp("de" &lt;*&gt; &lt;sam-&gt; &lt;np-close&gt;)       DE
|   |-D:par
|     |-CJT:g(np)
|     | |-D:pron(det "o" &lt;*&gt; &lt;-sam&gt; &lt;artd&gt; DET F S)   A
|     | |-H:n("informação" &lt;*&gt; F S §TP)   INFORMAÇÃO
|     |-CO:conj("e" &lt;*&gt;)  E
|     |-CJT:g(np)
|       |-H:n("linguagem" &lt;*&gt; F S)        LINGUAGEM
|       |-D:adj("jurídico" &lt;*&gt; F P)       JURÍDICAS
|-CJT:g(np)
  |-D:gr*(pp)
  | |-CO:conj("e" &lt;*&gt;)    E
  |-D:pron(det "seu" &lt;poss 3S&gt; &lt;*&gt; DET F S) SUA
  |-H:n("aplicação" &lt;*&gt; F S §TH)  APLICAÇÃO
  |--D:g(pp)
    |-H:prp("em" &lt;*&gt; &lt;sam-&gt; &lt;np-close&gt;)       EM
    |-D:g(np)
      |-D:pron(det "o" &lt;*&gt; &lt;-sam&gt; &lt;artd&gt; DET F S)     A
      |-H:n("representação" &lt;*&gt; F S §AG)  REPRESENTAÇÃO
      |-D:g(pp)
        |-H:prp("de" &lt;*&gt; &lt;sam-&gt; &lt;np-close&gt;)   DE
        |-D:g(np)
          |-D:pron(det "o" &lt;*&gt; &lt;-sam&gt; &lt;artd&gt; DET F S) A
          |-H:n("informação" &lt;*&gt; F S §TP) INFORMAÇÃO
          |-D:adj("jurídico" &lt;*&gt; F S)     JURÍDICA
```

Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

O SN “e sua aplicação na representação da informação jurídica” iniciou-se com a conjunção “e”. Correa e Celerino (2019) fizeram uma adaptação de critérios propostos por Lopes (2012) e Martins (2014), dentre os quais há um voltado para a eliminação de preposições e conjunções que se encontrem no início e no fim de sintagmas.

Na figura 17, observa-se outro exemplo de categorização equivocada de palavras. No caso, o termo “periódicos”, foi classificado como “adjetivo”, quando, no contexto, deveria ser marcado como nome (substantivo), gerando um sintagma nominal.

Figura 17 – Equívoco na etiquetagem feita pelo PALAVRAS

Trecho analisado:

[...] os materiais foram recuperados na Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação, Service Electronic Library Online, Library and Information Science Abstracts, Web of Science, revistas e **periódicos** especializados em Arquivologia, Epistemologia e Organização do Conhecimento, eventos científicos [...]

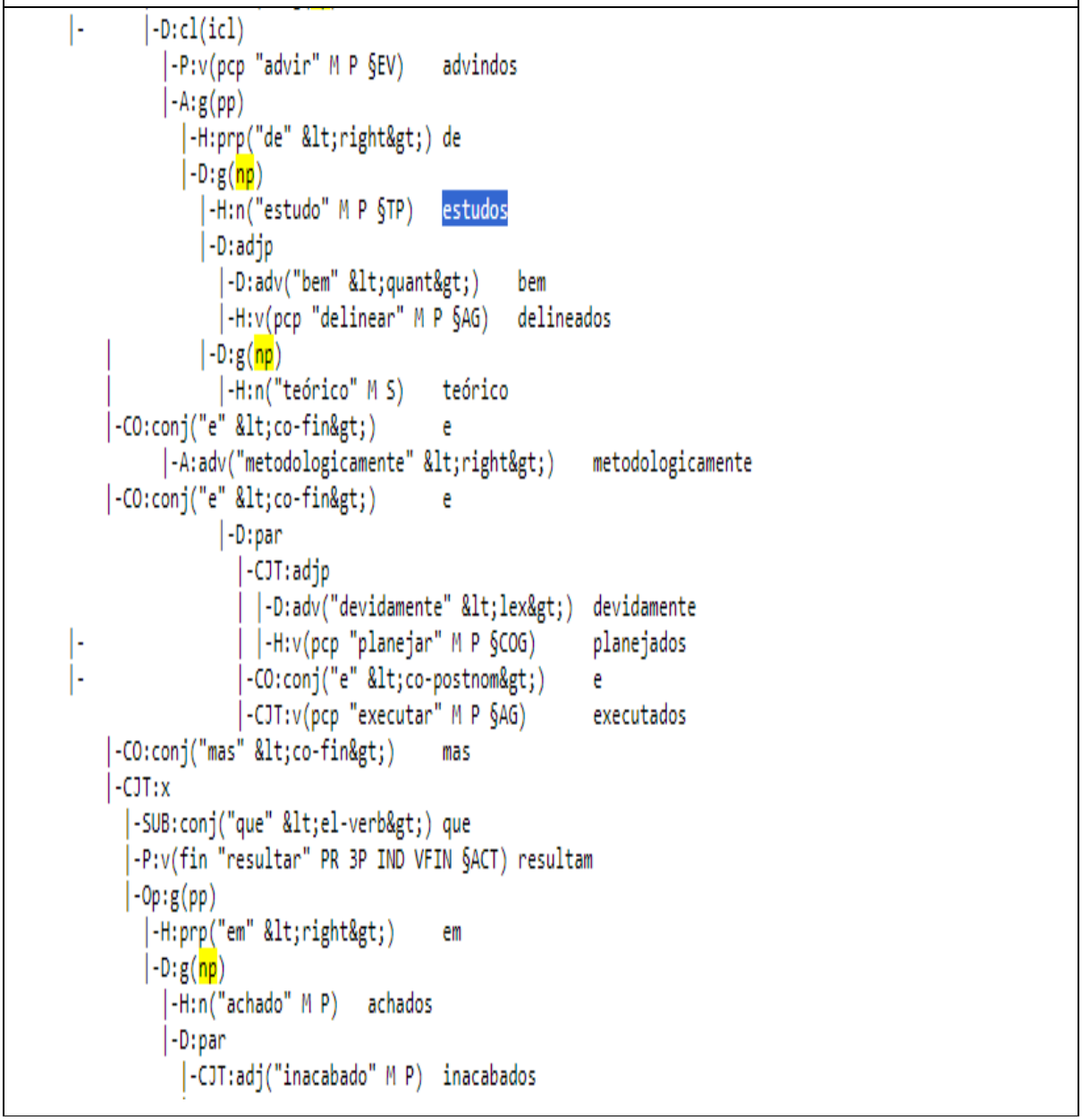
```
|           |          |-D:par
|-          |-D:n("revista" F P)    revistas
|-S:par
  |-CJT:adjp
    |-CO:conj("e" &lt;co-pcv&gt;)   e
    | |-H:adj("periódico" &lt;n&gt; M P)    periódicos
    |-CJT:cl(icl)
      |-P:v(pcp "especializar" M P §COG)    especializados
      |-A:g(pp)
        |-H:prp("em" &lt;right&gt;) em
        |-D:par
          |-CJT:n("arquivologia" &lt;*&gt; F S)     Arquivologia
          |-H:n("perspectiva" F P)  perspectivas
          |-CJT:n("epistemologia" &lt;*&gt; F S)    Epistemologia
          |-CO:conj("e" &lt;co-prparg&gt;)  e
          |-CJT:g(np)
            |-H:prop("Organização_do_Conhecimento" &lt;*&gt; F S £occ)      Organização_do_Conhecimento
            |-P:v(pcp "apresentar" F P §AG) apresentadas
            |-D:g(np)
            | |-H:n("evento" M P)   eventos
            | |-D:adj("científico" M P)     científicos
            | |-D:g(pp)
            |   |-H:prp("como" &lt;np-long&gt;)     como
|-          |   |-D:g(np)
```

Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

Houve dificuldade do referido *software* também em analisar advérbios coordenados, visto que tais advérbios, quando coordenados, aceitam que apenas o último possua a terminação “mente”. Embora essa regra não resulte na identificação de SN, ela permite que o *software* apresente um comportamento mais robusto e eficiente na análise de textos.

Figura 18 – Etiquetagem equivocada de estruturas formadas por advérbios coordenados

Trecho analisado: [...] podendo ou não possuir natureza aberta, advindos de estudos bem delineados**, teórico e metodologicamente**, e devidamente planejados e executados, mas que resultam em achados inacabados ou inesperados, fato que não se traduz em falta de qualidade científica.



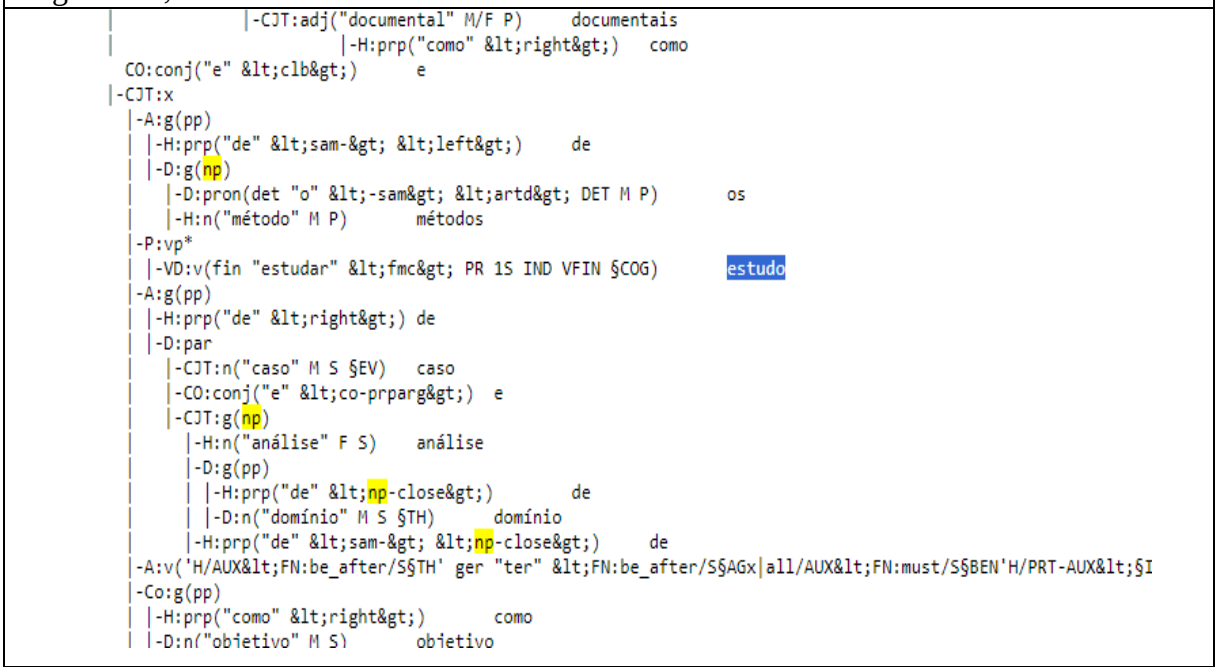
Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

Mais adiante, na Figura 19, verifica-se que o termo “estudo”, ora classificado como “verbo” não o deveria ser, visto que, no contexto, tal termo está sendo empregado como substantivo, ou seja, como nome, logo deveria remeter a um sintagma nominal.

Figura 19 – Omissão de Sintagmas Nominais devido a erro de etiquetagem

Trecho analisado:

[...] abordagem qualitativa, a qual se vale de procedimentos bibliográficos e documentais, e dos métodos **estudo de caso** e análise de domínio, tendo como objetivo refletir o papel da garantia cultural em SOCs, partindo da averiguação da diversidade cultural e das variações linguísticas, sob o viés da Arca do Gosto.



Fonte: VISL. Disponível em: https://edu.visl.dk/visl/en/parsing/automatic/trees.php.

De modo geral, boa parte das omissões de SN, os quais deveriam ser identificados pelo *software*, se deve a equívocos na etiquetagem das palavras, gerando, assim, errônea identificação de sintagmas nominais.

Análise das Etapas 6 – Ordenação dos SN, 7 – Análise da relevância semântica dos SN e 8 – Analisar os termos extraídos e classificados como sintagmas autorizados[27]

Análise detalhada dos critérios de seleção aplicados aos SN do experimento

[27] Etapa 6 - Ordenação dos SN conforme os critérios: a frequência absoluta, a frequência normalizada, a frequência relativa de ocorrência de termos, a estrutura (nível) dos sintagmas e o uso de *stop words*; Etapa 7 – Análise da relevância semântica dos SN e da viabilidade dos critérios aplicada na Etapa 6; e a Etapa 8 – Analisar os termos extraídos e que foram classificados como sintagmas autorizados

Ainda que o PALAVRAS tenha apresentado falhas na identificação de alguns SN, esse *software*, considerado pela literatura como um dos melhores para identificação de SN (Maia, 2008; Morellato, 2010; Lopes, 2012; Silva, 2014; Martins, 2014), de modo geral, alcançou um bom desempenho na identificação dos Sintagmas. O PALAVRAS identificou 1.235 estruturas como sendo SN, sendo que, deste total, 1.212 constituíam expressões que realmente eram SN, nomeados, nesta pesquisa, de **legítimos SN,** e 23 não constituíam Sintagmas Nominais, representando, assim, um total de 98% de acerto em relação aos seus próprios resultados. Esse percentual confirma a taxa obtida pelo seu idealizador Bick (2000), que foi de 98% de acerto na identificação de SN. Essas taxas de acertos na identificação de SN não ficam tão distantes das obtidas por Silva (2014) e Nascimento (2015), que foram de 94% de acerto em relação aos seus próprios resultados em cada uma das pesquisas.

Posteriormente, esses 1.212 SN foram transcritos para planilhas eletrônicas, compondo uma lista com legítimos SN para cada documento do *corpus* desta pesquisa. À reunião destas listas com os SN, identificados automaticamente e extraídos manualmente, deu-se o nome de **Lista com legítimos SN verdadeiros**. Em seguida, prosseguiu-se com a inclusão das palavras-chave atribuídas pelos autores à referida planilha, para que, assim, pudessem ser manuseadas mais facilmente.

Antes de verificar o desempenho de cada critério, realizou-se o cotejamento dos SN de cada documento com as palavras-chave atribuídas pela indexação dos autores com o intuito de verificar o desempenho do *software* em relação aos seus próprios resultados. Dessa forma, os SN que eram semelhantes às palavras-chave ou que continham tais palavras em suas estruturas foram considerados Sintagmas Nominais Descritores, e os que não se encaixavam nesses dois critérios foram considerados apenas Sintagmas Nominais. A escolha pelas palavras-chave para cotejamento dos SN e categorização dos SN em descritores ou não, se deu pelo o corpus desta pesquisa ser da área de organização e representação do conhecimento, logo os autores dessa área possuem o domínio mínimo para a prática da indexação. Assim, calcularam-se as taxas de revocação e precisão sem a aplicação de cada critério para que, em seguida, fosse possível realizar comparações entre as taxas de revocação e precisão com e sem a aplicação de cada critério.

Na análise da utilidade dos critérios na separação dos sintagmas nominais descritores dos não descritores, levando em conta todos os sintagmas nominais extraídos dos documentos da coleção, resultou em um *corpus* constituído de 267 sintagmas nominais descritores (semelhantes às palavras-chave ou as contêm em suas estruturas) e 945 sintagmas nominais

não descritores. Esta análise resultou numa precisão de 22% e revocação de 100% quando da não aplicação de nenhum critério.

Para a análise do comportamento de cada critério, procedeu-se com a aplicação por critério a cada um dos 24 artigos que compuseram este experimento, para que, assim, fosse possível levantar as medidas de revocação e precisão com a aplicação de cada critério. Toda a análise foi operacionalizada por meio de planilhas do excel. Na Figura 20, há a amostra de uma das planilhas utilizadas para a análise de cada critério e sua aplicação a cada sintagma de cada documento dos 24 artigos.

Figura 20 – Exemplo de planilha utilizada para a análise de cada sintagmas nominais e aplicação dos critérios

	A	B	C	D	E	F	G	H	I	
1		**Número de Sintagmas:**	**63**							
2										
3	**ORDEM**	**LEGÍTIMOS SINTAGMAS NOMINAIS**	**Nível**	**Freq. Doc.**	**Freq. Norm**	**Freq. Coleção**		**IDF**	**Stoplist**	**CONTÉM OS**
4	62	a área da organização do conhecimento	3	1	0,01587302	1		1,380211242		Contém
5	61	dois pesquisadores da área da organização do conhecimento	4	1	0,01587302	1		1,380211242		Contém
6	29	a análise discursiva dos trabalhos	2	1	0,01587302	1		1,380211242		
7	21	a análise do discurso	2	1	0,01587302	1		1,380211242		Contém
8	60	a aparente influência de dois pesquisadores da área da organização do conh	5	1	0,01587302	1		1,380211242		Contém
9	52	a biblioteconomia	1a	1	0,01587302	2		1,079181246		
10	47	a formação discursiva de organização da informação	3	1	0,01587302	1		1,380211242		Contém
11	55	a formação discursiva de organização do conhecimento	3	1	0,01587302	1		1,380211242		Contém
12	2	a formação ideológica	1b	1	0,01587302	1		1,380211242		
13	50	a ligação histórica da ciência da informação com a biblioteconomia	3	1	0,01587302	1		1,380211242		Contém
14	43	a organização com a ciência da informação	3	1	0,01587302	1		1,380211242		Contém
15	38	a organização da informação	2	1	0,01587302	2		1,079181246		Contém
16	42	a relação da organização com a ciência da informação	4	1	0,01587302	1		1,380211242		Contém
17	39	ambas influenciadas por suas próprias epistemologias e prismas	3	1	0,01587302	1		1,380211242		
18	16	as posições discursivas	1b	1	0,01587302	1		1,380211242		
19	28	auxílio	1a	1	0,01587302	1		1,380211242		

... Art. 1 | Art.2 | Art.3 | Art.4 | Art.5 | Art.6 | Art. 7 | Art.8 | Art.9 | Art.10 | Ar.11 | Art.1 ...

Fonte: Desenvolvido pelo autor.

Critério: Frequência absoluta e frequência normalizada

Verificar a frequência de ocorrência de um termo ou sintagma nominal em um documento é uma medida bastante utilizada nas estratégias de indexação semiautomática. A DF, da sigla em inglês “Document Frequency”, verifica a quantidade de vezes que cada termo ocorre no documento, partindo do pressuposto de que, quanto maior a sua ocorrência, maior a possibilidade de este termo ser mais representativo do que outros que apresentam baixa frequência de ocorrência no documento.

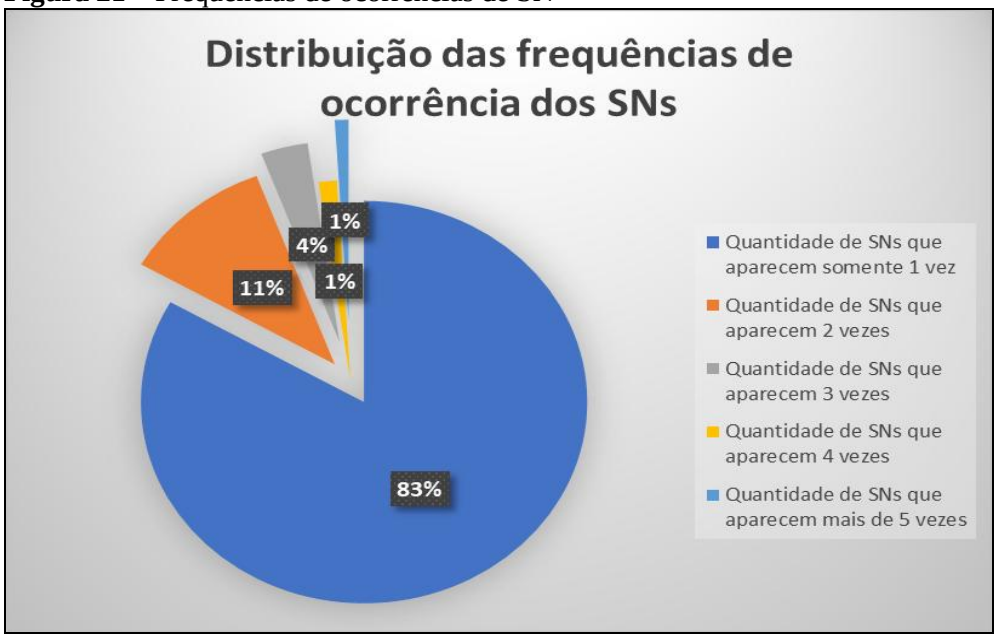
Com base na lei de Zipf, Hans Peter Luhn (1957) sugere que as palavras, conforme as suas frequências de ocorrência, possuem maiores pesos para fins de representação do

documento do qual são retiradas. Nem todas as palavras que ocorrem num documento representam semanticamente o conteúdo tratado no documento, palavras com frequência muito baixa seriam pouco significativas para fins de representação do documento. Ao lado da frequência absoluta de cada termo, a frequência normalizada busca corrigir qualquer distorção introduzida pelo comprimento do documento, considerando, assim, a extensão do documento para atribuição de pesos aos termos desse documento. Tal medida é alcançada pela divisão da frequência absoluta do SN no documento pelo número total de SN ocorridos no documento

Do mesmo modo que ocorre com as palavras isoladas, a frequência de ocorrência e a frequência normalizada também são aplicadas às estruturas linguísticas mais complexas e mais precisas em termos de sentido, ou seja, aos sintagmas nominais. No experimento que constituiu esta pesquisa, tanto a frequência de ocorrência quanto a frequência normalizada foram aplicadas a cada documento constituinte do *corpus* da pesquisa.

Na Figura 21, é possível ver a distribuição de taxas de frequências de ocorrências dos sintagmas nominais do *corpus* desta pesquisa, constituído de 1.212 sintagmas, os quais foram extraídos de 24 artigos publicados na edição 2022 do Enancib.

Figura 21 – Frequências de ocorrências de SN



Fonte: Dados da pesquisa (2024)

Na Figura 21, é possível perceber que a maioria dos sintagmas de todo o *corpus* aparece apenas uma vez. Esse fato pode estar relacionado à extensão do *corpus* utilizado que, nesta pesquisa, foi constituído de "títulos, subtítulos e resumos". De todo o *corpus*, 1.008 sintagmas apareceram apenas uma vez, 134 ocorreram duas vezes, 42 sintagmas ocorreram

três vezes, 16 sintagmas apareceram quatro vezes, e 12 sintagmas apresentaram cinco ou mais ocorrências.

No Quadro 14, é possível verificar os percentuais de SN descritores em cada faixa de ocorrência.

Quadro 14 - Percentuais de SN descritores com base na frequência de ocorrência dos SN nos documentos

QUANTIDADE DE SN QUE APARECEM	TOTAL DE SN DESCRITORES	TOTAL DE SN NÃO DESCRITORES	PERCENTUAL DE SN DESCRITORES
somente 1 vez	189	819	18,70%
2 vezes	48	86	35,80%
3 vezes	9	33	21,40%
4 vezes	16	0	100%
mais de 5 vezes	7	5	42%

Fonte: Dados da pesquisa (2024)

Os percentuais presentes no Quadro 14 corroboram o que a literatura evidencia acerca da frequência de ocorrência de termos em documentos, uma vez que o percentual de SN descritores aumentam consoante aumentam as frequências de ocorrências do SN no documento. Isso fica nítido quando se analisam os percentuais de SN descritores para os SN que aparecem uma vez e duas vezes, respectivamente, 18,70% e 35,80%.

Comparando os SN que aparecem uma única vez e todas as outras frequências de ocorrências, percebem-se maiores percentuais, ou seja, a frequência de ocorrência do termo, nesse caso do SN, influencia diretamente no seu potencial como descritor para fins de representação temática.

Ao analisar todas as frequências de ocorrências, os que apareceram duas vezes apresentaram maior percentual de SN descritores, quando comparados com os que aparecem apenas uma vez. Observar a frequência absoluta, ou seja, a frequência de ocorrência de um termo (SN) no documento influencia diretamente na relevância de um sintagma, ou seja, a relevância aumenta conforme a frequência de ocorrência, entretanto há de se observar as frequências demasiadamente altas, as quais podem estar associadas a uma tendência de saturação. Conforme Souza (2005), a saturação está associada aos termos com frequência

demasiado alta podem ser descritores insignificantes; expressões por demais comuns, com pouco poder discriminatório.

Estabelecer pontos de cortes para essas frequências de ocorrências é importante para alcançar mais objetividade, entretanto é preciso ter cuidado para adotar níveis "seguros". De modo geral, infere-se que todas as frequências acima de uma ocorrência apontam para sintagmas nominais candidatos mais fortes que descritores documentais. Sobre esse aspecto, Souza (2005, p. 117) aponta que "a frequência de ocorrência de cada SN é diretamente proporcional à relevância como descritor, com a indicação de possível saturação".

Os percentuais presentes no Quadro 14 corroboram essa constatação de Souza (2005) que também se debruçou no uso dos SN na indexação. Ademais, em Nascimento (2015), a frequência de ocorrência de um sintagma no documento também se mostrou pertinente e evidenciou que a quantidade de ocorrência do SN estava diretamente relacionada ao seu potencial como descritor.

Sintagmas que ocorrem com maior frequência tendem a ser melhores candidatos a descritores documentais. São expostos abaixo os valores percentuais de revocação e precisão para os sintagmas nominais que aparecem uma vez, duas vezes, três vezes, quatro vezes e cinco ou mais vezes.

Quadro 15 - Taxa de revocação e precisão dos SN conforme a frequência de ocorrência do SN

QUANTIDADE DE SN QUE APARECEM	TOTAL DE REVOCAÇÃO	TOTAL DE PRECISÃO
01 vez	70,2%	18,70%
02 vezes	17,8%	35,80%
03 vezes	3,3%	21,40%
04 vezes	5,9%	100%
mais de 05 vezes	1,8%	42%

Fonte: Dados da pesquisa (2024)

Ao considerar a revocação, os SN que apareceram uma vez e duas vezes apresentaram melhores resultados. Já para a precisão, os melhores índices concentraram-se nas frequências de 2, 3, 4 e 5 ocorrências, apresentando melhor percentual de precisão, quando da não aplicação de critérios que foi de 22%. Verifica-se que a frequência de ocorrência absoluta

acima de um demonstra melhores resultados em comparação com a frequência de ocorrência de apenas uma única vez.

Embora a taxa de revocação não tenha apresentado um bom desempenho no tocante à frequência de ocorrência do SN no documento, foi possível ratificar a ideia, já evidenciada na literatura da área, de que a frequência de ocorrência de um determinado termo tem muito a ver com o poder discriminatório do SN. Ao olhar para a precisão, as taxas alcançadas com todos os cálculos evidenciaram nitidamente a utilidade de se olhar a frequência de ocorrência de um SN no documento, uma vez que conseguiu ser mais preciso em selecionar SN descritores e evitar os SN considerados não descritores. Os baixos percentuais relacionados à revocação podem estar relacionados ao *corpus* trabalhado neste experimento devido às próprias características dos elementos textuais utilizados.

Ao lado da frequência absoluta, tem-se a frequência normalizada, a qual considera a extensão do documento. Essa frequência considera o tamanho do documento, por esse motivo, tal frequência é bastante variável.

Quadro 16 - Exemplo de frequência normalizada do Art.3 que constituiu o *corpus* deste experimento

Art. 3		**Art. 4**	
QUANTIDADE DE SN QUE APARECEM	FREQUÊNCIA NORMALIZADA		FREQUÊNCIA NORMALIZADA
01 vez	0,015873016	Quantidade de SN que aparecem somente 1 vez	0,02857143
2 vezes	0,031746032	Quantidade de SN que aparecem 2 vezes	0,05714286
3 vezes	0,047619048	Quantidade de SN que aparecem 3 vezes	0,08571429

Fonte: Dados da pesquisa (2024)

A frequência normalizada varia conforme a quantidade de sintagmas de cada documento como pode ser vista no Quadro 16. Logo, a frequência absoluta se torna mais fácil de ser operacionalizada em uma proposta de indexação semiautomática. A frequência normalizada é viável na atribuição de pesos aos SN, visto que considera, ao mesmo tempo, a frequência absoluta e a extensão de cada documento.

Embora a frequência normalizada não tenha sido crucial neste experimento, constatação percebida em outra pesquisa de Nascimento (2015), fazem-se necessários mais estudos acerca desse critério no contexto específico dos Sintagmas Nominais. Em Souza e

Raghavan (2014), a frequência normalizada se mostrou viável, alcançando melhores resultados quando comparados com a não aplicação da frequência normalizada.

Além da frequência do SN no documento, há de se considerar a frequência do sintagma na coleção como um todo, evidenciando, assim, o possível caráter genérico de um sintagma com alta frequência absoluta e normalizada. Nos parágrafos seguintes, é observada a frequência do sintagma nominal na coleção como um todo, partindo do pressuposto de que um sintagma que é frequente nos documentos de uma coleção tende a ser genérico demais e, assim, possuir pouco poder discriminatório já que é bastante genérico. Essa frequência inversa supre o inconveniente dos sintagmas que, embora sejam frequentes no documento, são genéricos demais por estarem presentes demasiadamente na coleção como um todo.

Critério: Frequência relativa de ocorrência de termos

Observar a frequência de ocorrência de um sintagma em um documento é fundamental para o julgamento de tal unidade com potencialmente descritiva, contudo, a sua ocorrência em demásia em outros documentos pode indicar o oposto, evidenciando que tal sintagma é bem mais genérico do que específico. Borges (2009), acerca desse critério de seleção, aponta que a frequência com que um termo aparece na base de dados é mais importante do que a frequência com que o mesmo termo aparece em um documento.

Dessa forma, os SN não tão comuns possuem maior potencial descritivo do que os que são frequentes em vários documentos. É possível observar, no Quadro 17, que, de fato, os SN mais frequentes na coleção são bem mais genéricos do que outros termos que apresentaram frequência de ocorrência baixa na coleção como um todo.

No Quadro 17, é possível verificar os SN que apresentaram uma frequência de ocorrência alta no conjunto do *corpus* deste experimento:

Quadro 17 – Sintagmas Nominais com maior frequência de ocorrência no *corpus* do experimento

SINTAGMAS NOMINAIS	FREQUÊNCIA DE OCORRÊNCIA EM TODO O *CORPUS*
o conhecimento	14
organização do conhecimento	9
a organização do conhecimento	7
a informação	6
a ciência da informação	6
os resultados	5

o Brasil	5
este estudo	4
esta pesquisa	4
a metodologia	4

Fonte: Dados da pesquisa (2024)

O sintagma nominal “o conhecimento” apareceu em 14 documentos do total de 24 que constituiu o *corpus* do experimento, demonstrando, assim, o seu caráter genérico, ou seja, constitui-se em um sintagma que é recorrente no domínio trabalhado, possuindo pouco poder informacional (valor discriminatório)[28]. Um SN, com frequência demasiadamente alta, tende a perder o seu valor discriminatório, ou seja, possui pouco valor para fins de representação temática de um determinado documento, visto que é comum em vários documentos.

Outro sintagma frequente na coleção é “organização do conhecimento”, ocorrendo em 9 documentos. Embora esse SN seja bastante comum e usado dentro do domínio estudado, ele se trona genérico demais por ser frequente em boa parte da coleção, ou seja, há uma perda de poder discriminatório no referido sintagma. Isso evidencia a pertinência de se considerar os valores relativos ao “inverso da frequência de ocorrência no *corpus*”. Não basta que um determinado SN ocorra com frequência no documento, é preciso verificar se esse mesmo SN é recorrente na base como um todo.

A associação do critério “frequência do termo no documento indexado” ao “inverso da frequência de ocorrência (ou seja, frequência na coleção)” permite a aferição de forma mais abrangente e precisa do SN no que se refere ao seu valor descritivo. Para que o SN seja considerado pertinente para a representação temática de um documento, é preciso que ele não seja tão frequente na coleção, visto que essa popularidade extensa demonstra um caráter genérico e comum do SN.

Segundo Fujita e Corrêa (2024), TF (da sigla em inglês para Term Frequency) representa a frequência normalizada de um termo, em um determinado, utiliza-se IDF (da sigla em inglês para *Inverse Document Frequency*), para demonstrar a frequência inversa nos documentos da coleção. A associação da frequência normalizada com o inverso da frequência de ocorrência - TF-IDF - busca representar a multiplicação da frequência normalizada do termo em um documento pela frequência inversa na coleção.

Lopes (2012) evidencia claramente o desempenho desse critério, o qual também foi utilizado por Nascimento (2015) e apresentou bom desempenho na seleção de SN para fins de

[28] Optou-se nesta tese por utilizar a expressão “poder informacional” (ou valor discriminatório) para representar o potencial do SN para fins de representar o poder discriminatório de um SN em relação a outro.

representação temática de documentos. No Quadro 18, são expostos dados quantitativos de SN descritores e não descritores que apareceram nas faixas de frequência 1, 2, 3, 4 ou 5 ou mais documentos do *corpus* do experimento.

Quadro 18 – Quantitativo de SN descritores e ou não descritores que apareceram em 1, 2, 3, 4 ou 5 ou mais documentos

COLEÇÃO - QUANTIDADE DE SN	TOTAL DE SN DESCRITORES	TOTAL DE SN NÃO D DESCRITORES	PERCENTUAL DE SN DESCRITORES	IDF
01 documento	229	776	22,7%	1,380211242
02 documentos	14	58	19,4%	1,079181246
03 documentos	6	33	15,3%	0,903089987
04 documentos	0	12	0,0%	0,77815125
05 ou mais documentos	14	70	16,6%	0,681241237 / 0,234083206

Fonte: Dados da pesquisa (2024)

Com base no Quadro 18, verifica-se que, de todos os SN descritores da coleção, a maior parte deles aparece em apenas um documento dos 24 que compõem o *corpus*. Analisando a quarta coluna do referido quadro, pode-se perceber que, à medida que os SN aparecem em mais documentos, o percentual de SN descritores diminui, evidenciando, desse modo, o que foi visto na literatura acerca da frequência demasiada de SN. Logo, SN que aparecem com frequência em muitos documentos tendem a possuir pouco poder discriminatório para fins de representação temática do documento, não servindo como descritores documentais.

Verifica-se também que os percentuais de SN descritores apresentam melhores taxas para as faixas de ocorrência, em 1 e 2 documentos, que possuem, respectivamente, os seguintes percentuais de SN descritores: 22,7% e 19,4%. No tocante ao IDF, nota-se que o IDF igual ou acima de um é o que tem um melhor rendimento em termos de selecionar SN descritores com maior potencial discriminatório quando analogamente há SN comuns na coleção, mais genéricos.

A seguir, são demonstradas as taxas de revocação e precisão alcançadas para cada faixa de ocorrência dos SN em um, dois, três, quatro e em cinco ou mais documentos.

Esses dados encontram-se no Quadro 19:

Quadro 19 – Taxas de revocação e precisão para as diferentes frequências de ocorrência na coleção

TAXAS DE REVOCAÇÃO E PRECISÃO PARA CADA FREQUÊNCIA DE OCORRÊNCIA NA COLEÇÃO		
COLEÇÃO – QUANTIDADE DE SN QUE APARECEM EM	**REVOCAÇÃO**	**PRECISÃO**
01 documento	87,0%	22,7%
02 documentos	5,0%	19,4%
03 documentos	2,0%	15,3%
04 documentos	0,0%	0,0%
05 ou mais documentos	5,0%	16,6%

Fonte: Dados da pesquisa (2024)

Se se considerar a taxa de revocação e precisão, os SN que apareceram em apenas um documento da coleção apresentaram melhores valores. Observando especificamente a precisão, verifica-se que os SN que aparecem em até dois documentos da coleção apresentam percentuais razoáveis, inclusive a taxa de precisão para os SN que aparecem em um documento da coleção é mais alta quando da não aplicação de nenhum critério que é de 22%. Isso reforça a importância de considerar a frequência do SN na coleção. É possível observar que os valores são decrescentes conforme os SN são mais frequentes na coleção, evidenciando a viabilidade da medida IDF para fins de seleção de SN que atuam como descritores documentais.

Os percentuais para a última faixa de frequência de ocorrência apresentam uma discrepância, com percentuais maiores do que aqueles para as frequências de ocorrência em três e quatro documentos. Esse desvio está associado ao uso dos sintagmas “organização do conhecimento”, “organização da informação” e “ciência da informação”, os quais, embora sejam abrangentes, foram utilizados com frequência pelos autores dos textos analisados e, ao mesmo tempo, foram recuperados como sintagmas com frequência em alguns documentos da coleção.

No Quadro 20, a título de exemplificação, apresenta-se a aplicação do índice IDF no artigo 6, constituinte do *corpus* deste experimento.

Quadro 20 – Exemplo de SN ordenados conforme índice IDF

SINTAGMAS NOMINAIS	Freq. Coleção	IDF	Posição	SN semelhante ou contém palavras-chave dos autores
a integração de dados	1	1,380211	1ª	Contém
integração de dados	1	1,380211	2ª	Contém
o contexto da ciência da informação	1	1,380211	3ª	
a discussão sobre o tema	1	1,380211	5ª	
uma temática abordada na área	1	1,380211	6ª	
os tópicos	1	1,380211	7ª	
ambientes digitais	1	1,380211	8ª	
o uso das tecnologias da web semântica	1	1,380211	9ª	Contém
as tecnologias da web semântica	1	1,380211	10ª	Contém
a web semântica	1	1,380211	11ª	Contém
os dados interligados	1	1,380211	12ª	Contém
um termo	1	1,380211	13ª	
os sistemas	1	1,380211	14ª	
as unidades de informação	1	1,380211	15ª	
uma visão integrada de um conjunto de recursos de informação	1	1,380211	16ª	
um conjunto de recursos de informação	1	1,380211	17ª	
recursos de informação	1	1,380211	18ª	
a web	2	1,079181	19ª	
o objetivo deste trabalho	2	1,079181	20ª	
o tema	2	1,079181	21ª	
a integração	2	1,079181	22ª	
a área	2	1,079181	23ª	
a integração	2	1,079181	24ª	
este trabalho	3	0,90309	25ª	
a ciência da informação	6	0,60206	26ª	
a organização do conhecimento	7	0,535113	27ª	

o conhecimento	14	0,234083	28ª	

Fonte: Dados da pesquisa (2024)

No Quadro 20, é possível visualizar a funcionalidade e a aplicabilidade do índice IDF na seleção de SN que possam servir como descritores documentais em um sistema de indexação semiautomático. É possível perceber que os SN semelhantes às palavras-chave atribuídas pelos autores do documento ou que contém tais palavras-chave em sua estrutura apresentam IDF acima de um e frequência, na coleção de ocorrência, em apenas um documento de toda a coleção.

Os últimos sintagmas, respectivamente, "a ciência da informação", "a organização do conhecimento" e "o conhecimento" apresentam frequência em demásia na coleção, possuindo, assim, IDF abaixo de um. Por meio desses sintagmas e de suas respectivas posições, conforme o índice IDF, fica evidente a funcionalidade do referido índice na seleção de SN candidatos a descritores documentais.

A aplicação do índice IDF se mostrou viável na seleção de SN. Tal índice associado a outros critérios resultam em SN com maior poder discriminatório. Desse modo, fica evidente a viabilidade de se aplicar tal critério na indexação automática baseada em sintagmas nominais. Sintagmas nominais bastante frequentes em vários documentos tendem a possuir pouco valor discriminatório, logo, é preciso considerar tal critério em propostas de indexação semiautomática.

Para um documento que trate da "integração de dados na *web* a partir do desenvolvimento dos catálogos", é perceptível que os sintagmas "o uso das tecnologias da web semântica", "as tecnologias da *web* semântica" e a "a *web* semântica" possuem mais potencial discriminatório do que os seguintes sintagmas "a ciência da informação", "a organização do conhecimento" e "o conhecimento", com baixo índice IDF quando comparados com aqueles.

O índice IDF apresentou bom desempenho no *corpus* utilizado. Tal desempenho também foi percebido no domínio do Direito, quando da realização da pesquisa Nascimento (2015). Isto posto, a frequência de ocorrência do SN nos diversos documentos de uma coleção deve ser vista como um critério a ser incorporado em propostas de indexação semiautomática baseada em SN, visto que tal critério contribui diretamente para a seleção de SN com maior potencial descritivo.

É pertinente ressaltar que a adoção desse critério na indexação semiautomática, por meio de SN, não exclui a necessidade de outros critérios, ou seja, a associação desse critério

com os expostos a seguir contribui efetivamente para a seleção dos SN com maior poder discriminatório para fins de representação temática da informação.

Critério: Estrutura (nível[29]) dos sintagmas

Observar a frequência de ocorrência de um SN, no documento e na coleção como um todo, é fundamental para propostas de indexação semiautomática, uma vez que tais índices já apresentaram bom desempenho em pesquisas que utilizaram palavras isoladas como fontes de informação. De forma semelhante às palavras isoladas, esses critérios apresentaram bom comportamento também na seleção de SN, não obstante é preciso, no caso de uso de SN na indexação, observar outro critério, não mais estatístico, mas semântico e sintático de cada sintagma nominal.

Sobre esse critério, Souza (2005, p. 118) explica que "[...] a complexidade da estrutura do SN e o seu nível são proporcionais à sua densidade informacional [...]". Assim, é preciso considerar esse critério na ponderação dos SN para a valoração da relevância de cada SN. A atribuição do nível para cada Sintagma contribui diretamente para a verificação do potencial discriminatório que um determinado sintagma de nível 2 ou 3, por exemplo, possui em comparação com um Sintagma de nível 1, constituído por um determinante e um nome, que é, normalmente, bem geral e comum, possuindo pouco valor discriminatório.

Ao se observar os SN "o uso das tecnologias da web semântica" e "a web semântica", são nítidas a precisão e a especificidade presentes naquele sintagma quando análogo a este último. O primeiro sintagma é de nível três, pois possui outros dois sintagmas embutidos em sua estrutura, já o sintagma "a web semântica" é de nível 1, pois não possui outro sintagma embutido em sua estrutura. Observando tais sintagmas, é possível notar o grau de especificidade daquele quando comparado ao de nível 1.

Outro exemplo que evidencia a ideia de o nível do SN estar associado ao seu potencial como descritor documental são os sintagmas: "recuperação da informação" e "a informação". Este pode ser interpretado dentro de diferentes contextos, por ser genérico demais, enquanto aquele "recuperação da informação" já apresenta um sentido mais específico e pode ser interpretado como pertencente ao contexto da computação, da biblioteconomia ou ciência da informação, por exemplo.

[29] Nesta pesquisa, a classificação dos sintagmas nominais em nível é baseada nas relações de encadeamento de cada sintagma nominal conforme ressaltado por Kuramoto (1995). Para esse autor, por meio dessas relações de encadeamento, é possível classificar o nível dos sintagmas nominais pela quantidade de outros sintagmas que esses englobam.

Para categorização do nível de cada sintagma nominal, utilizaram-se a categoria e os valores estabelecidos por Souza e Raghavan (2014),[30]a saber, o sintagma de nível 1 (Determinante + Nome) recebeu a categoria 1a, o sintagma de nível 1 (formado por qualquer estrutura, exceto "Determinante + Nome") recebeu a categoria 1b, o sintagma de terceiro nível, formado por qualquer estrutura, recebeu a categoria 3, o sintagma de quarto nível, também formado por qualquer estrutura, recebeu a categoria 4, e o sintagma de nível 5 ou superior, constituído por qualquer estrutura, recebeu a categoria >4.

O Quadro 21 mostra alguns Sintagmas e seus respectivos níveis, demonstrando a distinção em termos de potencial semântico de cada sintagma e seu nível, por exemplo, um sintagma de nível 2 possui maior densidade informacional do que o sintagma de nível 1 formado por **determinante[31] + nome**.

Quadro 21 – Exemplos de Sintagmas Nominais e seus respectivos níveis do art. 06 do *corpus* que constituiu o experimento desta pesquisa

SINTAGMAS NOMINAIS	Nível do SN
a integração de dados	2
a web	1a
o desenvolvimento dos catálogos	2
os catálogos	1a
o contexto dos ambientes digitais	2
os ambientes digitais	1b
uma inquietação	1a
a história dos padrões de catalogação	3
os padrões de catalogação	2
a contextualização dessa história	2
essa história	1a
este trabalho	1a
essa integração	1a
uma definição	1a
o termo	1a

[30] O detalhamento das categorias e valores atribuídos por estes autores encontra-se na página 151 desta tese.

[31] Os elementos que constituem o SN são detalhados no Quadro 10 desta tese e baseados em Perini (2016). São os elementos que aparecem depois do predeterminante e, na ausência dele, aparecem como primeiro elemento do SN. São eles: a, um, esse, aquele, algum, nenhum, cada, que, qual etc.

integração de dados	2
o contexto da ciência da informação	3
a ciência da informação	2
a ciência da informação	2
a integração de dados	2
essa integração	1a
a necessidade e importância de sua realização	3
a integração	1a
um levantamento bibliográfico	1b
a integração	1a
uma temática abordada na área	2
a área	1a
os tópicos	1a
a organização do conhecimento	2
o conhecimento	1a
ambientes digitais	1b
o uso das tecnologias da web semântica	3
as tecnologias da web semântica	2
a web semântica	1b
os dados interligados	1b
a integração	1a
um objetivo	1a
seus benefícios mencionados	1b
o nível das fontes de informação dos sistemas	4
as fontes de informação dos sistemas	3
os sistemas	1a
as unidades de informação	2
uma visão integrada de um conjunto de recursos de informação	4
um conjunto de recursos de informação	3
recursos de informação	2

Fonte: Dados da pesquisa (2024)

Os sintagmas presentes no Quadro 21 constituem uma parte dos sintagmas extraídos do art. 6, que constituiu o *corpus* do experimento realizado nesta pesquisa. O referido artigo possui o seguinte título: “DEFININDO INTEGRAÇÃO DE DADOS NA *WEB* A PARTIR

DO DESENVOLVIMENTO DOS CATÁLOGOS". Tal estudo busca compreender em que consiste a integração de dados, a sua necessidade e a importância de sua realização no contexto da Ciência da Informação. Considerando essa pequena síntese do conteúdo tratado no art. 6 e analisando os sintagmas destacados em verde (nível 1b e nível 2) e amarelo (nível 1a), verifica-se a distinção no tocante ao potencial discriminatório dos diferentes níveis de sintagmas, visto que os sintagmas destacados em verde são mais específicos e tendem a ser fortes candidatos a descritores documentais, enquanto os sintagmas destacados em amarelo são mais gerais, ou seja, mais fracos em termos de potencial informativo para fins de representação temática da informação.

Por meio dos dados expostos no Quadro 21, corrobora-se o que Souza (2006, p. 118) ressaltou acerca da importância do nível do SN para fins de representação temática da informação, ou seja, "[...] a complexidade da estrutura do SN e o seu nível são proporcionais à sua densidade informacional". Ainda segundo esse autor, "Há também que se considerar que os SN 'extensos', como os de nível 4 ou superior, e os SN com estruturas sintáticas muito complexas não são bons candidatos a descritores por lhes faltar certa concisão, desejável nos descritores".

No Quadro 22, expõem-se os dados quantitativos referentes aos SN descritores e não descritores e seus respectivos níveis (1a, 1b, 2, 3, 4, >4). É possível perceber em quais níveis há maior percentual dos SN.

Quadro 22 – Quantitativo de SN descritores e não descritores e os seus respectivos níveis

NÍVEL	TOTAL DE SN DESCRITORES	TOTAL DE SN NÃO DESCRITORES	PERCENTUAL DE SN DESCRITORES
1a	37	325	10,20%
1b	22	192	10,20%
2	96	255	27,30%
3	63	108	36,80%
4	27	47	36,40%
>4	19	21	47,50%

Fonte: Desenvolvido pelo autor (2024)

Os percentuais presentes no Quadro 22 corroboram a ideia de que SN de nível 2, por exemplo, possuem maior potencial discriminatório do que os de nível 1a e 1b. Souza (2008, p. 11), de forma semelhante aos resultados expressos no quadro acima, conclui que "[...] podemos afirmar que a densidade informacional do SN cresce com seu nível (ao menos até os de terceiro nível [...])" e "[...] a menor densidade informacional ocorre entre os SN de estrutura (D+n)". Ao correlacionar os percentuais de SN descritores aos níveis dos sintagmas,

verifica-se uma crescente a partir dos SN de nível 2, ou seja, aqueles que possuem um sintagma embutido em sua estrutura, um exemplo básico é o sintoma “os padrões de catalogação” (n2) e o sintagma “a catalogação” (1a). Desse modo, ao selecionar SN para atuarem como descritores documentais, faz-se necessário que se dê atenção ao critério do nível do sintagma nominal, afinal, precisa-se escolher os descritores que melhor representam o conteúdo informacional de um documento e, para isso, deve-se traçar estratégias que conduzam o sistema à escolha dos SN com maior potencial para isso.

Ainda acerca dos dados sobre os níveis dos SN, torna-se pertinente esclarecer que, embora os SN de nível 4 e os de nível acima de 4 tenham alcançado bons percentuais de SN descritores, essas estruturas não são interessantes para fins de representação temática da informação, visto que, na indexação semiautomática, automática e manual busca-se usar, normalmente, descritores objetivos e concisos. Sintagmas muito extensos e com alta complexidade sintática não são bons candidatos a descritores documentais. Embora no experimento desta tese, os SN de nível 1b e 1a tenham apresentado percentuais iguais, em Nascimento (2015) e em Souza (2008), encontra-se melhor desempenho dos sintagmas de nível 1b quando comparados aos de nível 1ª. Dessa forma, ao estabelecer pesos aos sintagmas e seus respectivos níveis, Souza (2008) e Souza e Raghavan (2014) atribuem peso maior aos SN de nível 1b.

Assim, considerando os dados deste experimento, bem como os dados expostos por Nascimento (2015) e Souza (2008), os sintagmas de nível 1b, 2 e 3 tendem a ser melhores candidatos a descritores documentais, quando comparados aos de nível 1 (formado pela estrutura Determinante + Nome) e os de nível 4 ou mais. Corroborando o exposto por Souza (2008), a densidade informacional do SN está diretamente relacionada ao nível de cada sintagma.

No Quadro 23, adiante, são expostas as taxas de revocação e precisão para cada nível analisado, considerando as categorias de Souza e Raghavan (2014).

Quadro 23 – Taxas de revocação e precisão de cada nível de sintagma nominal

	TAXA DE REVOCAÇÃO E PRECISÃO DE CADA NÍVEL	
NÍVEL	REVOCAÇÃO	PRECISÃO
1a	14%	10,20%
1b	8,30%	10,20%
2	36,30%	27,30%
3	23,80%	36,80%
4	10,20%	36,40%

>4	7,10%	47,50%

Fonte: Dados da pesquisa (2024)

Para alcançarem-se as taxas de revocação e precisão expressas no Quadro 23, levou-se em consideração o total de SN descritores e de não descritores. A revocação foi alcançada por meio da divisão do total de SN descritores de cada nível individualmente pelo total de SN descritores de toda a coleção (soma de todos os SN descritores de cada documento). Por sua vez, a precisão foi alcançada, dividindo-se o total de SN descritores de cada nível pelo total de SN descritores e não descritores que cada nível apresentou.

Ao analisar simultaneamente as duas taxas (revocação e precisão), é possível inferir melhores desempenhos para os sintagmas de nível 2 e 3. Se se considerar apenas a revocação, os sintagmas de nível 2 e 3 apresentam melhores índices, já na análise apenas da precisão, os SN de nível 2, 3, 4 e >4 apresentam altos índices.

Os dados expostos nos quadros anteriores (quadros 21, 22, 23), acerca do nível do SN, permitiram inferir que o nível do SN deve ser levado em conta como recurso para subsidiar a seleção de SN descritores dentre uma variedade de SN extraídos de um documento indexado semiautomaticamente, uma vez que a estrutura e o nível do SN está vinculado à sua capacidade de representação temática, ou seja, se se constatar que SN de nível 2 apresentam melhor comportamento quando comparados aos de nível 1, por exemplo, é justificável que se use esse critério na indexação semiautomática de documentos por meio de SN.

Ademais, outro critério frequentemente utilizado em pesquisas sobre a automação da indexação é o uso de listas de termos (palavras) vazias de conteúdo e frequentemente comum nos textos acadêmicos: chamados de *stop words*.

Nas seções que seguem, discute-se sobre tal critério.

Critério: Uso de *stop words* de SN menos relevantes

É comum encontrar nos textos palavras ou grupos de palavras que são generalistas demais e possuem pouco poder discriminatório, ou seja, em nada contribuem para fins de representação do conteúdo do documento no qual estão inseridas. Essas palavras compõem uma lista de *Stop Words*. Normalmente são sintagmas bastante utilizados na escrita formal e que se repetem nas diversas manifestações textuais.

Por exemplo, em textos científicos, é comum a presença de expressões como “esta pesquisa”, “a metodologia da pesquisa”, “o presente trabalho” etc. Essas palavras são bastante

genéricas e com pouca densidade informacional do ponto de vista da indexação e representação temática da informação. Nas pesquisas de indexação automática e semiautomática, essas palavras (expressões) são desconsideradas pelos motivos já explicitados. Da mesma forma ocorre com os Sintagmas Nominais que são extraídos automaticamente, mas em nada contribuem para a representação do documento de que foram retirados. Por esse motivo, é viável que se considere eliminar os Sintagmas não descritores que se encaixam numa possível lista de *stop words*.

Isto posto, procedeu-se a aplicação do critério "uso de *stoplist* de SN menos relevantes e com pouca densidade informacional" no experimento desta tese. O referido critério foi aplicado manualmente a todos os 1.212 sintagmas submetidos à indexação semiautomática. É importante ressaltar que a criação de listas de *stop words* deve ser considerada com bastante cuidado, pois cada domínio apresenta uma maneira distinta de apresentar, ou seja, uma maneira diferente de escrita, embora haja termos comuns em diversos domínios no âmbito da escrita científica.

Na Figura 22, é possível verificar o percentual de SN descartados com a eliminação de estruturas que, embora possuam a estrutura de um sintagma nominal, em pouco contribuem para a representação temática de documentos.

Figura 22 – Eliminação de SN que em pouco contribuem para a representação temática de documentos

Fonte: Dados da pesquisa (2024)

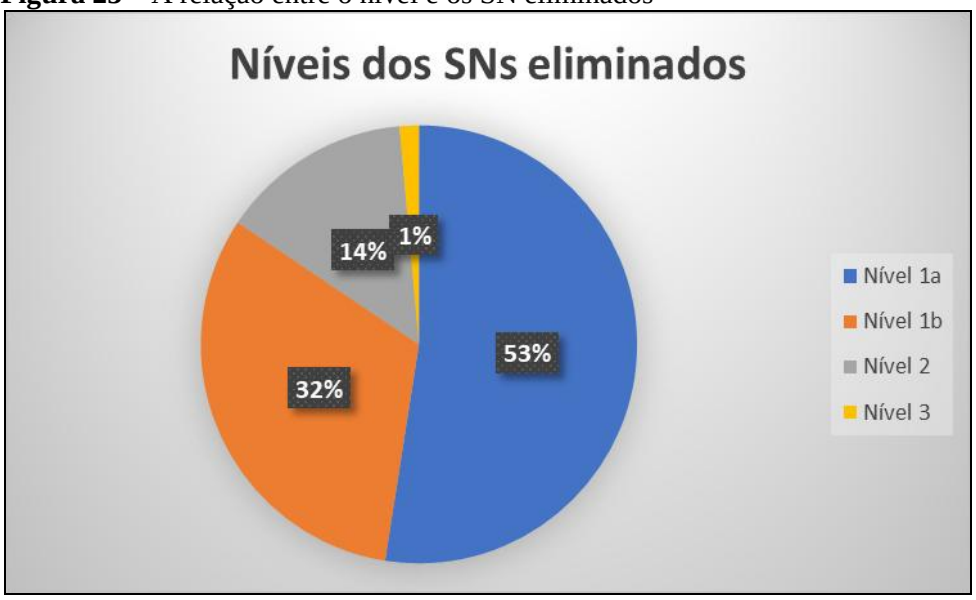
Durante o processo de aplicação dos critérios utilizados neste experimento, cada SN foi analisado manualmente com o intuito de verificar a sua relevância semântica para fins de

construção de uma lista de *stop words*. Nesse sentido, conforme se aplicavam cada critério, sinalizavam-se os sintagmas que comporiam a referida lista, ou seja, como os termos frequentes em textos científicos não são relevantes para indexação e recuperação de documentos formam a lista de *stop words*.

De 1.212 sintagmas nominais, a análise manual de cada estrutura sintagmática permitiu eliminar 141 sintagmas[32], os quais compuseram a lista de *Stop Words*, no caso específico desta tese, considerando os Sintagmas Nominais como candidatos a descritores documentais. A análise manual dos 141 sintagmas permitiu inferir que, de fato, a criação de *stop words* em propostas de indexação semiautomática contribui efetivamente para a eliminação de sintagmas com pouca densidade informacional, ou seja, com pouca relevância conceitual para fins de representação temática da informação.

A eliminação desses sintagmas economiza tempo no processamento e aplicação dos outros critérios nos documentos indexados, deixando a lista de sintagmas candidatos a descritores documentais mais refinada e condensada. Boa parte dos SN eliminados possuem nível 1 (A e B), conforme pode ser visto na Figura 23:

Figura 23 – A relação entre o nível e os SN eliminados



Fonte: Dados da pesquisa (2024)

Mais da metade dos sintagmas eliminados são do nível 1 (A e B), ou seja, possuem estrutura simples, geralmente formado por "determinante + nome", como, "este artigo", "a

[32] Encontram-se no Apêndice A desta tese os sintagmas que compuseram a lista de *Stop Words*.

pesquisa", "o presente trabalho", "este estudo" etc. A lista com os SN eliminados encontra-se no apêndice A, no final desta tese.

Apesar de ser relativamente curta, devido ao *corpus* trabalhado, esta lista serve como ponto de partida para a construção de outras listas ou condensação com listas utilizadas neste mesmo domínio em outras pesquisas, contribuindo, assim, diretamente para propostas de indexação semiautomática, eliminando palavras ou expressões sem relevância conceitual para fins de representação temática da informação.

Os dados no Quadro 24, de modo mais preciso, corroboram a viabilidade de aplicação desse critério na indexação semiautomática por meio de Sintagmas Nominais.

Quadro 24 – Percentual de SN descritores e não descritores eliminados pelo critério "Eliminação de *Stop Words*"

CRITÉRIO DE ELIMINAÇÃO DE SN COMO STOP WORDS			
	TOTAL DE SN DESCRITORES	TOTAL DE SN NÃO DESCRITORES	PERCENTUAL DE SN DESCRITORES
QUANTIDADE DE SN QUE NÃO ATENDEM AO CRITÉRIO	263	808	**24,5%**
QUANTIDADE DE SN QUE ATENDEM AO CRITÉRIO	1	140	**0,7%**
TAXAS DE REVOCAÇÃO E PRECISÃO			
REVOCAÇÃO: 99,6%			
PRECISÃO: 24,5%			

Fonte: Dados da pesquisa (2024)

Os dados do Quadro 24 mostram boas taxas de revocação e precisão, uma vez que aquelas acima de 90% e estão com um percentual considerável. De 141 sintagmas que atenderam ao referido critério, ou seja, que foram categorizados como sintagmas com pouca densidade informacional e comumente encontrado em textos acadêmicos, apenas um constituía um sintagma nominal descritor, representando apenas 0,7% de SN descritor eliminado pelo referido critério. Isso mostra a total viabilidade de se criar listas de *stop words* em uma indexação semiautomática por meio de Sintagmas Nominais.

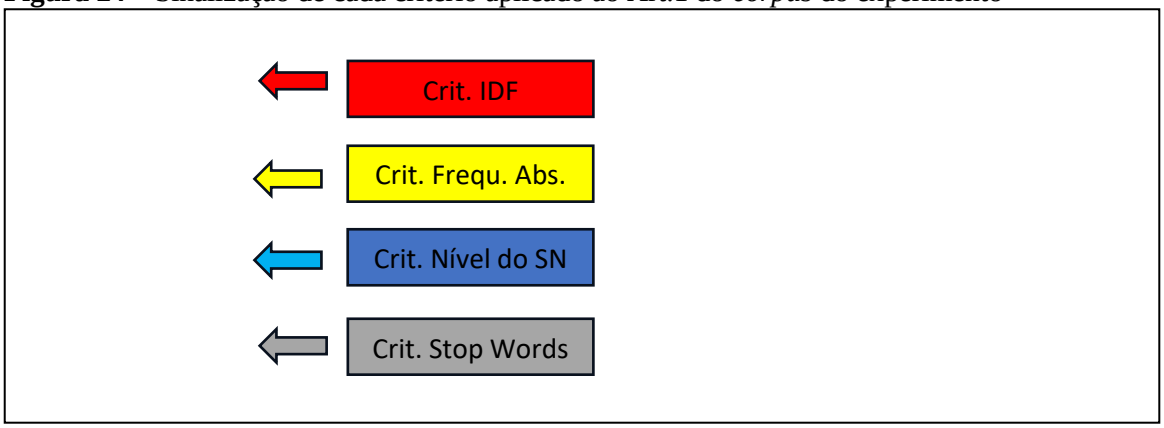
Análise detalhada da etapa 7 – Análise da relevância semântica dos SN e da Etapa 8 – Analisar os termos extraídos e que foram classificados como sintagmas autorizados[33]

Após a submissão de todos os documentos ao processo de indexação semiautomática, representado pelas etapas anteriores, foi possível perceber que, de fato, os sintagmas nominais possuem um potencial semântico diferenciado, uma vez que, ainda que sejam extraídos de seus contextos maiores, constituem-se em unidades portadoras de informação. Significa dizer que essas unidades sinalizam com precisão, porque se eximem da possibilidade de apresentar vários significados possíveis, como a apalavra isolada "integração" que pode se apresentar em diferentes contextos semânticos. Contudo, ao usarmos o sintagma "integração de dados", tal termo adquire um sentido particular porque está contextualizado.

Ademais, as taxas de revocação e precisão expostas de forma detalhada, nas seções anteriores, ratificam a ideia de que os sintagmas nominais, quando bem selecionados e refinados por meio de critérios, apresentam-se como unidades viáveis para fins de representação temática da informação em propostas semiautomáticas de indexação.

A título de exemplificação da aplicação conjunta dos critérios e da análise semântica dos sintagmas nominais autorizados, ou seja, dos que restaram após a aplicação dos critérios de seleção e que passarão pela análise do indexador humano, expõe-se abaixo um quadro com os sintagmas nominais extraídos do art. 1 (Quadro 25). Em seguida, apresenta-se a aplicação dos referidos critérios (Quadro 26, 27, 28 e 29) e, por fim, a lista final de termos autorizados (Quadro 30) para o documento em questão.

Figura 24 – Sinalização de cada critério aplicado ao Art.1 do *corpus* do experimento



Fonte: Desenvolvido pelo autor (2024)

[33] A Etapa 7 buscou analisar a relevância semântica dos SN, bem como a viabilidade dos critérios aplicados na Etapa 6. Por fim, a Etapa 8, intrinsicamente ligada à anterior, buscou analisar os termos extraídos e que foram classificados como sintagmas autorizados.

A Figura 24 busca tornar mais fácil a visualização de aplicação de cada critério conjuntamente ao Art. 1 do *corpus* deste experimento. No Quadro 25, são expostos todos os 65 sintagmas identificados e extraídos automaticamente do referido artigo. Os sintagmas marcados em verde são aqueles que são iguais às palavras-chave atribuídas pelos autores do documento ou que contêm as palavras-chave em suas estruturas.

Ao lado de cada sintagma, consta uma seta colorida referente a cada critério. Essa seta indica que tal sintagma foi eliminado pelo referido critério.

Quadro 25 – SN extraídos do Art. 1, que constituíram o *corpus* do experimento desta tese

estudo da formação ideológica				
a formação ideológica				
formação discursiva do GT2 DO Enancib				
GT2 do Enancib				
o Enancib	←			
organização	←	←		
representação do conhecimento	←			
o conhecimento	←	←		
os estudos sobre organização				
organização	←	←		
uma relação histórica e epistemológica com a ciência da informação				
a ciência da informação	←	←		
o grupo de trabalho 2 do encontro nacional de pesquisa e pós-graduação em ciência da informação				
encontro nacional de pesquisa e pós-graduação em ciência da informação				
pós-graduação em ciência da informação				
as posições discursivas				
os trabalhos desenvolvidos pelos pesquisadores				
os pesquisadores	←	←		
este grupo	←	←		
os aportes	←	←		
a análise do discurso				
origem francesa				
seu precursor Michel Pêcheux				
Michel Foucault	←			
os dois principais autores	←			
marco teórico desta pesquisa	←			
esta pesquisa	←	←	←	
auxílio	←			
a análise discursiva dos trabalhos				
os trabalhos	←	←		

o software linguístico sketch engine				
o corpus analisado				
os trabalhos publicados				
o grupo de trabalho de organização e representação do conhecimento				
os anos de 2014 a 2019				
duas formações discursivas diferentes				
a organização do conhecimento				
a organização da informação				
ambas influenciadas por suas próprias epistemologias e prismas				
suas próprias epistemologias e prismas				
prismas				
a relação da organização com a ciência da informação				
a organização com a ciência da informação				
a ciência da informação				
o contexto do grupo de trabalho 2				
o grupo de trabalho 2				
a formação discursiva de organização da informação				
organização da informação				
uma relação discursiva influenciada pela ligação histórica da ciência da informação com a biblioteconomia				
a ligação histórica da ciência da informação com a biblioteconomia				
a ciência da informação				
a biblioteconomia				
um posicionamento ideológico pautado no positivismo				
o positivismo				
a formação discursiva de organização do conhecimento				
organização do conhecimento				
dois discursos				
um positivismo-institucional				
o pragmatismo				
a aparente influência de dois pesquisadores da área da organização do conhecimento				
dois pesquisadores da área da organização do conhecimento				
a área da organização do conhecimento				
a organização do conhecimento				

Fonte: Dados da pesquisa (2024)

Ao aplicar o critério IDF e estabelecer como ponto de corte a eliminação de sintagmas com IDF abaixo de 1, eliminar-se-iam os sintagmas que podem ser visualizados no Quadro 26.

Quadro 26 – Sintagmas eliminados com a aplicação do critério IDF

representação do conhecimento
o conhecimento

esta pesquisa
a organização do conhecimento
a ciência da informação
organização do conhecimento

Fonte: Dados da pesquisa (2024)

Quanto ao critério de estrutura e nível de cada sintagma, ao eliminar os sintagmas mais genéricos, ou seja, os de nível 1a, eliminar-se-iam os sintagmas que podem ser visualizados no Quadro 27.

Quadro 27 – Sintagmas eliminados coma aplicação do critério nível

o Enancib
organização
o conhecimento
os pesquisadores
este grupo
os aportes
Michel Foucault
esta pesquisa
auxílio
os trabalhos
prismas
a biblioteconomia
o positivismo
dois discursos
o pragmatismo

Fonte: Dados da pesquisa (2024)

Quanto ao critério de frequência de ocorrência, se se eliminassem os sintagmas que se repetem duas ou mais vezes, eliminar-se-iam os sintagmas que podem ser visualizados no Quadro 28.

Quadro 28 – Sintagmas eliminados com o uso do critério frequência de ocorrência

a organização do conhecimento
Organização

a ciência da informação

Fonte: Dados da pesquisa (2024)

Ao se aplicar o critério de *stop words*, ou seja, sintagmas pouco relevantes para fins de representação temática e que são comumente encontrados em textos acadêmicos, eliminar-se-iam os sintagmas que podem ser visualizados no Quadro 29.

Quadro 29 – Alguns sintagmas eliminados com o uso de *stop words*

esta pesquisa
este grupo
marco teórico desta pesquisa
o corpus analisado
os aportes
os dois principais autores
os pesquisadores
os trabalhos
os trabalhos publicados

Fonte: Dados da pesquisa (2024)

A título de exemplo de aplicação conjunta dos critérios de seleção expostos no experimento desta tese, temos o artigo 1, que constituiu o *corpus* do referido experimento. Inicialmente foram identificados e extraídos 63 sintagmas nominais, dos quais, ao final da aplicação dos critérios, restaram 35 sintagmas. Destes, 15 considerados sintagmas nominais relevantes (por serem semelhantes ou conterem as palavras-chave dos autores dos documentos) foram preservados, ou seja, não foram eliminados pelos critérios supracitados.

Quadro 30 – Sintagmas restantes após a aplicação dos critérios de seleção (refinamento)

estudo da formação ideológica	o software linguístico sketch engine
a formação ideológica	o grupo de trabalho de organização e representação do conhecimento
formação discursiva do GT2 DO Enancib	os anos de 2014 a 2019
GT2 do Enancib	duas formações discursivas diferentes
os estudos sobre organização	a organização da informação
	ambas influenciadas por suas próprias epistemologias e prismas
uma relação histórica e epistemológica com a ciência da informação	suas próprias epistemologias e prismas
	a relação da organização com a ciência da informação
o grupo de trabalho 2 do encontro nacional de pesquisa e pós-graduação em ciência da informação	a organização com a ciência da informação
encontro nacional de pesquisa e pós-graduação em ciência da informação	o contexto do grupo de trabalho 2
pós-graduação em ciência da informação	o grupo de trabalho 2

as posições discursivas	a formação discursiva de organização da informação
os trabalhos desenvolvidos pelos pesquisadores	uma relação discursiva influenciada pela ligação histórica da ciência da informação com a biblioteconomia
a análise do discurso	a ligação histórica da ciência da informação com a biblioteconomia
origem francesa	um posicionamento ideológico pautado no positivismo
seu precursor Michel Pêcheux	a formação discursiva de organização do conhecimento
a análise discursiva dos trabalhos	um positivismo-institucional
	a aparente influência de dois pesquisadores da área da organização do conhecimento
	dois pesquisadores da área da organização do conhecimento
	a área da organização do conhecimento

Fonte: Dados da pesquisa (2024)

É importante ressaltar que os quadros 26, 27, 28, 29 e 20 buscam apenas evidenciar a aplicação conjunta dos critérios como forma de verificar a lista de termos autorizados para o documento do qual tais sintagmas foram identificados, extraídos e selecionados.

É possível perceber que boa parte dos sintagmas (destacados em verde) considerados relevantes por serem semelhantes às palavras-chave dos autores ou por contê-las em suas estruturas permaneceram inalterados e não foram eliminados. Ademais, outros sintagmas permaneceram como autorizados, embora não sejam iguais às palavras-chave dos autores do documento. No caso do artigo 1, ainda restaram 35 sintagmas, dentre os quais o indexador humano terá a função de analisá-los e escolher os que serão selecionados para representar o documento em questão.

Um aspecto que dever ser considerado em propostas de indexação semiautomática por meio de sintagmas nominais é a apresentação dos sintagmas autorizados de forma ranqueada, ou seja, pontuada. Para isso, pode-se realizar cálculos que pontuem cada sintagma conforme os critérios utilizados para refiná-los, por exemplo, o sintagma que apresenta um IDF menor deve ter um peso maior do que o que apresenta idf alto. Da mesma forma ocorre com a frequência normalizada e os outros critérios.

No exemplo acima (Quadro 30), em relação ao critério “estrutura e nível do SN”, não pontuamos os SN de nível 1b e 2 com maior peso do que os de nível 3 e 4, mas, ao desenhar uma proposta de indexação semiautomática, é pertinente que se considere a possibilidade de pontuar cada sintagma para que, ao final, eles apareçam em uma lista ranqueada, facilitando a validação de cada SN pelo indexador humano.

Foi possível constatar que, de fato, a aplicação dos critérios de seleção resulta em SN com melhores potenciais para fins de apresentação temática, os quais no final do processo deve ser validado e escolhido pelo indexador humano. Com base em todo o exposto, foi

possível levantar algumas diretrizes que devem ser consideradas na elaboração de uma proposta de indexação semiautomática baseada em sintagmas nominais. Tais diretrizes são expostas na seção que segue.

6 DIRETRIZES PARA A INDEXAÇÃO SEMIAUTOMÁTICA BASEADA EM SINTAGMAS NOMINAIS

Como forma de nortear a prática da indexação semiautomática baseada em SN, buscou-se expor nesta seção as diretrizes, os caminhos a serem considerados no desenho de qualquer proposta de indexação semiautomática que se baseie em SN como fontes de representação da informação. Dessa forma, as diretrizes expostas nesta seção foram fundamentadas em: estudos já realizados que envolveram a indexação automática e a semiautomática; utilização de Sintagmas Nominais como fontes de informação para indexação; critérios exaustivamente utilizados com o intuito de refinar cada vez mais os termos descritores de documentos em sistemas de informação; e fundamentos teóricos presentes na literatura de representação temática da informação, sobretudo no que diz respeito à indexação automática e semiautomática.

A pesquisa realizada, durante o mestrado, pelo autor desta tese também contribuiu diretamente para que se chegasse a essas diretrizes. A realização de experimentos no referido curso e agora no doutorado forneceram subsídios para a reflexão de diretrizes a serem consideradas em propostas de indexação semiautomática por meio de SN. É imperativo ressaltar que tais diretrizes constituem uma sugestão de percurso a ser tomado no que se refere a modelos de indexação semiautomática. Logo, faz-se necessário considerar que adequações, inserções de outros critérios, uso de outros instrumentos terminológicos, por exemplo, podem se fazer necessários consoante os objetivos propostos ou o domínio a que se refere a proposta de indexação.

As diretrizes levantadas nesta tese foram agrupadas, considerando as etapas da indexação automática por meio de SN associadas à parte manual, essencial a propostas semiautomáticas. Nascimento (2015) e Corrêa e Celerino (2019) propuseram as etapas da indexação totalmente automática por meio de SN. Tomando como referência tais etapas, é possível vinculá-las em uma proposta semiautomática, em que a parte automática de identificação, extração e seleção de SN é complementada pela parte manual, esta última sendo acrescentada à proposta inicial daqueles autores. A incorporação da etapa manual apoia-se no fato de que, por ser uma proposta de indexação semiautomática, há a intervenção do indexador humano para, conforme Lancaster (2004) e Moreiro González (2004), validar os termos selecionados automaticamente ou, conforme Pinto (2001), fazer ajustes e/ou complementações nos descritores selecionados de forma automática.

A indexação semiautomática, por meio de SN, apoia-se nos benefícios advindos da indexação automática, dos estudos acerca dos sintagmas nominais e do olhar do indexador humano. Visto como indexação automática, por si só, a indexação semiautomática já evidencia uma economia de tempo. A indexação baseada em SN evidencia maior preocupação com os descritores contextualizados e potencialmente informativos e, por fim, o indexador humano escolhe efetivamente os SN que representarão o documento indexado.

Convencionalmente, a indexação semiautomática é visualizada pelos textos da área como sendo constituída por duas etapas: na primeira, um sistema analisa o texto e sugere descritores; e, na segunda, o indexador valida os termos, escolhendo quais serão utilizados para descrever o documento. De modo geral, a indexação semiautomática é executada pelas referidas etapas, não obstante é pertinente ressaltar que tais etapas, na prática, podem ser desmembradas em outras etapas. Por exemplo, quando se afirma que a primeira etapa consiste na análise por parte do *software* para identificar os descritores documentais de determinado documento, é evidente que são necessárias outras atividades para que tal análise ocorra pelo *software*. Isso quer dizer que o indexador precisa escolher o *software*, precisa submeter o documento a ser indexado, escolher o formato do documento etc. Da mesma forma acontece com a seleção dos descritores por parte do *software*, ou seja, é preciso que o analisador seja programado, por meio de regras, instruções para selecionar os descritores documentais, visto que nem todos os sintagmas que se encontram em um documento são representativos do conteúdo de tal documento.

Com base no exposto, as diretrizes expostas a seguir foram organizadas e detalhadas como forma de permitir uma visualização completa das etapas que constituem a indexação semiautomática por meio de SN. Embora algumas diretrizes sejam executadas quase que de

forma simultânea, sobretudo na parte automática, o detalhamento de etapa por etapa buscou deixar mais didática a visualização de cada diretriz, bem como evidenciar a pertinência que cada etapa e sua diretriz representa no desenvolvimento de uma proposta semiautomática.

1ª Etapa - Escolha do *software* para indexação semiautomática

Os sistemas de indexação semiautomática são projetados para analisar os textos dos documentos e, com base em distintos critérios, selecionar termos descritores de tais documentos. A etapa inicial de uma proposta de indexação semiautomática é a escolha do *software* que realizará toda a parte automática da indexação, ou seja, o *software* escolhido ficará responsável pela análise do texto e, com base em critérios estatísticos, semânticos e sintáticos, fará o reconhecimento das estruturas tidas como Sintagmas Nominais, da identificação e da extração de tais unidades. Nesta tese, a proposta também é semiautomática, entretanto faz-se uso aqui de sintagmas nominais no lugar de palavras isoladas e destina-se a um domínio específico.

Há analisadores sintáticos (*softwares*) que, conforme as possibilidades de acesso, realizam a identificação, seleção e extração dos SN. No experimento realizado nesta tese, optou-se, conforme descrito na subseção "**4.3.3 Etapa 3 - Submissão dos textos ao *software* para identificação dos SN**", pelo PALAVRAS por apresentar bom desempenho em pesquisas anteriores.

Assim, a escolha do *software* que fará a análise do texto e, consequentemente, a extração dos SN será crucial para todo o processo semiautomático, visto que os SN que estarão dispostos para o indexador no final do processo serão determinados pelo desempenho do referido *software*. A extração de palavras isoladas já vem sendo utilizada com bastante frequência em propostas automáticas e semiautomáticas, entretanto, no tocante à extração de SN, há menos estudos e sistemas que fazem uso dessas unidades como descritores documentais. Isso indica que tais estudos já evidenciam o potencial dos SN para fins de representação e recuperação da informação, sobretudo quando comparados com as palavras isoladas.

Após a escolha do *software*, passa-se para a segunda etapa, a qual diz respeito à escolha do *corpus* a ser indexado.

2ª Etapa – Escolha do *corpus* documental

Neste momento, o indexador irá escolher o conjunto de documentos a serem indexados semiautomaticamente naquele momento, igualmente como ocorre na indexação manual (intelectual). Após a seleção do material a ser indexado, passa-se para a terceira etapa, que está descrita a seguir.

3ª Etapa – Submissão do documento ao *software* de indexação semiautomática

A forma como será feita a submissão do documento ao *software* irá variar conforme o próprio analisador, pois é ele que definirá a forma de entrada do documento a ser analisado, ou seja, em qual formato o texto deve estar para ser indexado automaticamente. Normalmente o analisador sintático recebe o documento em formato TXT, como o PALAVRAS utilizado nesta tese, o *software* PyPLN etc.

Considerando uma indexação feita com título, subtítulo (quando houver), resumo e principais partes de um documento e não o documento completo, esta etapa não demanda muito tempo, principalmente porque a conversão desses elementos em formato de texto analisável pelo *software* é feita rapidamente de forma automática. Após a conversão do formato do documento para o formato aceito pelo sistema, submete-se o texto ao *software*. Por exemplo, o LX-*Parser* analisa a sentença até o primeiro ponto final encontrado, sendo necessário que o título do resumo e cada sentença do corpo do resumo seja submetido um a um. Já o PALAVRAS e o OGMA, por exemplo, analisam de uma só vez. Feita a submissão, inicia-se a parte automática de processamento do texto.

4ª Etapa – Identificação dos SN através das subetapas de "etiquetagem" e de "cotejamento dos léxicos etiquetados com as regras dos SN"

Os *softwares*, no processo de indexação semiautomática, assumem a função de selecionar os termos candidatos a descritores para cada documento analisado. Essa seleção é feita por meio de regras e programações repassadas ao *software*. Inicialmente é feita a etiquetagem de cada palavra analisada, em seguida faz-se a identificação das estruturas classificadas como sintagmas nominais, essa identificação leva em consideração as regras de constituição de SN e, por fim, faz-se a extração desses SN. Por meio dessas tarefas, os *softwares* classificam cada palavra conforme a classe gramatical a que pertence, realizam a análise sintático-semântica de todas as estruturas e extraem os SN, mostrando-as em listas, por exemplo.

A tarefa inicial (etiquetagem) é de extrema importância, pois é a partir dela que as outras etapas serão desenvolvidas, obtendo ou não sucesso. A etiquetagem (classificação morfológica) das palavras é uma atividade que requer cuidado ao ser executada em sistemas de indexação, uma vez que uma mesma palavra pode, conforme o contexto em uso, pertencer a distintas das classes gramaticais. Se, na linguagem humana, algumas vezes é difícil essa categorização, imagine para sistemas computacionais. A correta classificação das palavras, ou seja, a sua adequada etiquetagem das palavras tem um papel fundamental, visto que é, a partir dessa etiquetagem, que os sintagmas nominais são identificados ou não. A homonímia e a polissemia[34] são fenômenos da língua com os quais a indexação automática precisa lidar constantemente para que alcance a etiquetagem adequada das palavras nos contextos em que elas se encontram e, por conseguinte, sejam identificados os sintagmas nominais. Um exemplo claro é demonstrado na Figura 12, em que a palavra "estudo" foi equivocadamente categorizada como "verbo" (conjugado na 1ª pessoa), quando deveria ser classificada como nome (substantivo).

O reconhecimento das categorias gramaticais é um dos objetivos do Processamento de Linguagem Natural (PNL). Tal processamento é fundamental para a identificação, por exemplo, de sintagmas nominais, visto que tais sintagmas possuem, em sua formação substantivo, pronomes ou palavras substantivadas. Logo, a atribuição correta das categorias gramaticais às palavras, presentes em um documento, é responsável, consequentemente, pela identificação de todas as formas de sintagmas, sejam verbais, preposicionais, nominais etc.

Ao lado da categorização das palavras em classes gramaticais, os analisadores sintáticos (*parsers*) realizam as análises das palavras em suas relações umas com as outras, ou seja, estuda as relações entre os constituintes de uma frase, os quais podem ser chamados de sintagmas verbais, nominais, adverbiais etc.

O processo de identificação automática e a extração manual dos SN, realizados no experimento desta tese, evidenciaram que, embora o *software* seja capaz de reconhecer os SN presentes no *corpus* de estudo, alguns equívocos de categorização levaram à omissão de SN e à identificação equivocada de possíveis SN candidatos a descritores documentais. Alguns problemas, como identificação de SN iniciando com conjunções, preposições, podem ser

[34] A análise da distinção entre a homonímia e a polissemia não é tarefa fácil. Bechara (2020, p. 604) enfatiza que estudiosos apontam a dificuldade de nem sempre se poder distinguir a polissemia da homonímia. Para isso, têm sido propostos critérios para aclarar se se trata de uma mesma palavra com dois ou mais significados diferentes (polissemia) ou de duas palavras distintas com idênticos fonemas (homonímia). Nessa seara, Pestana (2023) aponta que a identificação da classe gramatical é um critério para diferenciar polissemia de homonímia, se pertencerem à mesma classe gramatical, estar-se diante de polissemia, se se estiver diante de classes gramaticais distintas, há a presença da homonímia.

sanados com melhoramento do referido *software*, sobretudo com ênfase às regras gramaticais de formação de sintagmas.

Em suma, é possível condensar os problemas apresentados pelo PALAVRAS no momento de identificação de SN nas seguintes situações: equívocos de etiquetagem; omissão de SN; equívocos na etiquetagem identificando erroneamente SN; identificação de nomes, mas não como SN; problemas relacionados à identificação de SN iniciando com preposições e conjunções; e problemas com análise de numerais e símbolos.

Como a classificação errônea de uma unidade linguística (palavra ou grupo de palavras) em SN e a omissão na identificação de SN estão ligadas à correta categorização das palavras nas classes gramaticais, fica evidente que o melhoramento do *software*, no reconhecimento adequado das palavras em suas categorias gramaticais, reduzirá substancialmente os inconvenientes associados ao PALAVRAS no que se refere à identificação de SN. Os desafios para os analisadores morfológicos e sintáticos não são poucos, sobretudo no que diz respeito a fenômenos próprios da linguagem natural, especialmente, a polissemia, a homonímia. Por exemplo, pode-se pontuar a palavra "periódico" que pode ser "adjetivo" e "substantivo" e que, conforme a Figura 10, foi analisada como adjetivo, quando não o deveria ser, uma vez que a ocorrência de tal palavra no excerto "revistas e periódicos especializados em Arquivologia" configura-se como um "substantivo". Dentre as estratégias para esses fenômenos da linguagem natural, é possível que os *softwares* aprimorem continuamente a etapa de etiquetagem das palavras, atualizem constantemente seus vocabulários, bem como revisem as regras de formação de estruturas linguísticas, abrangendo as diferentes possibilidades de classificação morfológica de uma mesma palavra, por exemplo.

Como forma de lidar com a ambiguidade na etiquetagem das palavras, a estratégia utilizada pelo *software* OGMA, segundo Maia (2008), é formar uma lista com todas as combinações encontradas de etiquetas gramaticais para uma frase e, em seguida, submeter cada frase etiquetada às regras para extração dos SN, depois, submeter os SN encontrados a uma lista geral de sintagmas nominais da frase, eliminando, assim, a duplicidade.

Após a etiquetagem de cada unidade léxica, é feita identificação dos SN por meio do cotejamento entre os léxicos etiquetados com as regras de formação dos sintagmas nominais. Silva (2014), baseado em Miorelli (2001) e Santos (2005), elaborou um quadro com um conjunto de regras de formação dos SN, conforme o Quadro 31**.**

Quadro 31- Elementos 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 3: DET + N	A informação
Regra 4: N	Informação
Regra 5:DET + N + DET + N + MOD	A filosofia e a ciência juntas
Regra 6:DET + DET + N + MOD	A minha recuperação da informação
Regra 7:MOD + N + MOD	Grande área da informação
Regra 8:DET+ DET + N	Uma certa área

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

DET é a abreviação para "determinante"; MOD, a abreviação para "modificador"; N, a abreviação para "nome". Os modificadores, diferentemente dos determinantes que sempre se encontram antepostos ao núcleo, podem estar antepostos ou pospostos ao núcleo, possuindo a função de caracterizar, enfatizar, quantificar e sendo compostos por adjetivos, locuções adjetivas, advérbios e locuções adverbiais (Silva e Corrêa, 2015). Após a identificação dos SN, passa-se para a extração desses sintagmas, também de forma automática.

5ª Etapa – Extração dos SN de cada documento

A extração consiste em exportar os SN identificados para um recurso passível de visualização por parte do indexador, por exemplo, em uma lista com os SN. Essa forma de saída dos SN depende de cada *software*. Um exemplo de *software* que executa, além de identificar, a extração é o OGMA.

Segundo Maia (2008), o *parser* PALAVRAS (Bick, 2000), é um programa que faz parte de um conjunto de várias ferramentas multilíngues, que compõe o Visual Interactive Syntax Learning (VISL). Nesse programa, o usuário submete um texto e, após a análise de seus elementos, recebe o referido texto com as marcações. O *parser* PALAVRAS apresenta o resultado da análise de texto em três formatos: forma gráfica; forma arbórea; formato de representação da análise de texto. Conforme Silva e Corrêa (2015, p. 9), "[...] o VISL trabalha com uma gramática de restrições que é feita a partir de uma análise em que se observam os morfemas, os grupos de palavras e a composição da oração. Isso permite uma observação da ortografia, da semântica e da sintaxe". Feita a identificação dos morfemas, o PALAVRAS elimina as ambiguidades encontradas em cada léxico. Tal eliminação se dá por meio de uma

aplicação de regras que identificam e eliminam as possibilidades de estruturas sintáticas inexistentes.

O PyPLN é um *software* que faz a etiquetagem, a identificação e a extração dos SN em análises de textos. Essa plataforma de Processamento de Linguagem Natural faz uso do PALAVRAS como analisador sintático-semântico e permite realizar diferentes análises linguísticas de textos.

A extração dos SN, de modo geral, consiste em mostrar os SN fora dos textos ao indexador. Feita a extração, têm-se os SN, não obstante se sabe que nem todo SN presente em um texto é potencialmente informativo para representar esse texto. Fica clara a distinção do potencial informativo entre os sintagmas "representação da informação" e "a metodologia da pesquisa" retirados de um documento que trate de indexação, por exemplo. Aquele sintagma nitidamente possui maior potencial informativo do que este último. A esse potencial representativo, Souza (2005), em sua tese, usa a expressão "densidade informacional". Para o referido autor, a densidade informacional e o poder discriminatório estão relacionados a diferentes aspectos dos SN, por exemplo, à estrutura e ao nível de um sintagma, à frequência de ocorrência do sintagma etc.

No Quadro 32, são expostos os sintagmas extraídos do doc. 23 do *corpus* do experimento desta tese

Quadro 32 – A densidade informacional e o poder discriminatório dos SN

fotografias do patrimônio histórico	**Sintagma de nível 2**
o patrimônio histórico	Sintagma de nível 1
fotografias de esculturas sacras	Sintagma de nível 2
esculturas sacras	Sintagma de nível 1

Fonte: Dados da pesquisa (2024)

Ao analisar os sintagmas presentes no Quadro 32, é perceptível que a densidade semântica de um sintagma está diretamente ligada, dentre outros aspectos, ao nível do SN. Souza (2005), no contexto do *corpus* utilizado em sua tese, afirma que "A densidade informacional do SN cresce com seu nível (ao menos até os de terceiro nível). […]. A menor densidade informacional ocorre entre os SN de estrutura (D + N)". Com base no exposto, fica claro que, em uma proposta de indexação semiautomática que use sintagmas nominais, não basta a identificação e extração dos SN, é preciso que se apliquem critérios para selecionar os sintagmas com maior poder discriminatório, com maior densidade informacional, posto que cada sintagma, conforme diferentes aspectos, apresenta potencial distinto.

É imperativo ressaltar que, na prática de indexação semiautomática, a identificação, a extração e a seleção de SN, esta última exposta na 6ª Etapa, são executadas totalmente de forma automática. O detalhamento de cada etapa aqui se mostra como forma de evidenciar, de forma clara, a necessidade de cada etapa: identificação, extração e seleção.

6ª Etapa – Seleção dos SN, com base em critérios preestabelecidos nesta pesquisa

Embora possuam um potencial informativo maior do que as palavras isoladas, alguns SN possuem baixo ou não possuem valor descritivo para fins de representação temática da informação, por exemplo, o sintagma nominal "o objetivo deste trabalho" extraído do artigo 19, que constituiu o *corpus* do experimento realizado nesta tese. Isto porque em nada contribui, em termos de representação temática, para o documento que trata do "impacto das categorias de Ranganathan e Dalberg na organização do conhecimento". É nesse sentido que ganha relevância a aplicação de critérios para selecionar os melhores SN, ou seja, os sintagmas considerados potencialmente informativos e que descrevam substancialmente os documentos de que serão procuradores.

O desenvolvimento da indexação automática está intimamente ligado à utilização de critérios, uma vez que os *softwares*, aos serem preparados para selecionar termos representativos dos documentos, apoiam-se em um ou mais critérios. Conforme ressalta Vieira (1988, p. 44), "A indexação automática é um processo que pode utilizar diferentes métodos desenvolvidos para programas de computador. Essa operação, ainda segundo Robredo, é objetiva, pois utiliza sempre os mesmos programas para extração de termos significativos dos documentos". Para Moreiro González (2004, p. 3 *apud* Bufrem, 2005),

> [...] A essência do processo é a identificação automática de palavras-chave no texto pela freqüência com que aparecem e sua fundamentação teórica tem origem na lei de Zipf. Novas formulações desta Lei originaram outras técnicas de discriminação dos termos, sobre as quais discorre o autor, destacando a indexação estatística de termos por freqüência, conhecida pela sigla IDF, a Term frequency, inverse document frequency (TFIDF), o método N-grams, que modifica a lei de Zipf para possibilitar o tratamento de palavras compostas e os Stemmers, que utilizam a freqüência com que aparecem seqüências de letras no corpo de um texto para extrair a raiz das palavras. Além dessas possibilidades, as relações semânticas entre os termos lingüísticos podem ser estabelecidas por métodos de agrupamento e classificação.

No contexto dos sintagmas nominais, recentes pesquisas dedicaram-se a analisar o comportamento de critérios de seleção já aplicados a palavras isoladas. Por exemplo, o critério de "frequência absoluta de ocorrência" é bastante utilizado nos sistemas automatizados de indexação e já apresentou bom desempenho na seleção de sintagmas nominais. Trabalhos como os de Souza (2005, 2006), Souza, Alvarenga Neto e Mendes (2007), Maia (2008), Maia e Souza (2010), Lopes (2012), Souza e Raghavan (2014) e Martins (2014), Nascimento (2015), Nascimento e Corrêa (2019) utilizaram o critério de "frequência absoluta de ocorrência" na seleção de sintagmas nominais. Esse critério é comumente utilizado para a classificação do sintagma nominal como descritor ou não, sendo considerado fundamental para a seleção de sintagmas nominais. *Frequência absoluta de ocorrência*, *frequência de ocorrência normalizada do sintagma nominal em um documento, inverso da frequência de ocorrência dos sintagmas nominais em um conjunto de documentos ou corpus (IDF), Estruturas e níveis dos sintagmas nominais, Ocorrência do sintagma nominal em tesauro* etc. são alguns exemplos de critérios utilizados com o intuito de selecionar e refinar os sintagmas nominais para fins de representação e recuperação da informação.

O intuito desta seção é, com base nos estudos e experimentos já realizados, sobretudo com uso de SN, propor critérios de seleção para a indexação semiautomática por meio de sintagmas nominais. Tais critérios apoiam-se em aspectos estatísticos, semânticos e sintáticos dos textos. É importante ressaltar que os critérios expostos a seguir, quando analisados individualmente, podem apresentar comportamentos distintos a depender da área de conhecimento, bem como mostrar critérios que são úteis em várias áreas do conhecimento. Essa variação está diretamente ligada à forma como determinado domínio se comporta em termos de escrita. Por exemplo, um critério que elimina SN que possui numeral, aplicado ao domínio jurídico, certamente eliminará possíveis sintagmas relevantes, uma vez que nesse domínio é comum o uso de numerais se referindo às leis.

Torna-se necessário ressaltar que as diretrizes correspondentes aos critérios de seleção expostos aqui envolvem critérios gerais, que já foram aplicados em pesquisas e experimentos anteriores, especificamente com SN. Dessa forma, expõem-se critérios gerais que podem ser complementados por novos critérios que surjam com as próximas pesquisas sobre indexação automática e semiautomática por meio de SN. Boa parte desses critérios foi proposta para termos, palavras isoladas, no entanto, em estudos recentes tais critérios foram aplicados em SN. Dessa forma, no Quadro 33 faz-se uma adaptação para o contexto específico dos SN.

Quadro 33 – Critérios a serem utilizados na indexação semiautomática por meio de SN

Critério	Funcionamento
Frequência absoluta de ocorrência do SN	Este critério ordena os SN conforme a frequência de ocorrência no documento, considerando que, quanto maior a frequência de um SN, maior o potencial desse SN em ser representativo do documento.
Inverso da frequência dos sintagmas nominais em um conjunto de documentos ou *corpus* (IDF)	Verifica a frequência de ocorrência do SN no *corpus*, apoiando-se no fato de que a ocorrência de um determinado SN em demasia em vários documentos pode demonstrar que esse SN é comum e possui pouco poder discriminatório.
Frequência de ocorrência normalizada do sintagma nominal em um documento	Considera a extensão do documento em que se encontra o SN. Esse critério calcula a frequência absoluta de um determinado termo, dividida pela quantidade total de Sintagmas extraídos do documento. Essa frequência normalizada leva em consideração o tamanho do documento, uma vez que o comprimento de cada documento influencia a frequência de ocorrência de um determinado SN.
Estruturas e níveis dos sintagmas nominais	A estrutura e o nível do SN são proporcionais à sua densidade informacional.
Eliminação de sintagmas nominais em *stoplist* de SN menos relevantes	Busca-se, com este critério, eliminar estruturas linguísticas que, embora sejam SN, não são representativas dos documentos. A essa estrutura atribuiu-se o nome de *stoplist*, e às palavras deu-se o nome de *stop-words*. São exemplos de SN vazios em termos de representatividade: “este trabalho”, “esta pesquisa”, “o presente trabalho” etc.

Fonte: Desenvolvidos pelo autor (2024)

A associação dos critérios apresentados no Quadro 33 é essencial para a seleção dos SN com maior densidade informacional, ou seja, com alto poder discriminatório, uma vez que são complementares e abrangem diferentes aspectos dos SN. A aplicação de diferentes critérios busca chegar aos SN com maior potencial informativo para fins de representação e recuperação da informação. Tais critérios foram aplicados no experimento desta tese e apresentaram bom desempenho na seleção de adequados sintagmas nominais, contribuindo, assim, para a indexação semiautomática de documentos.

A frequência absoluta de ocorrência produz SN que, com base na repetição de ocorrência, apresentam maior potencial informativo quando comparados com SN que apresentam baixa frequência. Não obstante é preciso atentar para o fato de que um SN que apresenta altas frequências de ocorrência em um documento e na coleção como um todo tende a ter menor relevância como descritor, uma vez que não é tão específico e que os SN escolhidos para indexação possuem o papel de individualizar cada documento de toda a coleção. Assim, SN com ocorrência mais “concentrada” tendem a ser mais significativos, ou seja, valoriza-se o SN que ocorre frequentemente em poucos documentos quando comparados com o SN que aparece em demasia nos outros documentos da coleção.

A frequência absoluta de ocorrência do SN, no documento indexado, se mostra um critério relevante para a seleção de SN. Embora seja um critério fundamental para a seleção de SN, as pesquisas evidenciam que tal critério deve ser acompanhada de outras estratégias de

seleção, conforme ressaltam Correa et al. (2011, p. 20), ao apontarem que "[...] a extração de sintagmas deve ser acompanhada de estratégias de ordenação por relevância dos sintagmas, levando em conta critérios de frequência e posicionamento, semelhantemente às propostas existentes para palavras isoladas".

Borges (2009, p. 83) ressalta que "Pode-se acreditar, então, que a parceria da utilização do critério frequência absoluta de ocorrência da palavra no texto com outros critérios que consideram aspectos semânticos pode suprimir o uso de outros critérios puramente estatísticos.

Baeza-Yates e Ribeiro-Neto (2011) ressaltam que, embora seja simples, esse critério se mostra fundamental para os sistemas de recuperação da informação atuais. Quanto maior a quantidade de vezes que um determinado termo ocorre no documento, mais importante esse termo tende a ser para representar tematicamente o documento. A frequência normalizada calcula a razão entre a frequência absoluta dos SN no documento e o número total de SN que constam no documento, essa frequência busca corrigir distorções relacionadas à extensão do documento

Ao lado da frequência do termo do documento, está a frequência inversa nos documentos. Quanto mais documentos contiverem determinado termo, menos importante esse termo será para descrever qualquer desses textos, uma vez que não são tão específicos e aparecem com frequência em distintos documentos da coleção. Conforme Correa e Fujita (2024), utiliza-se DF (*Document Frequency*) para indicar a frequência absoluta de um termo, em um determinado documento, TF (*Term Frequency*) é usado para representar a frequência normalizada de um termo em um determinado documento, e a IDF (*Inverse Document Frequency*) é utilizada para indicar a frequência inversa nos documentos da coleção. Unindo-se os conceitos de TF e IDF, tem-se a formulação TF-IDF, que é usada para representar a multiplicação da frequência do termo normalizada em um documento pela frequência inversa nos documentos.

Souza e Raghavan (2014) e Nascimento (2015) constataram que a frequência normalizada alcançou bons resultados, demonstrando, assim, que esse critério contribui diretamente para a seleção de SN para a representação e recuperação da informação. O experimento realizado nesta tese também evidenciou a viabilidade de tais critérios.

A frequência normalizada calcula a razão entre a frequência absoluta dos SN no documento e o número total de SN que constam no documento. Essa frequência normalizada busca corrigir distorções relacionadas à extensão do documento. A TF-IDF busca analisar a relevância de um SN consoante a sua frequência na coleção, considerando que, quanto mais

documentos contiverem determinado SN, menos importante esse SN será para descrever qualquer desses textos.

Um SN pode aparecer com bastante frequência em um determinado documento e, ao mesmo tempo, aparecer também com frequência em toda a coleção. A IDF corresponde a uma medida estatística utilizada com o intuito de verificar o quanto um SN é importante para um documento em relação a uma coleção (*corpus*). Essa relevância aumenta proporcionalmente com o número de vezes que a palavra apareça no documento e diminua na coleção (Maia, 2008).

Ademais, é preciso aplicar um critério voltado à estrutura do SN, uma vez que, conforme as características de construção, cada sintagma apresenta potencial semântico distinto. Souza (2005) observou que a estrutura sintática dos SN está relacionada à sua relevância como descritor. Por exemplo, os sintagmas formados por "determinante + nome (D+N)" são bastante semelhantes às palavras-chave adotadas pelos autores no que se refere à densidade informacional.

Em sua tese, Souza (2005, p. 111) afirma que "a densidade informacional do SN cresce com seu nível (ao menos os de terceiro nível). [...] e a menor densidade informacional ocorre entre os SN de estrutura (D+N)". Essa densidade informacional pode ser visualizada nos SN extraídos do Artigo 22, que constituiu o *corpus* do experimento realizado nesta tese.

No quadro abaixo, é perceptível o potencial semântico mais específico de SN de nível 2, quando análogos aos de nível 1.

Quadro 34 – SN extraídos do Artigo 22, que constituiu o *corpus* do experimento desta tese

a gestão do fisco	**SN de Nível 2**
o fisco	SN de Nível 1
um modelo de arquitetura da informação	SN de Nível 3
arquitetura da informação	SN de Nível 2
a informação	SN de Nível 1

Fonte: Desenvolvido pelo autor (2024)

Quando se comparam os SN "a informação" e "arquitetura da informação", verifica-se o maior grau de especificidade deste último, uma vez que o SN "a informação" é bem mais genérico do que "arquitetura da informação".

Ao analisar o SN de nível 3 “um modelo de arquitetura da informação”, corroborou-se o que Souza (2005) evidencia acerca do crescimento da densidade informacional do SN até o de terceiro nível, ou seja, quanto maior for o nível do SN, mais delimitada será a informação representada pelo SN. Souza e Raghavan (2014), após testar diversos valores para cada Categoria de Sintagma Nominal – CNP, encontraram resultados satisfatórios com os seguintes valores que podem ser visualizados no Quadro 35.

Quadro 35 – Valores otimizados para a Categoria do Sintagma Nominal – 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**)

Ao lado dos critérios de frequência absoluta de ocorrência, frequência normalizada e inverso de frequência, a estrutura e o nível dos SN se mostram essenciais, sobretudo por abranger diretamente os aspectos semânticos e sintáticos do documento, alcançando, assim, os SN com maior poder descritivo do documento indexado.

Além dos critérios até aqui expostos, recomenda-se a utilização de outro critério: eliminação de sintagmas nominais em *stop list* de SN menos relevantes. Esse critério pode ser alcançado por meio da criação de uma lista de sintagmas nominais que frequentemente ocorrem em textos acadêmicos e que possuem pouca relevância semântica. A aplicação de tal critério é essencial, pois elimina sintagmas comuns em textos científicos, que possuem pouco valor em termos de representação temática. A utilização de uma *stop list* de SN menos relevantes contribui diretamente para refinar a seleção de sintagmas, retirando os que não serviriam como termos de indexação, limpando consideravelmente os sintagmas irrelevantes de um grupo de sintagmas.

Comparando os resultados da pesquisa de Souza e Raghavan (2014) com os da pesquisa anterior dos mesmos autores (2006), no tocante ao uso de *stop list*, verifica-se que os

resultados desta foram mais satisfatórios. Esses resultados podem estar relacionados, segundo os autores, à presença de uma lista de *stop words* para descartar SN menos relevantes, que foi utilizada nesta e não naquela. A extensão da lista de *stop words* está relacionada ao tamanho do *corpus* utilizado em cada experimento ou ainda à extensão da base de dados em que a indexação semiautomática esteja sendo realizada.

A cada documento analisado, há a possibilidade de crescimento da *stop list*, visto que podem surgir novos SN frequentes em textos científicos e com pouco valor descritivo. A utilização de *stop list* apresenta uma lista de sintagmas frequentes que possuem reduzido valor informacional, ou seja, possuem pouco poder representativo na indexação e recuperação de documentos.

7ª Etapa - Validação dos SN representativos do conteúdo do documento pelo indexador humano

Toda proposta de indexação semiautomática pressupõe a participação de um indexador humano que valida os descritores selecionados automaticamente. Essa validação se faz necessária, visto que, mesmo com todos os avanços tecnológicos, os *softwares* ainda apresentam limitações quando da seleção de sintagmas nominais que funcionem como descritores documentais.

No tocante à pertinência da associação da parte humana a propostas de indexação automática, Lancaster (2004) e Moreiro González (2004) corroboram a ideia de que os *softwares* de indexação automática estão aptos a realizar a indexação dos documentos, ressaltando, no entanto, as limitações inerentes a tal processo, as quais são supridas pela parte humana, ou seja, pela participação do olhar humano no momento de decisão final da seleção dos descritores que realmente serão utilizados como pontos de acesso ao documento indexado. Nessa seara, Martins (2014) evidencia que, no tocante ao uso de um sistema de indexação manual ou automático, quando é feita uma pesquisa exaustiva, dependendo da área de conhecimento, "[...] um ou outro sistema pode favorecer os resultados, no entanto é demonstrada uma melhoria quando os descritores atribuídos à indexação manual são combinados com um esquema de indexação automática". Por fim, Santos (2017, p. 54) ressalta que "[...] a partir da junção desses dois modelos, podem surgir grandes avanços no sentido de tratar as informações de forma mais ágil e, ao mesmo tempo, aprimorar tal atividade a partir do caráter intelectual do modelo de indexação manual.

Isto posto, nesse momento, o indexador humano irá validar os sintagmas nominais sugeridos pelo sistema de indexação semiautomática, escolhendo, assim, os que de fato serão adotados na representação temática do documento indexado. Com essa lista de SN candidatos à representação, o indexador, além de escolher os que funcionarão como descritores, poderá criar e/ou atualizar vocabulários controlados de SN com a inclusão, por exemplo, de SN que possuem potencial informativo e que foram sugeridos pela indexação semiautomática, mas que ainda não constam no tesauro.

8ª Etapa - Elaboração de lista de termos autorizados

Após as etapas anteriores, o indexador humano estará com a lista de termos autorizados à representação do documento submetido à indexação semiautomática por meio de SN. Os descritores que constam nessa lista servirão para a representação do documento no momento de entrada no Sistema de Recuperação de Informação, bem como para a representação das expressões de busca utilizadas pelos usuários.

As diretrizes sugeridas, na presente tese, apoiaram-se em:

- ✓ Pesquisas acerca da indexação automática e semiautomática;
- ✓ Propostas de indexação automática baseadas em Sintagmas Nominais;
- ✓ Estudos que evidenciam a necessidade emergente de propostas automáticas que abranjam fatores estatísticos, semânticos e sintáticos presentes em textos;
- ✓ Experimentos realizados em diferentes contextos de uso dos SN;
- ✓ Estudos que demonstram os benefícios de se associar a parte automática à manual na indexação e recuperação de documentos; e
- ✓ Experimentos realizados no curso de mestrado e doutorado do autor desta tese.

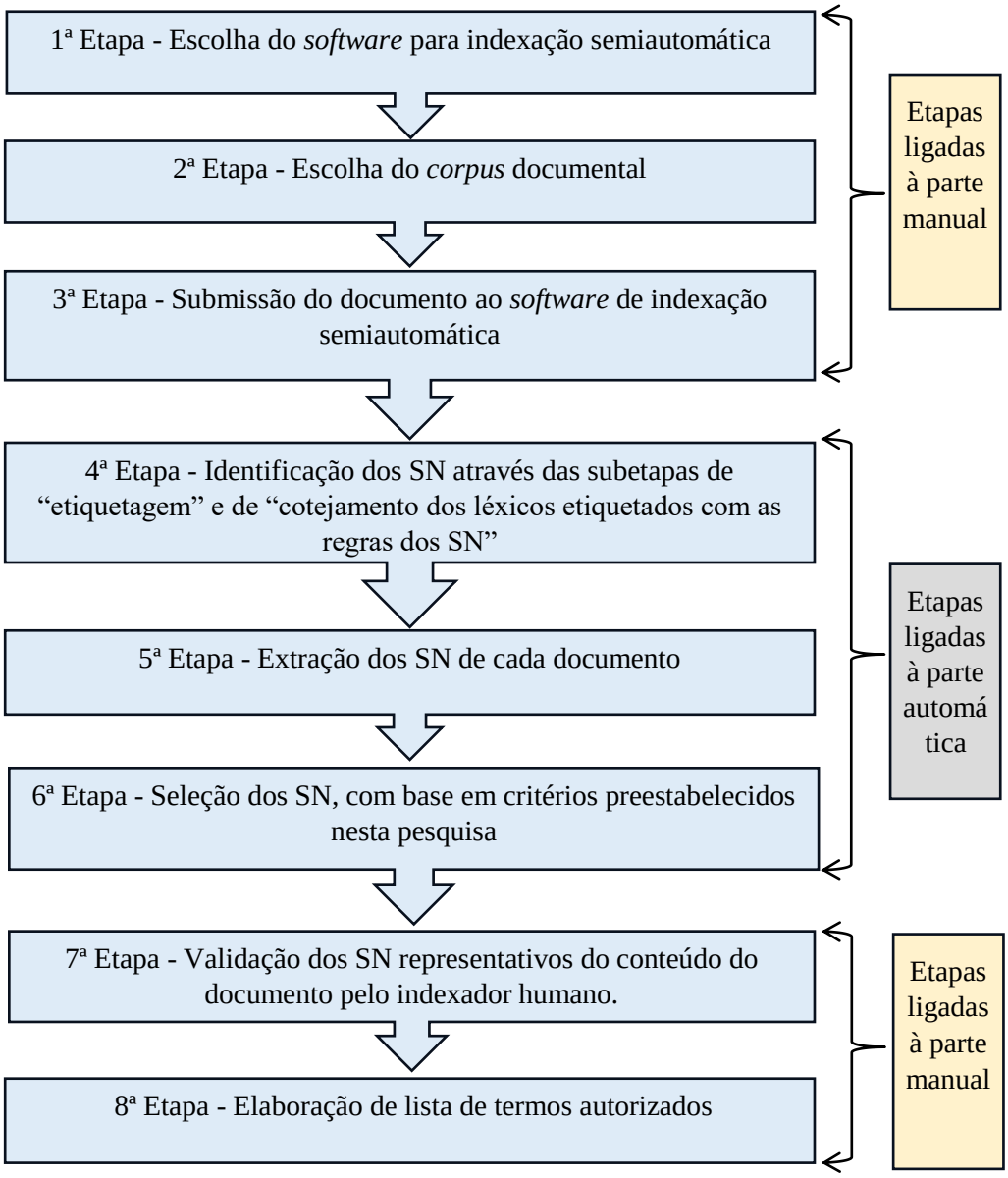
É pertinente ressaltar que as diretrizes aqui expostas assumem o caráter instrucional, buscando orientar a construção de propostas de indexação semiautomática em unidades de informação e sistemas de recuperação da informação. Tais diretrizes devem ser vistas como orientações gerais flexíveis, visto que cada domínio pode exigir um ou outro ajuste, seja na extração de SN, seja na seleção de SN etc.

O *corpus* utilizado, no experimento desta tese, foi extraído do domínio “Organização e Representação do Conhecimento”, entretanto tais orientações devem ser interpretadas como orientações gerais, que, conforme o domínio, podem ser utilizadas sem alteração, bem como

com adaptação quanto aos critérios utilizados na seleção de cada SN candidato a descritor documental.

Na Figura 25, expõem-se, de forma sintetizada, as diretrizes a serem seguidas na indexação semiautomática por meio de SN:

Figura 25 – Síntese das diretrizes à indexação semiautomática por meio de SN



Fonte: Desenvolvido pelo autor (2024)

As etapas apresentadas na Figura 25 compõem uma síntese das diretrizes a serem seguidas em uma proposta de indexação semiautomática por meio de Sintagmas Nominais. Tais diretrizes não encerram em si mesmas, mas funcionam como um norte, uma vez que podem ser feitas adaptações a cada uma das etapas sugeridas. Tais diretrizes se dividem em dois momentos gerais: a etapa totalmente automática e a manual, esta sendo executada no início e no final do processo de indexação.

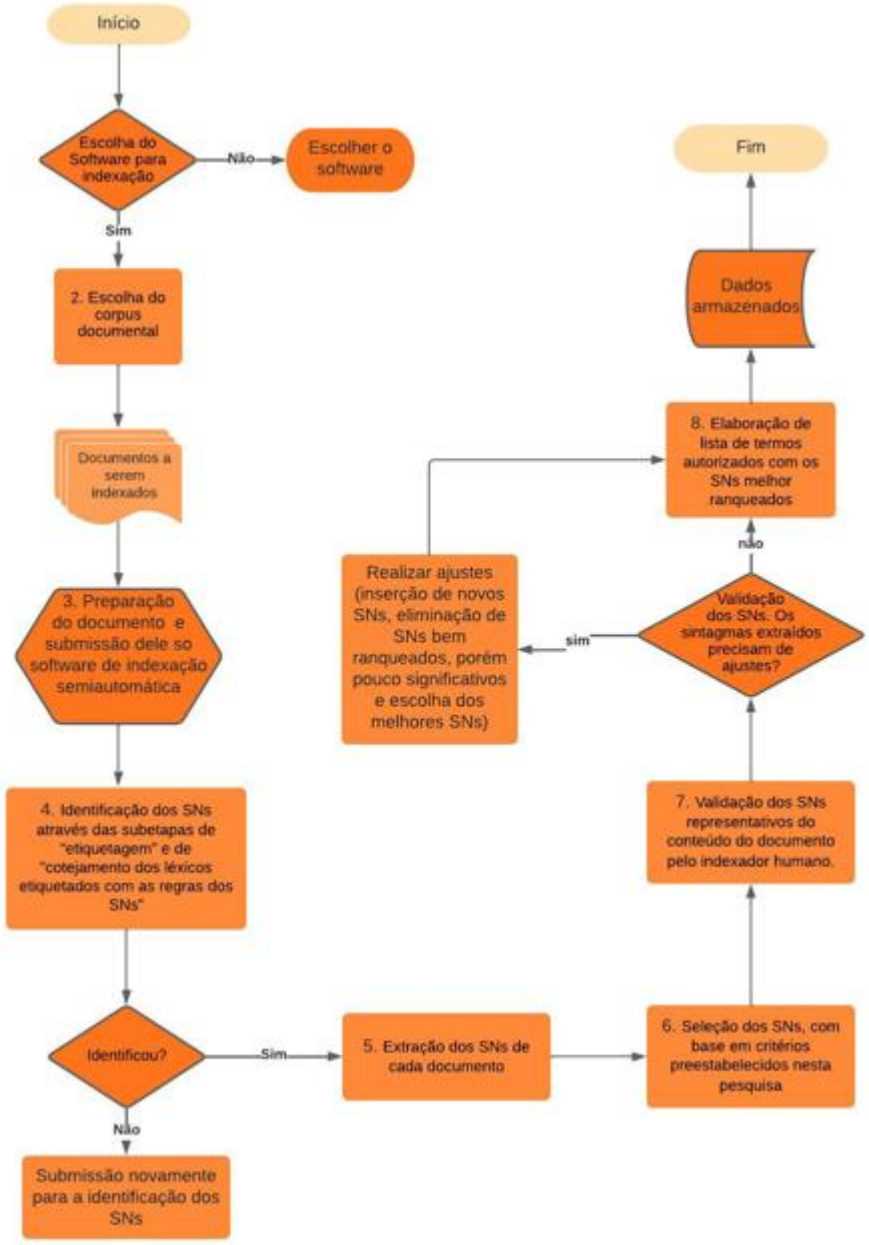
No que se refere à Etapa 3, é pertinente ressaltar que a operacionalização de tal etapa deverá considerar dois aspectos: o *software* a ser utilizado para a identificação, extração e seleção dos SN; e a fonte de informação a ser utilizada na indexação. Ou seja, deverá considerar o documento na íntegra ou as suas partes mais relevantes. Por exemplo, no experimento desta tese, os textos submetidos à indexação foram convertidos em formato .txt. Em síntese, a Etapa 3 se ajustará conforme a fonte a ser utilizada para indexação e o *software* utilizado em tal processo.

Merece destaque a Etapa 6 que, no experimento desta tese, se debruçou sobre quatro critérios, os quais apresentaram bom desempenho, não apenas no contexto desta tese, mas também em outras pesquisas. Todavia, há de se ressaltar que outros critérios podem ser inseridos em tal etapa. Para a definição dos critérios de ponderação dos SN e ranqueamento deles, é preciso que se observe o domínio a ser indexado, bem como os documentos que serão submetidos a tal processo, se serão partes do documento ou se serão o documento na íntegra.

No tocante à Etapa 7, há duas possibilidades de tomada de decisão: considerar os SN melhor ranqueados (com base nos critérios já mencionados) como descritores documentais do documento indexado ou fazer ajustes, como: inserção de outros sintagmas não identificados pelo *software;* modificação de SN; e eliminação de SN bem ranqueados, mas que, na percepção do indexador, não são bons descritores documentais para o documento indexado. Nesta etapa, o bibliotecário fará a validação dos SN, aceitando-os, corrigindo-os ou eliminando-os, uma vez que, como já exposto nesta tese, nem todo sintagma nominal representa o conteúdo do documento do qual foi extraído. Essa etapa é crucial para a efetiva atribuição dos SN ao documento indexado, tomando como fonte para isso os SN extraídos e selecionados automaticamente.

Na Figura 26, é possível ver um fluxograma da indexação semiautomática baseada nas diretrizes expostas nesta tese.

Figura 26 – Fluxograma da Indexação semiautomática por meio de SN baseadas nas diretrizes propostas desta tese



Fonte: Desenvolvido pelo autor (2024)

O fluxograma exposto na Figura 26 descreve o processo de indexação semiautomática por meio de sintagmas nominais. Tal proposta pode ser ajustada conforme necessidade do sistema de informação, bem como do domínio trabalhado.

7 CONSIDERAÇÕES FINAIS

Ante o exposto, foi possível corroborar o papel singular que a representação da informação assume no âmbito da Ciência da Informação, sobretudo da representação temática da informação no âmbito da recuperação da informação. Nesse contexto, a indexação manual (intelectual) ocupa lugar consolidado dentro das atividades voltadas para a representação dos conteúdos documentais. Ao lado dessa prática, vêm ganhado espaço as propostas automatizadas da indexação, buscando proporcionar mais agilidade e objetividade na representação temática de informação, mormente em ambientes digitais.

A automação da indexação apoia-se no fato de ser mais ágil e mais objetiva do que a indexação manual, a qual requer mais tempo e recursos humanos para ser executada nas unidades e sistemas de informação. Entretanto, a automação por completo da indexação ainda não é uma realidade isenta de inconvenientes, principalmente ligados à natureza da linguagem e a características próprias da operação de indexar, a qual busca representar, sinalizar o conteúdo veiculado no documento. Novas propostas de indexação automática surgem dando ênfase à semântica dos textos, dentre elas as que usam como unidade representativa o sintagma nominal. Tal unidade apresenta um potencial semântico maior quando comparadas com as palavras isoladas dos textos. As recentes pesquisas envolvendo sintagmas nominais reforçam essa afirmativa, uma vez que os sintagmas nominais, embora sejam retirados do contexto maior (o texto completo), carregam com si uma semântica precisa e que diz mais do que uma unidade lexical isolada.

Bebendo dos benefícios da automação da indexação, sem eximir dessa atividade o papel ímpar que assume o ser humano, a indexação semiautomática ganha relevância em um cenário em que é crescente a produção de informação e em curtos espaços de tempo. Essa atividade faz uso dos benefícios advindos do processamento automático da informação, combatendo a morosidade da prática manual, sem com isso deixar o viés humanístico necessário a uma atividade que por natureza é subjetiva.

Isto posto, desenvolveram-se pesquisas que conjugaram a automação da indexação e o uso de sintagmas nominais como fortes candidatos a descritores documentais. Essas pesquisas vêm ganhando espaço e mostrando resultados satisfatórios, embora ainda apresentem inconvenientes ora ligados à parte automática, ou seja, à identificação e extração dos sintagmas nominais que constituem um documento, ora ligados à própria natureza da linguagem e, por conseguinte, aos sintagmas nominais. Isto porque não basta que a estrutura seja reconhecida como sintagma nominal para que seja representativa do documento, faz-se necessário que possua potencial conceitual, ou seja, densidade informacional para representar o conteúdo veiculado no documento do qual foi retirado.

Considerando a singularidade da prática da indexação nas unidades e sistemas de informação, bem como o cenário atual acerca da automação da indexação e do potencial dos sintagmas nominais, nesta tese foi dada atenção à representação temática da informação, com ênfase na indexação semiautomática com base na identificação, extração e seleção de sintagmas nominais. Debruçou-se sobre a automação da indexação sem com isso isentar a parte humana de tal processo e sobre a potencialidade de estruturas linguísticas mais complexas e com maior potencial semântico quando análogas às palavras isoladas de um texto.

O exame detalhado dos sintagmas nominais evidenciou que essas estruturas são unidades potenciais para funcionarem como descritores documentais, visto que, ao serem extraídas dos documentos, ainda carregam uma semântica precisa e, em sua maioria, apontam para o conteúdo tratado no documento, ou seja, são aptas a atuarem como pontos de acesso dos documentos. Embora não se possa considerar que todos os sintagmas nominais extraídos de um documento possuam igual valor para fins de representação, os sintagmas nominais, se bem selecionados (com base em critérios preestabelecidos), se configuram como excelentes candidatos a descritores documentais. Ademais, os sintagmas nominais representam a própria linguagem dos usuários, dos autores, dos textos correntes e que circulam na sociedade.

A proposta de uma indexação semiautomática foi desenvolvida, considerando o cenário atual das pesquisas acerca dessa temática, as quais, quando operacionalizadas de forma totalmente automática, ainda apresentam limitações, sobretudo no que tange à dinamicidade da linguagem e dos fenômenos linguísticos presentes nos textos. A semiautomação da indexação busca agregar os benefícios de se processar grandes volumes de textos em curtos espaços de tempo à função essencial do indexador humano. Dessa forma, a indexação semiautomática se mostra promissora em um cenário de crescente produção de informação em meio digital, repositórios e bases de dados.

Boa parte das propostas de automação da indexação giram em torno da automação por completa de tal tarefa, contudo foi possível observar que a semiautomação associada ao uso dos sintagmas nominais se mostram oportunas, evidenciando resultados positivos e trazendo os benefícios expostos pelas pesquisas que se debruçam sobre a automação por completo desta atividade: a objetividade na análise da informação e a redução da morosidade no processamento temático da informação, além da melhoria na precisão e consistência da indexação. Isto posto, a vinculação da semiautomação da indexação ao uso de sintagmas nominais contribui diretamente para que os profissionais indexadores, ao realizarem a

indexação semiautomática, se aproximem cada vez mais dos temas tratados nos documentos indexados.

No que se refere aos sintagmas nominais, esta pesquisa apoia-se nos fundamentos teóricos e experimentais que se debruçaram nos estudos desses elementos. Os Sintagmas Nominais se apresentam como potenciais unidades na representação temática da informação, carecendo apenas de serem submetidas a critérios de refinamento e seleção dos sintagmas com maior potencial conceitual, ou seja, representativo. O uso dessas unidades linguísticas na indexação se mostra bem mais promissor do que a utilização de palavras isoladas, visto que aquelas unidades, ao serem extraídas dos documentos, possuem uma semântica mais precisa, como pôde ser visto nas diretrizes propostas nesta tese para uma indexação semiautomática baseada em SN, perpassando a escolha do *software*, a submissão dos documentos para a identificação e extração dos SN, a aplicação de critérios que refinem os SN para que restem os melhores SN candidatos a descritores documentais, a validação desses sintagmas pelo indexador humano e, por fim, a construção dos termos autorizados a representarem determinado documento.

Desenhar diretrizes para uma indexação semiautomática requer um olhar cuidadoso para os estudos já realizados que envolvem essa temática, os quais, somados a experimentos realizados em ambiente acadêmico, fornecem embasamento teórico e prático para refletir sobre tais processos. Conforme ressaltados no referencial teórico deste trabalho, bem como nos experimentos desenvolvidos no curso de mestrado e de doutorado do autor desta tese, a junção da indexação automática e o uso dos sintagmas nominais numa proposta semiautomática se mostra viável e promissora. Ao lado de todo o aporte teórico, os experimentos realizados e mencionados nesta tese buscaram avaliar os resultados alcançados pela parte automática de uma proposta semiautomática, verificando os índices de revocação, precisão dos SN, bem como os critérios de seleção que apresentam bom desempenho em propostas que usem SN e não palavras isoladas descontextualizadas.

Ante o exposto, acredita-se que os objetivos propostos nesta tese foram alcançados, uma vez que as análises permitiram um aprofundamento teórico acerca da indexação semiautomática e dos sintagmas nominais, permitindo, assim, alcançar o objetivo geral desta tese, que foi a proposição de diretrizes para que se realize a indexação semiautomática por meio de SN. Retomando os objetivos específicos desta tese: a) Identificar na literatura científica nacional e internacional estratégias de indexação semiautomática; b) Investigar composição dos Sintagmas Nominais, enquanto unidades portadoras de informação; c) Analisar o funcionamento do *software* PALAVRAS na identificação/extração de sintagmas

nominais por meio da análise de artigos científicos da área de Organização e Representação do Conhecimento; e d) Traçar procedimentos para a identificação semiautomática de SN que funcionem como elementos descritores do domínio Organização e Representação do Conhecimento, verifica-se que tais propósitos foram alcançados, os quais alicerçaram o alcance do objetivo geral.

Em relação aos objetivos específicos, o primeiro objetivo específico foi alcançado, ao passo que se identificaram no referencial teórico desta tese as propostas de automação da indexação até então desenvolvidas, as quais serviram de embasamento teórico para reflexão da viabilidade de se propor uma indexação semiautomática. O segundo objetivo específico foi alcançado por meio de uma análise cuidadosa e detalhada acerca da estrutura e composição dos sintagmas nominais. Tal análise se mostrou fundamental para compreender o funcionamento dos SN, já que a proposta nesta tese se apoia nessas unidades linguísticas. O terceiro objetivo específico permitiu analisar o comportamento do *software* na identificação e extração dos SN, identificando as falhas e limitações ainda presentes na parte automática da indexação semiautomática. Ademais, foi possível verificar que tais *softwares*, ainda que possuam limitações, conseguem extrair boa parte dos sintagmas nominais de um documento. O quarto e último objetivo específico permitiu desenhar os procedimentos a serem executados na indexação semiautomática realizada no experimento desta tese, os quais foram essenciais para mais na frente propor as diretrizes para uma indexação semiautomática por meio de SN.

Com base no exposto sobre os estudos nacionais e internacionais apresentados no referencial teórico, esta tese apoia-se no papel impar que exerce a indexação nas unidades e sistemas de informação, na contribuição da automação da indexação para o processamento temático da informação hodiernamente e, por fim, no potencial já evidenciado dos sintagmas nominais como unidades representativas dos documentos. O arcabouço teórico acerca dos estudos da indexação, da automação da indexação e dos sintagmas nominais e os experimentos realizados, associados ao estudo realizada nesta pesquisa, permitiram o aprofundamento do estudo da indexação semiautomática por meio de sintagmas nominais. Face ao exposto, esta tese propôs diretrizes a serem consideradas na elaboração de propostas de indexação semiautomática por meio de SN. Essas diretrizes envolvem aspectos ligados à ação manual realizada pelo indexador humano, que ainda é fundamental tanto na submissão dos documentos a serem indexados como na validação dos descritores representativos dos documentos, bem como envolvem também diretrizes acerca das operações totalmente automatizadas, como a identificação, extração e seleção de SN e aplicação de critérios de refinamento dos SN.

As diretrizes expostas nesta tese devem ser compreendidas como um norte a ser tomado no desenho de propostas de indexação semiautomática. Tais diretrizes não estão engessadas, uma vez que podem ser adaptadas a diferentes domínios com alterações em maior ou menor grau. Nesta tese, são propostos quatro critérios de seleção que servirão para melhor selecionar os SN com maior potencial semântico para fins de representação. Todavia, é pertinente que se ressalte a possibilidade de incorporação de novos critérios, após aprofundamento de estudos, bem como de adaptação dos aqui propostos. A utilização de critérios de seleção é essencial em qualquer proposta de indexação semiautomática, tendo em vista que cada documento é composto por inúmeros sintagmas e nem todos são representativos em termos de conteúdo do documento do qual são extraídos. Por esse motivo, faz-se necessária não só a aplicação de critérios de seleção como também a análise manual por parte do indexador humano. A validação pelo indexador não exigirá o mesmo tempo demandado na indexação totalmente manual, visto que esse profissional terá que avaliar apenas uma determinada quantidade de SN e os melhores classificados conforme os critérios supracitados.

Ao retomar a questão de pesquisa exposta no início desta tese “Como realizar a indexação semiautomática por meio da extração de Sintagmas Nominais de artigos científicos?”, verifica-se que as discussões, os resultados experimentais e, sobretudo, as diretrizes para a realização de uma indexação semaiutomática por meio de sintagmas nominais expostas nesta tese respondem à referida questão norteadora. Ademais, foi possível perceber o quão pertinente é o papel singular do indexador humano em processos automatizados de indexação.

Com base no exposto, a proposição de diretrizes para a indexação semiautomática por meio de sintagmas nominais evidencia a singularidade desta tese, uma vez que boa parte das pesquisas envolvendo a automação desta atividade se debruçam sobre a automação por completo. Contudo, esta atividade ainda não se isenta de inconvenientes, ora ligados às limitações tecnológicas, ora à própria natureza da linguagem. Nessa seara, a indexação semiautomática surge como uma saída, sobretudo por considerar o olhar humano em uma atividade tão delicada e complexa como a representação temática da informação. Ademais, a conjugação da semiautomação com o uso dos sintagmas nominais reforça ainda mais o caráter ímpar desta tese.

Em face do exposto, espera-se que esta pesquisa possa contribuir para o desenvolvimento de propostas de indexação semiautomática com base em sintagmas nominais em diferentes contextos, permitindo maior agilidade na indexação realizada em unidades e

sistemas de informação, economizando tempo do indexador humano e trazendo mais objetividade a esta tarefa de representação temática da informação. As diretrizes propostas nesta tese podem ser implementadas na indexação, bem como podem ser adaptadas e aprimoradas conforme estudos posteriores, sobretudo no que diz respeito às relações entre a indexação semiautomática por meio da extração de sintagmas nominais e as contribuições das ferramentas de Inteligência Artificial (IA) nesse processo (Aprendizado de Máquina, Processamento de Linguagem Natural, mineração de texto, análise de sentimentos, por exemplo).

No contexto desta pesquisa e dos trabalhos futuros, é possível fazer uso da IA em parte da indexação semiautomática, permitindo, assim, o processamento mais rápido de textos no que se refere à análise de assunto, por exemplo. A IA pode executar a parte do software utilizado para a identificação, extração e seleção de sintagmas nominais em um processo de indexação semiautomática, ressaltando a relevância da validação e/ou revisão final pela inteligência humana, alcançando, assim, a inteligência aumentada, a qual mostra-se viável para tarefas repetitivas como ocorre no processo de indexação.

É importante ressaltar que, embora a IA seja promissora na prática da indexação, a inteligência humana atua em situações em que a IA não consegue ou identifica erroneamente os significados ainda que estejam explícitos nos textos. Nessa perspectiva, seria mais viável a aplicação da junção das duas inteligências (artificial e humana), ou seja, a inteligência aumentada, no processo de indexação sobretudo pela própria natureza subjetiva dessa atividade.

Isto posto, a IA deve ser vista como uma ferramenta de assistência que deve ser explorada pelos estudos sobre indexação e pelos indexadores como ferramenta que contribui para o alcance de bons resultados em processos de representação temática da informação. Com o uso da inteligência aumentada (a vinculação da inteligência artificial à inteligência humana), alcança-se um equilíbrio entre a qualidade da indexação feita pelo humano com a agilidade alcançada com a indexação automática. Embora o processamento e a análise de assuntos sejam feitas pela inteligência artificial, a tomada de decisão cabe ao humano, o qual já terá economizado bastante tempo, uma vez que a máquina já realizou as tarefas repetitivas e que demandam mais tempo.

REFERÊNCIAS

ALBRECHTSEN, H. Subject analysis and indexing:from automated indexing to domain analysis. **The indexer**, v.18, n.4, p.219-224, 1993.

ALLEN, B. Cognitive research in information science: implications for design. **AnnualReview of Information Science and Technology**, v. 26, p. 3-37, 1991.

ANDRADE E CRUZ, M. C.; FERNEDA, E.; FUJITA, M. S. L. A disponibilização de vocabulário controlado aos usuários para a recuperação da informação. **RICI: R. Ibero-amer. Ci. Inf.,** v. 15, n. 1, jan./abril, 2022.

ARAUJO, E. A.; OLIVEIRA, M. A produção de conhecimentos e a origem das bibliotecas. In.: OLIVEIRA, M. (Org). **Ciência da Informação e Biblioteconomia**: novos conteúdos e espaços de atuação. 2. ed. Belo Horizonte: UFMG, 2011.

ARTANDI, S. Machine indexing: linguistic and semiotic implications. **Journal of the American Society for Information Science**, v. 27, n. 4, p. 235-239, July/Aug. 1976.

BANDIM, M. A. S.; CORREA, R. F. Indexação automática por atribuição de artigos científi cos em português da área de ciência da informação. **Transinformação,** v. 31, n., 2019. Disponível em: https://brapci.inf.br/#/v/217221. Acesso em: 10 marc. 2024.

BARANOW, U. G. Perspectivas na contribuição da linguística e de áreas afins à ciência da informação. **Ciência da Informação**, v. 12, n. 1, p. 23-35, 1983.

BARITÉ, M. **Referenciales teóricos vigentes en el área de tratamiento temático de la información y su expresión metodológica**. Porto Alegre: ABEBD, 1998. Relatório técnico do II Encontro de Dirigentes dos cursos superiores de Biblioteconomia dos países do Mercosul, Buenos Aires, nov. 1997.

_____. Sistemas de organización del conocimiento: una tipologia actualizada. **Informação & Informação**, Londrina, v. 16, n. 3, p. 122-139, 2011. Disponível em: http://ojs.uel.br/offline.html. Acesso em: 24 jul. 2022.

BEGHTOL, C. Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. **Jornal of Documentation**, London, v. 42, n. 2, p. 84-113, 1986. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/eb026788/full/html?skipTracking=true. Acesso em: 06 abril 2022.

BLAIR, D. C. **Language and representation in information retrieval.** New York: Elsevier Science, 1990.

BORGES, G. S. B.; LIMA, G. N. B. O. Desenvolvimento de softwares de indexação automática: breve avaliação dos principais critérios. **Informação & Tecnologia**, v. 2, n. 2, p. 49-70, 2015. Disponível em: http://hdl.handle.net/20.500.11959/brapci/41528. Acesso em: 17 nov. 2021.

_____.; MACULAN, B. C. M.; LIMA, G. Â. B. de. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade:** estudos, João Pessoa-PB, v. 18, n.2, p. 181-193, mai./ago. 2008.

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3–5, 1968. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090190103. Acesso em: 10 jun. 2022.

BOCCATO, V. R. C. A linguagem documentária vista pelo conteúdo, forma e uso na perspectiva de catalogadores e usuários. *In:* FUJITA, M. S. L. (Org.) et al. **A indexação de livros**: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais [online]. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. 149 p. Disponível em: http://books.scielo.org. Acesso em: 10 fev. 2022.

BOUTIN-QUESNEL, R. *et al*. **Vocabulaire systématique de la terminologie**. Québec: Publications du Québec, 1985 (Cahiers de l'Office de la langue française).

BRASCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento? Comunicação oral apresentada ao GT-02 - Organização e Representação do Conhecimento. *In:* ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais** [...]. São Paulo, 2008. Disponível em; http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/809/17.pdf?sequence=1. Acesso em 03 abril 2021.

BRUZINGA, G. S.; MACULAN, B. C. M. dos S.; LIMA, G.Â. B. de O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Salvador, **Anais** [...]. Salvador, 2007. Disponível em: http://www.enancib.ppgci.ufba.br/artigos/GT2--117.pdf. Acesso em: 14 jan. 2022.

CAIXETA, M.; SOUZA, R. R. Representação do conhecimento: história, sentimento e percepção. **Informação & Informação**, [S.l.], v. 13, n. 2, p. 34-55, nov. 2008. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/1815/1688. Acesso em: 02 abr. 2021.

CARNEIRO, M. V. Diretrizes para uma política de indexação. **Revista da Escola de Biblioteconomia da UFMG**, v. 14, n. 2, 1985. Disponível em: http://hdl.handle.net/20.500.11959/brapci/73170. Acesso em: 09 maio 2021.

CELERINO, V. G.; CORRÊA, R. F. A revocação na indexaçãoautomática por sintagmas nominais de artigos de periódicos em Ciência da Informação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 18. 2017. Marília. Disponível em: http://enancib.marilia.unesp.br/index.php/XVIII_ENANCIB/ENANCIB/paper/viewFile/40/1155. Acesso em 14 fev. 2023.

CESARINO, M. A. da. N.; PINTO, M. C. M. F. Cabeçalho de assunto como linguagem de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 7, n. 2, p. 268-288, set. 1978.
CESARINO, M. A. N.; PINTO, M. C. M. F. Análise de assunto. **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 1, p. 32-43, jan./jun. 1980. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/2089/2219. Acesso em: 11 maio 2021.

CINTRA, A. M. M. *et al*. **Para entender as linguagens documentárias**. 2ª edição revista e ampliada. São Paulo: Polis, 2002.

CORREA, R. F.; BANDIM, M. A. S. A consistência na indexação automática por atribuição de artigos científicos na área de ciência da informação. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 23, n. 53, 2018. Disponível em: https://brapci.inf.br/#/v/36926. Acesso em: 12 out. 2023.

CORREA, R. F.; LAPA, R. C. Panorama de estudos sobre indexação automática no âmbito da Ciência da Informação no Brasil (1973-2012). **Ciência da Informação**, v. 42, n. 2, p. 255-273, 2013.

CORRÊA, R. F.; CELERINO, V. G. Método de normalização de sintagmas nominais na indexação automática. **Em Questão**, Porto Alegre, v. 25, n. 1, p. 321–344, 2019. DOI: 10.19132/1808-5245251.321-344. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/81901. Acesso em: 18 dez. 2023.

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.

CORRÊA, R. F.*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.

CORRÊA, R. F.; BAZÍLIO, L. H. T. Análise da extração de descritores como sintagmas nominais através do software OGMA. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 22, n. 50, p. 44-58, set. 2017.

COOPER, W. S. I inter-indexer consistency a hobgoblin? **American Documentation**, 20, 1969.

CHAUMIER, J. Indexação: conceito, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n. 1/2, p. 63-79, jan./jun. 1988. Disponível em: https://pdfcoffee.com/indexacao-conceito-etapas-e-instrumentos-pdf-free.html. Acesso em: 21 mar. 2022.

CRUZ, C. M. da. A análise morfossintática e o estudo dos sintagmas: sugestões metodológicas. **Palimpsesto**, Rio de Janeiro, n. 19, out - nov. 2014, p. 399- 413. Disponível em: http://www.pgletras.uerj.br/palimpsesto/num19/estudos/palimpsesto19estudos01.pdf. Acesso em: 04 jun. 2021.

DAHLBERG, L. A referent-oriented, analytical concept theory for INTERCONCEPT. International Classification, 5 (3): 142-51,1978. Disponível em:https://www.nomos-elibrary.de/10.5771/0943-7444-1978-3-142/a-referent-oriented-analytical-concept-theory-for-interconcept-volume-5-1978-issue-3. Acesso em: 21 ago. 2021.

DANTAS, E. R. F.; SAMPAIO, D. A.; ALBUQUERQUE, M. E. B. C. DE. Avaliação da consistência de descritores: a representação da informação relacionada à temática responsabilidade social nas dissertações do PPGCI-UFPB. **Folha de Rosto**, v. 6, n. 1, p. 72-

84, 29 abr. 2020. Disponível em: https://brapci.inf.br/index.php/res/download/146559. Acesso em 12 jan. 2022.

DIAS, E. W.; NAVES, M. M. L. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007. 116 p. (Estudos Avançados em Ciência da Informação, 3).

DUBOIS, J. *et al*. **Dicionário de Lingüística.** São Paulo: Cultrix, 1978.

PEQUENO Dicionário Houaiss da língua portuguesa / Instituto Antonio Houaiss de Lexicografia. São Paulo: Moderna, 2015.

DODEBEI, V. L. **Tesauro**: linguagem de representação da memória documentária. Niterói: Intertexto, 2002.

DUBOIS, J. *et al*. **Dicionário de linguística**. São Paulo, Cultrix, 1997.

DUBUC, R. **? Qué es la terminologia**?: manual de terminologia. Providencia: Ril Ed., 1999.

ELISEU, André. **Sintaxe do Português**. [S.l]: Editorial caminho, 2008. (Coleção O essencial sobre língua portuguesa). Disponível em: https://www.ispsn.org/sites/default/files/documentos-virtuais/pdf/sintaxe_do_portugues_andre_eliseu_z-library.pdf. Acesso em: 05 jan. 2025.

FÁVERO, L. L.; KOCK, I. G. V. **Linguística textual**: introdução. São Paulo: Cortez, 1988.

FIORIN, J. L. (Org.) *et al*. **Introdução à Linguística**. São Paulo: Contexto, 2003.

FERNEDA, E. **Introdução aos modelos computacionais de Recuperação de Informação.** Rio de Janeiro: Ciência Moderna, v.1, 2012.

FIDEL, R. User-centered indexing. **Journal of the American Society for Information Science**, New Jersey, v. 45, n. 8, p. 572-576, 1994. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.4232&rep=rep1&type=pdf. Acesso em: 06 maio 2022.

FREITAS JUNIOR, N. *et al*. Indexação semiautomática de publicações através de técnicas de mineração de texto. CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, 12 & INOVARSE, 3 - Responsabilidade Social Aplicada. 2016. Disponível em: https://www.inovarse.org/sites/default/files/T16_222.pdf. Acesso em: 19 nov. 2021.

FUJITA, M. S. L. A identificação de conceitos no processo de análise de assunto para indexação. Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 1, n. 1, p. 60-90, jul./dez. 2003. Disponível em: http://www.sbu.unicamp.br/seer/ojs/index.php/sbu_rci/issue/archive. Acesso em: 20 abr. 2022.

FUJITA, M. S. L.; RUBI, M. P.; BOCCATO, V. R. C. As diferentes perspectivas teóricas e metodológica sobre indexação e catalogação de assuntos. *In*: FUJITA, M. S. L. (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. Disponível em https://repositorio.unesp.br/bitstream/handle/11449/109109/ISBN9788579830150.pdf?sequence=2&isAllowed=y. Acesso em: 02 abril 2021.

GARDIN, J.-C. *et al*. **La logique du plausible**: essais d'epistemologie pratique. Paris: Maison de Sciences de L'Homme, 1981.

GARDIN, J.-C. *et al.* (1968). **L'automatisation des recherches documentaires**: un modele général "Le SYNTOL". 2.ed. revue et augmentée. Paris: Gauthier-Villars.

GONÇALVES, J. A.; SOUZA, R. R. Relações e conceitos em ontologias: contribuições das teorias de farradane e dahlberg.,2008. Disponível em: http://hdl.handle.net/20.500.11959/brapci/176547. Acesso em: 12 abr. 2022.

GIL LEIVA, I. La automatización de la indización, propuesta teórico-metodológica: aplicación al área de Biblioteconomía y Documentación.1997. 268f. Tese (Doutorado em Filosofia e Letras) – Universidad de Murcia, Murcia, España, 1997. Disponível em: https://www.tesisenred.net/handle/10803/10917#page=1. Acesso em: 21 out. 2021.

GIL LEIVA, I. **La automatización de la indización de documentos**. Gijón: Trea, 1999.

GIL LEIVA, I. **Manual de indización. Teoría y práctica**. Gijón: Trea, 2008,

GIL LEIVA, I. Aspectos conceituais da indexação. *In*: GIL LEIVA, Izidoro Gil; FUJITA, Mariângela Spotti Lopes. (Editores). **Política de Indexação**. Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 04 abril 2021.

GIL LEIVA, I.; FUJITA, M. S. L. (Editores). **Política de Indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 04 abril 2021.

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

GUIMARÃES, J. A. C.; SALES, R.; GRACIO, M. C. **DataGramaZero** – **Revista de Ciência da Informação**, v. 13, n. 06, dez. 2012.

GUIMARÃES, J. A. C. A dimensão teórica do tratamento temático da informação e suas interlocuções com o universo científico da International Society for Knowledge Organization (ISKO). **Revista Ibero-americana de Ciência da Informação (RICI)**, v.1 n.1, p.77-99, jan./jun. 2008. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/940/815. Acesso em: 03 abril 2021.

GUIMARÃES, J. A. C. **Indexação em um contexto de novas tecnologias**. [S.l.: s.n.], 2000. 10 p. Texto didático.

GUIMARÃES, J. A. C. Aspectos éticos em organização e representação do conhecimento: uma reflexão preliminar. *In*: GONZÁLEZ DE GÓMEZ, M. N.; DILL ORRICO, E.Goyannes (Org.). **Políticas de memória e informação: reflexos na organização do conhecimento**. Natal: EDUFRN, 2006.

GUINCHAT, C.; MENOU, M. **Introdução geral às ciências e técnicas da informação e documentação**. Brasília: IBICT, 1994.

KATZ, J. O Escopo da Semântica. In: **Fundamentos Metodológicos da Lingüística**. v. 3, Campinas, 1982.

KENEDY, E. Gerativismo. In: Mário Eduardo Toscano Martelotta. (Org.). In.: **Manual de lingüística**. São Paulo: Contexto, 2008, v. 1, p. 127-140.

KINTSCH, W. & VAN DIJK, T. A. **Strategies of discouse comprehension**. San Diego, California, Academic Press, 1983.

KOBASHI, N. Y. **Elaboração de informações documentárias: em busca de uma metodologia**. 1994. Tese (Doutorado) – Universidade de São Paulo, São Paulo, 1994.

KURAMOTO, H. Sintagmas Nominais: uma nova proposta para a recuperação de informação. **DataGramaZero – Revista de Ciência da Informação.** Rio de Janeiro, v. 3, n. 1, fev. 2002.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, v. 25, n. 2, p. 1- 18, 1995.

KURAMOTO, H. Proposta de um Sistema de Recuperação de Informação Assistido por Computador – SRIAC, **Revista de Biblioteconomia de Brasília**, Brasília, v. 21, n.2, p. 211-228, jul./dez. 1999. Disponível em: https://brapci.inf.br/index.php/res/v/75896. Acesso em: 18 nov. 2021.

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.

LAPA, R. C. **A indexação automática no Brasil no âmbito da Ciência da Informação (1973-2012)**: o estado da arte. 2014, 151 f. Dissertação (Mestrado em Ciência da Informação) – Centro de Artes e Comunicação. Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2014.

LAPA, R. C.; CORRÊA, R. F. Indexação automática no âmbito da ciência da informação no brasil. **Informação & Tecnologia**, v. 1, n. 2, p. 59-76, 2014. Disponível em: https://brapci.inf.br/index.php/res/v/41624. Acesso em: 28 jun. 2022.

LASSWELL, H. D. A estrutura e a função da comunicação na sociedade. *In*: COHN, G. **Comunicação e indústria cultural.** São Paulo: Nacional; EDUSP, 1971

LE GUERN, M. **Unanalyseur morpho-syntaxique pour I'indexation automatique**. Le Français Moderne, juin, 1991.

LIMA, V. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em ciência da informação do vocabulário controlado do sibi/usp nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, v. 14, n. 1, p. 131-151, 2009. Disponível em: http://hdl.handle.net/20.500.11959/brapci/36563. Acesso em: 26 jan. 2022.

LINDEN, L. L.; BRÄSCHER, M. O tratamento temático da informação em instrumentos normativos de descrição arquivística. **Em Questão**, v. 24, n. 3, p. 96-124, 2018. Disponível em: https://brapci.inf.br/index.php/res/v/89102. Acesso em: 28 nov. 2021.

LOPES, L. L.; LOPES, L. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação**, v. 31, n. 1, 2002. Disponível em: https://www.brapci.inf.br/#/v/21868. Acesso em: 15 jan. 2024.

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa.** 2012, 156 f. Tese (Doutorado em Ciencia da Computação). Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

LU, W.; LI, X.; ZHIFENG, L.; CHENG, Q. How do author-selected keywords function semantically in scientific manuscripts? Knowledge Organization, [s.l.], v. 46, n. 6, p. 403-18, 2019. Disponível em: https://www.researchgate.net/publication/336720289_How_do_Author-Selected_Keywords_Function_Semantically_in_Scientific_Manuscripts/link/5daf22eb92851c577eb99975/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19. Acesso em: 22 dez. 2023.

LYONS, J. **Semántica linguística**: uma introducción. Ediciones Paidós Ibérica, S.A., Mariano Cubí, Barcelona, 1997.

MACULAN, B. C. M. D. S.; LIMA, G. N. B. O. Buscando uma definição para o conceito de "conceito". **Perspectivas em Ciência da Informação**, v. 22, n. 2, p. 54-87, 2017. DOI: 10.1590/1981-5344/2963 Acesso em: 28 set. 2021.

MAIA, L.C. G. **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.

MAIMONE, G. D; KOBASHI, N. Y.; MOTA, D. A. R.*In*: SILVA, J. F.; PALETTA, F. C. **Tópicos Para o Ensino de Biblioteconomia**. Volume 1. Disponível em: http://www3.eca.usp.br/sites/default/files/form/biblioteca/acervo/producao-academica/002749723.pdf. Acesso em: 08 abril 2021.

MARTINS, A. L. **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.

MEDEIROS, M. B. B. Terminologia brasileira em Ciência da Informação: uma análise. **Ciência da Informação**, Brasília, n.15, v. 2, p.135-42, jul./dez. 1986135. Disponível em: http://revista.ibict.br/ciinf/article/view/234/234. Acesso em: 21 ago. 2021.

MIORELLI, S. T. **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.

MOREIRO GONZÁLEZ, J. A. **El contenido de los documentos textuales**: su análisis y representación mediante el lenguaje natural. Gijón (Astúrias): Trea, 2004.

MORELLATO, L. V. **SIDSN**: sistema identificador de sintagmas nominais, 2007. 58 f. Monografia (Bacharelado em Ciência da Computação) – Departamento de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2007.

MORELLATO, L. V. **Metodologia computacional para identificação de sintagmas nominais na língua portuguesa**. 2010. 112 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Espírito Santo, Vitória, 2010.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. **Informação & Sociedade**: Estudos, João Pessoa, v.19, n.2, p. 99-118, maio/ago. 2009. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/10577/WOS000269446000009.pdf?sequence=2&isAllowed=y. Acesso em: 02 out. 2021.

NASCIMENTO, G. D. **Dos sintagmas nominais aos descritores documentais**: estudo de caso na indexação de teses e dissertações da área de Direito. 2015, 198 f. Dissertação (Mestrado em Ciência da Informação) – Centro de Artes e Comunicação. Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2015.

NASCIMENTO, G. D.; CORREA, R. F. Seleção de sintagmas nominais na indexação automática. **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 24, n. 55, p. 1-19, maio 2019. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2019.e57927. Acesso em: 16 ago. 2021.

NASCIMENTO, G. D.; CORREA, R. F. Avaliação de critérios para seleção de sintagmas nominais com valor para a recuperação da informação. **Transinformação**, Campinas, v. 30, n. 2, p. 179-192, Aug. 2018. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862018000200179&lng=en&nrm=iso. Acesso em: 16 ago. 2021

NASCIMENTO, G. D. **Dos sintagmas nominais aos descritores documentais**: estudo de caso na indexação de teses e dissertações da área de direito. 2015, 198 F. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal da Paraíba, João Pessoa, 2015.

OLIVEIRA, L. P. de. Política de Indexação: concepções acerca do conceito e percepções em torno de sua elaboração. **Ciência da Informação em Revista**, Maceió, v. 4, n. 2, p. 39-58, set. 2017. Disponível em: <https://www.seer.ufal.br/index.php/cir/article/view/3463>. Acesso em: 09 maio 2021.

OTHERO, G. Á. **Grammar Play:** um parser sintático em Prolog para a língua

portuguesa. 2004, 265 f. Dissertação (Mestrado em Letras) – Faculdade de Letras, Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, Porto Alegre, 2004.

PERINI, M. A. **Gramática do português brasileiro**. São Paulo: Parábola editorial, 2010.

PINHEIRO, L. V. R. Medidas de consistência da indexação: interconsistência. **Ciência da Informação,** Brasília, v. 7, n. 2, p.109-114, 1978.

PINTO MOLINA, M. **Análisis documental**: fundamentos y procedimientos. 2. ed. rev. aum. Madrid: EUDEMA, 1993.

PINTO, V. B. Indexação documentária: uma forma de representação do conhecimento registrado. **Perspectivas em Ciência da Informação**, v. 6, n. 2, 2001. Disponível em:https://brapci.inf.br/index.php/res/v/37708. Acesso em: 28 jun. 2021.

PINTO, V. B. Indexação Documentárias: uma forma de representação do conhecimento registrado. **Rev. de Letras**,v. ½, n. 22, jan/dez. 2000. Disponível em: http://www.revistadeletras.ufc.br/rl22Art09.pdf. Acesso em: 06 set. 2021.

PINHEIRO, M.S. **Uma abordagem usando sintagmas nominais como descritores no processo de mineração de opiniões**. 2009. Tese (Doutorado em Engenharia Civil). Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2009. Disponível em: http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=155208. Acesso em: 08 dez. 2020.

PRET, R.; CORDEIRO, R. A influência dos estudos semânticos no processo da indexação. In T. Barros & N. Tognoli (Orgs.), **Estudos avançados em organização do conhecimento: organização do conhecimento responsável**: promovendo sociedades democráticas e inclusivas (pp. 166-175). ISKO Brasil. https://brapci.inf.br/index.php/res/download/123289

REDIGOLO, F. M. **O processo de análise de assunto na catalogação de documentos**: a perspectiva sociocognitiva do catalogador em contexto de Biblioteca Universitária. 2010. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, São Paulo, 2010.

ROBREDO, J. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. rev. e ampl. Brasília: Edição de autor, 2005.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, v. 20, n. 2, 1991. Disponível em: https://brapci.inf.br/_repositorio/2010/04/pdf_2b09a726d5_0009108.pdf. Acesso em: 18 nov. 2021.

ROCHA, C. M. C. As expressões idiom´ticas da Língua Portuguesa em Dicionários Monolingues. **UNOPAR Cient**., Ciênc. Human. Educ., Londrina, v. 12, n. 2, p. 11-18, Out. 2011.

ROWLEY, J. A biblioteca eletrônica. Brasília: Brinquet de Lemos/Livros, 2002.

RUBI, M. P. **Política de indexação para construção de catálogos coletivos em bibliotecas universitárias**. 2008. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista. Campus de Marília. Marília, 2008. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/103388/rubi_mp_dr_mar.pdf?se. Acesso em: 05 set. 2021.

RUBI, M. P. Os princípios da política de indexação na análise de assunto para catalogação: especificidade, exaustividade, revocação e precisão na perspectiva dos catalogadores e usuários.*In*: FUJITA, M.S.L., (org.), *et al.***A indexação de livros**: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais [online]. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. 149 p. ISBN 978- 85-7983-015-0. Disponível em: http://books.scielo.org/id/wcvbc/pdf/boccato-9788579830150-06.pdf. Acesso em: 09 maio 2021.

RUBI, M. P.; FUJITA, M. S. L. Elementos de política de indexação em manuais de indexação de sistemas de informação especializados. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 8, n. 1, p.66-77, jan./jun. 2003. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/375/193. Acesso em: 09 maio 2021.

RUIZ PÉREZ, R. **El análisis documental**: bases terminológicas, conceptualización y estructura operativa. Granada: Ed. Universidad de Granada, 1992.

SACCONI, L. A. Nossa gramáticcompleta. 34. Ed. São Paulo: Matrix, 2020.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: https://www.brapci.inf.br/_repositorio/2017/07/pdf_7810a51cca_0000015436.pdf. Acesso em10 jun. 2021.

SAUTCHUK, I. **Prática de Morfossintaxe** - Como e por que aprender análise (morfo)sintática. 2. ed. Barueri, SP: Manole, 2010.

SAUTCHUK, I. **Prática de morfologia:** como e por que aprender análise (morfo)sintática. Barueri: Manole, 2004.

SAUSSURE, F. de. **Curso de Lingüística Geral**. Tradução Antônio Chelini *et al*. 25a edição. São Paulo: Cultrix, 1996.

SILVA, A. R. da. **A Dimensão discursiva da organização do conhecimento na Ciência da Informação Brasileira**. 2017. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília. Brasília, 2017. Disponível em: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1089229/a-dimensao-discursiva-da-organizacao-do-conhecimento-na-ciencia-da-informacao-brasileira. Acesso em: 14 jun. 2021.

SILVA, M. D. R.; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, v. 16, n. 2, p. 133-161, 2004. Disponível em: https://brapci.inf.br/index.php/res/v/115657. Acesso em: 28 jan. 2022.

SILVA, T. J. **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, R. R. **Uma proposta de metodologia para a escolha automática de descritores utilizando sintagmas nominais**. 2005. 197 f. Tese (Doutorado) – Curso de Doutorado em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 2005.

SOUZA, R. R.; 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.

SILVA, S. R. B.; CORREA, R. F.; GIL-LEIVA, I. Avaliação direta e conjunta de sistemas de indexação automática por atribuição. **Inf. & Soc.**: Est., João Pessoa, v. 30, n. 4, p. 1-27, out./dez. 2020. Disponível em: https://brapci.inf.br/index.php/res/v/153384. Acesso em 08 maio 2023.

SILVA, R. C.; BRITO, J. F. Proposta de um manual de indexação para bibliotecas universitárias. **Informação@Profissões**, v. 7, n. 1, p. 92-113, jan./jun. 2018. Disponível em: https://brapci.inf.br/index.php/res/v/64054. Acesso em: 27 ago. 2021.

SILVA, T. J.; CORRÊA, R. F. Ferramentas para indexação automática: uma análise comparativa entre o ogma, parser palavras, lx-parser e a extração manual de sintagmas nominais. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais**[...] João Pessoa: UFPB, 2015.

SILVA, T. J.; CORRÊA, R. F. Ferramentas para indexação automática: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. In.: XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação, 2015, João Pessoa. **Anais do XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação.** João Pessoa: PPGCI/UFPB, 2015.

SILVA, B. F. M; CORREA, R. F. A aplicação da folksonomia assistida na construção de corpus de referência em Ciência da Informação. **Em Questão**, v. 26, n. 2, p. 413-436, 2020.

SOERGEL, D. Indexing and retrieval performance: the logiciel evidence. **JASIS,** v.45, n.8, p.589-599, 1994.

SOUSA, B. P.; FUJITA, M. S. L. A análise de assunto no processo de indexação: um percurso entre teoria e norma. Informação & Sociedade: Estudos, v. 24, n. 1, 2014. Disponível em:https://edisciplinas.usp.br/pluginfile.php/7880440/mod_resource/content/1/Souza%20e%20Fujita.pdf. Acesso em: 02 fev. 2022.

STREHL, L. Avaliação da consistência da indexação realizada em uma biblioteca universitária de artes. **Ciência da Informação**, Brasília, v. 27, n. 3, p. 00, set. 1998. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000300011&lng=en&nrm=iso. Acesso em: 22 mai. 2021.

TÁLAMO, M. F.G. M. **Elaboração de resumos.** São Paulo: ECA/USP, 1987.

TAVARES, W. Q.; CELERINO, V. G. A importância da Bibliometria para a Indexação Automática. **Folha de Rosto: revista de Biblioteconomia e Ciência da Informação**. v.4, n. 1, p. 7-15, jul./dez., 2018. Disponível em: https://brapci.inf.br/index.php/res/v/109164. Acesso em: 20 marc. 2023.

TRUJILLO, A. M. Semântica, pragmática e tradução. **Revista InterteXto**, v. 5, n. 2, 2012.

TROITIÑO RODRIGUEZ, S. et. al. Indexing In Records Management. In J. A. C. Guimarães, S. O. Milani & V. Dodebei (Eds.), **Knowledge Organization for a Sustainable World**: Challenges and Perspectives for Cultural, Scientific, and Technological Sharing in a Connected Society: advances in Knowledge Organization (Vol. 15, pp. 234-242). Ergon Verlag, 2016.

UNISIST, Princípios de indexação. **Revista da Escola de Biblioteconomia da UFMG**, v. 10, n. 1, 1981. Disponível em: https://brapci.inf.br/index.php/res/download/87644. Acesso em: 08 maio 2021.

VAN SLYPE, G. **Los lenguajes de indización**: concepción, construcción y utilización en los sistemas documentales. Tradução de Pedro Hípola e Félix de Moya. Madri: Fundación Germán Sánchez Ruipérez; Pirámide, 1991. (Biblioteca del Libro).

VIGNOLI, R. G.; SOUTO, D. V. B.; CERVANTES, B. M. N. Sistemas de organização do conhecimento com foco em ontologias e taxonomias. **Informação & Sociedade: Estudos**, v. 23, n. 2, 2013. Disponível em: https://brapci.inf.br/#/v/91940. Acesso em: 04 maio 2024.

VYGOTSKY, L. S. **A formação social da mente:** o desenvolvimento dos processos psicológicos superiores. 4ª ed. São Paulo: Martins Fontes, 1991.

XATARA, C.M. O resgate das expressões idiomáticas. **Alfa:** Revista de Lingüística**,** São Paulo, v.39, p.195-210, 1995.

WARD, M. L. The future of the human indexer. **Journal of the American Society for Information Science**, v. 28, n. 4, p. 217-225, 1996.

APÊNDICE A – LISTA DE *SOTP WORDS* DE SINTAGMAS NOMINAIS POUCO RELEVANTES PARA FINS DE REPRESENTAÇÃO TEMÁTICA DA INFORMAÇÃO

14 artigos	esse âmbito	o tema	temas
a área	esse modo	o trabalho	trabalho desenvolvido
a literatura	esse trabalho	objetivo	um arcabouço teórico

a metodologia	esta finalidade	objetivos específicos	um domínio
a metodologia adotada	esta pesquisa	os aportes	um estudo qualitativo, exploratório e descritivo
a metodologia utilizada	este artigo	os assuntos abordados	um levantamento bibliográfico
a pesquisa	este estudo	os autores	um objetivo
a pesquisa bibliográfica	este grupo	os dados de pesquisa	uma busca exploratória
a pesquisa de doutoramento	este sentido	os dois principais autores	uma estratégia
a pesquisa exploratória e comparativa	este trabalho	os estudos	uma pesquisa bibliográfica e exploratória
a pesquisa teórica bibliográfica	levantamento bibliográfico	os instrumentos	uma pesquisa de mestrado
a presente pesquisa	mais pesquisas com esta finalidade	os métodos	uma pesquisa dedutiva, descritiva, bibliográfica e qualitativa
a proposta	marco teórico desta pesquisa	os objetivos	uma pesquisa exploratória
a revisão de literatura	metodologia de base bibliográfica e qualitativa	os pesquisadores	uma proposta
a temática	métodos de pesquisa bibliográfica e documental	os principais resultados	uma temática abordada na área
a temática estudada	o autor	os procedimentos metodológicos	
a tese de doutoramento	o corpus analisado	os resultados	
abordagem qualitativa	o desenvolvimento do trabalho	os termos	
abordagem teórico-metodológica	o estudo	os tópicos	
as duas propostas	o levantamento bibliográfico	os trabalhos	
as temáticas abordadas	o objetivo	os trabalhos publicados	
campo delimitado definido	o objetivo desta pesquisa	parte da pesquisa de doutoramento	
caráter exploratório	o objetivo deste artigo	parte da tese de doutoramento	
caráter qualitativo	o objetivo deste trabalho	pesquisa bibliográfica e documental	
caráter qualitativo, exploratório e descritivo	o objetivo geral	pesquisa de caráter exploratório	
contribuição deste trabalho	o objeto de estudo	pesquisa descritiva, documental e de abordagem quali-quantitativa	
duas questões	o presente artigo	procedimentos	

		bibliográficos e documentais	
elementos	o presente trabalho	quatro critérios	
essa questão	o referencial teórico	resultados	
essas temáticas	o referencial teórico da pesquisa	revisão bibliográfica	