LUCINÉIA SOUZA MAIA

EXTRAÇÃO E EXPLICITAÇÃO DE RELAÇÕES SEMÂNTICAS PARA A REPRESENTAÇÃO DO CONHECIMENTO DE DOCUMENTOS ACADÊMICOS: UM ESTUDO DE CASO A PARTIR DE UMA ESTRUTURA CLASSIFICATÓRIA

BELO HORIZONTE

2018

LUCINÉIA SOUZA MAIA

EXTRAÇÃO E EXPLICITAÇÃO DE RELAÇÕES SEMÂNTICAS PARA A REPRESENTAÇÃO DO CONHECIMENTO DE DOCUMENTOS ACADÊMICOS: UM ESTUDO DE CASO A PARTIR DE UMA ESTRUTURA CLASSIFICATÓRIA

Tese apresentada ao Programa de Pós-Graduação em Gestão e Organização do Conhecimento da Escola de Ciência da Informação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Doutora em Gestão e Organização do Conhecimento.

Área de concentração: Ciência da Informação

Linha de Pesquisa: Arquitetura e Organização do Conhecimento

Orientadora: Profa. Dra. Gercina Ângela de Lima

BELO HORIZONTE

2018

M217e

Maia, Lucinéia Souza.

Extração e explicitação de relações semânticas para a representação do conhecimento de documentos acadêmicos [manuscrito] : um estudo de caso a partir de uma estrutura classificatória / Lucinéia Souza Maia. 2018.

248 f., enc. : il., color.

Orientadora: Gercina Ângela de Lima

Tese (doutorado) Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

Referências: f. 201-209.

Apêndices: f. 210-247.

1. Ciência da informação Teses. 2. Representação do conhecimento (Teoria da informação) Teses. 3. Linguagens formais Semântica Teses. I. Título. II. Lima, Gercina Ângela Borém de Oliveira. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4.03

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

FOLHA DE APROVAÇÃO

EXTRAÇÃO E A EXPLICITAÇÃO DE RELAÇÕES SEMÂNTICAS PARA A REPRESENTAÇÃO DO CONHECIMENTO DE DOCUMENTOS ACADÊMICOS: UM ESTUDO DE CASO A PARTIR DE UMA ESTRUTURA CLASSIFICATÓRIA

LUCINÉIA SOUZA MAIA

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Arquitetura e Organização do Conhecimento.

Aprovada em 17 de dezembro de 2018, pela banca constituída pelos membros:

Prof(a). Gercina Ângela de Lima (Orientadora)
ECI/UFMG

Prof(a). Celia da Consolação Dias
ECI/UFMG

Prof(a). Cristiane Mendes Netto
UNIVALE

Prof(a). Elisângela Cristina Aganette
ECI/UFMG

Prof(a). Mateus Ferreira Satler
DECSI/UFOP

Prof(a). Luiz Cláudio Gomes Maia
FUMEC

Prof(a). Benildes Coura Moreira dos Santos Maculan
ECI/UFMG

Belo Horizonte, 17 de dezembro de 2018.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

ATA DA DEFESA DE TESE DA ALUNA
LUCINÉIA SOUZA MAIA

Realizou-se, no dia 17 de dezembro de 2018, às 13:30 horas, Sala 1000 - ECI/UFMG, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *EXTRAÇÃO E A EXPLICITAÇÃO DE RELAÇÕES SEMÂNTICAS PARA A REPRESENTAÇÃO DO CONHECIMENTO DE DOCUMENTOS ACADÊMICOS: UM ESTUDO DE CASO A PARTIR DE UMA ESTRUTURA CLASSIFICATÓRIA*, apresentada por LUCINÉIA SOUZA MAIA, número de registro 2014655434, graduada no curso de COMPUTAÇÃO - SISTEMAS DE INFORMAÇÃO, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Gercina Ângela de Lima - ECI/UFMG (Orientadora), Prof(a). Cristiane Mendes Netto - UNIVALE, Prof(a). Mateus Ferreira Satler - DECSI/UFOP, Prof(a). Benildes Coura Moreira dos Santos Maculan - ECI/UFMG, Prof(a). Celia da Consolação Dias - ECI/UFMG, Prof(a). Elisângela Cristina Aganette - ECI/UFMG, Prof(a). Luiz Cláudio Gomes Maia - FUMEC.

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 17 de dezembro de 2018.

Prof(a). Gercina Ângela de Lima

Prof(a). Celia da Consolação Dias

Prof(a). Cristiane Mendes Netto

Prof(a). Elisângela Cristina Aganette

Prof(a). Mateus Ferreira Satler

Prof(a). Luiz Cláudio Gomes Maia

Prof(a). Benildes Coura Moreira dos Santos Maculan

Dedico esta tese ao meu irmão
Adeilson Souza Maia (*in memoriam*).
Saudade eterna.

AGRADECIMENTOS

Durante meus momentos difíceis no doutorado, imaginar os agradecimentos que eu iria escrever era o que me motivava a continuar. Então, agora esses agradecimentos se “materializaram”.

Como não poderia deixar de ser, agradeço primeiramente a Deus pela oportunidade de cursar um doutorado. Mais ainda, agradeço a Ele por ser minha fortaleza, por me conduzir quando eu não conseguia seguir em frente. Agradeço também à Nossa Senhora Aparecida todas as vezes em que senti seu manto me protegendo e me dando força.

Agradeço aos meus pais Ilenira Maia de Souza e José Hercílio de Souza, pela educação e pelos valores recebidos. Minha gratidão a eles por todo apoio durante minha vida acadêmica, especialmente no doutorado.

Agradeço aos meus irmãos Patrícia Souza Maia e Wallace Souza Maia. À minha cunhada Daniela Maia e ao meu cunhado Cristiano Almeida pelo apoio familiar que foi tão importante para mim.

Importante também nesse processo foram os meus sobrinhos Sabrina Rodrigues de Souza, Giovanna Rodrigues de Souza, Heitor Maia e Antonella Maia, que trouxeram alegria, descontração e luz, que são próprios deles.

Um agradecimento especial à Sabrina Rodrigues de Souza, quem me impulsionou, mesmo sem saber, a fazer o doutorado, pois assim quero dar o melhor a ela e à sua irmã, Giovanna Rodrigues de Souza.

Agradeço ao meu namorado Paulo Oliveira dos Santos, pela compreensão, apoio e companheirismo durante o doutorado.

Agradecimentos às minhas amigas Rosimeiry Carvalho, Tuliane Fernandes e Alana Sester, que me apoiaram muito nessa conquista.

Agradeço à Universidade Federal de Ouro Preto, onde sou professora, que me concedeu um período para me dedicar exclusivamente ao doutorado, afastamento que foi fundamental para meu empenho. Juntamente, agradeço aos meus colegas de trabalho, pelo suporte.

Agradeço à Escola de Ciência da Informação da Universidade Federal de Minas Gerais, que me acolheu no doutorado. Estendo o agradecimento para a secretária Gisele Reis, que sempre foi muito solícita.

Agradeço também todos aos professores que me deram uma gotinha de seus saberes. Em especial, à professora Benildes Maculan, sempre presente para ajudar, e ao professor Dagobert Soergel: *my thankfull to professor Dagobert Soergel. I will never forgetten your help.*

Agradecimento mais que especial à professora Gercina Ângela de Lima pela orientação, pelo aprendizado, pela dedicação. Juntas passamos por momentos que extrapolaram a relação aluno-orientador. Você foi muitas vezes amiga, confidente, orientadora de vida. Profissionalmente também aprendi muito com você. Te admiro mais do que eu consigo expressar.

Por fim, agradeço aos colegas que conheci na Escola de Ciência da Informação, especialmente a Cristiane Mendes Netto, Eduardo Felipe, Décio Berti, Vinícius Tolentino, Elaine Diamantino, Celsiane Araújo, Flávia Abreu, Danielle Rioga, Graciane Buzinga, Helder Firmino, Patrícia Lopes e Webert Araújo, dos quais muitos participam do grupo de pesquisa MHTX. Aos membros desse grupo, meu muito obrigada por todas as contribuições.

RESUMO

As relações semânticas são fundamentais para a compreensão da natureza da ligação entre dois conceitos em um domínio. Para documentos acadêmicos, como dissertações e teses, a representação do conhecimento inerente a eles, com relações semânticas explicitadas, apoiam a compreensão dos usuários sobre pesquisas, por vezes complexas, de determinado domínio. Assim, o objetivo desta tese é propor um modelo de extração de relações semânticas para a representação do conhecimento de documentos acadêmicos no contexto do idioma português brasileiro. Para a elaboração do modelo, foi realizado um levantamento bibliográfico com os assuntos que permeiam a temática da tese. Desse modo, um dos resultados foi uma compilação de classificações de relações semânticas de vários autores em uma taxonomia de relações semânticas em português. Foi realizada ainda uma revisão de literatura que apontou a carência de pesquisas sobre extração de relações semânticas no cenário nacional. Nesse contexto, um sistema de informação na *web* chamado Semantizar foi desenvolvido para dar suporte à extração de relações semânticas a partir de estruturas classificatórias que representam documentos acadêmicos específicos. Desse modo, o Semantizar realiza buscas, nesses documentos, de pares de conceitos em frases, indicando a existência de relação semântica. Um estudo de caso foi realizado para avaliar a extração e a explicitação de relações semânticas por meio do Semantizar, nesse sentido, o sistema trouxe contribuições importantes para as pesquisas sobre extração de relações semânticas. Os resultados demonstram que quando dois conceitos de uma estrutura classificatória existem em uma frase, uma relação semântica entre eles pode existir de fato. Outra contribuição importante foi a descoberta de um novo subtipo de relação semântica associativa, que recebeu o nome de agente-subordinado. Por fim, conclui-se que esta pesquisa é relevante para comunidade científica, sobretudo porque ela traz constatações importantes para a extração de relações semânticas para a representação do conhecimento de documentos acadêmicos para serem aplicadas no cenário brasileiro.

Palavras-chave: Relações Semânticas. Extração de Relações Semânticas. Explicitação de Relações Semânticas. Representação do Conhecimento. Semantizar.

ABSTRACT

Semantic relations are fundamental for understanding the nature of the connection between two concepts in a domain. For academic documents, such as dissertations and theses, the representation of the knowledge inherent to them, with explicit semantic relations, supports user comprehension of occasionally complex studies in a specific domain. Thus, the aim of this thesis is to propose a model of semantic relations extraction for knowledge representation of academic documents in Brazilian Portuguese. For the formulation of the model, a bibliographical survey about the subjects that compose the theme of the thesis was carried out. One of the outcomes was a compilation of semantic relation classifications of several authors, thus composing a taxonomy of semantic relations in Portuguese. A literature review was carried out as well, and it pointed to the lack of research about extraction of semantic relations in the national scenario. In this context, a web information system called Semantizar was developed, in order to support the extraction of semantic relations from classificatory structures that represent specific academic documents. Thus, Semantizar performs, in these documents, searches for concept pairs in sentences, indicating the existence of a semantic relation. A case study was executed, in order to evaluate the extraction and explanation of semantic relations through Semantizar; in this sense, the system has brought relevant contributions to the research on the extraction of semantic relations. The results reveal that when two concepts of a classificatory structure exist in a sentence, a semantic relation between them can, in fact, exist. Another significant contribution was the discovery of a new subtype of associative semantic relation, which was named agent-subordinate. Finally, the conclusion is that this study is relevant to the scientific community, mainly because it introduces important findings for the extraction of semantic relations for knowledge representation of academic documents, which can be applied in the Brazilian scenario.

Keywords: Semantic Relations. Extraction of Semantic Relations. Explanation of Semantic Relations. Knowledge Representation. Semantizar

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

BCI	Biblioteconomia e Ciência da Informação
CC	Ciência da Computação
CRG	*Classification Research Group*
CSR	*Composition of Semantic Relations*
CSS	*Cascading Style Sheets*
cTAKES	*clinical Text Analysis an Knowledge Extraction System*
EI	Extração de Informação
ER	Extração de Relações
HTML	*Hypertext Markup Language*
IA	Inteligência Artificial
kNN	*k-Nearest Neighbor*
LaSE	*Location and Semantic Extraction*
LHD	*Linked Hypernym Discovery*
MHTX	Modelo de Navegação Hipertextual
MUC	*Message Understanding Conference*
NOSC	*Naval Ocean System Center*
OC	Organização do Conhecimento
PDF	*Portable Document Format*
PHP	*Hypertext Preprocessor*
PLN	Processamento de Linguagem Natural
POS	*Part of Speech*
PMEST	*Propert, Matter, Energy, Space, Time*
RC	Representação do Conhecimento
RDF	*Resource Description Framework*
REN	Reconhecimento de Entidade Nomeada
RI	Recuperação da Informação
SDG	*Structured Data Graph*
SGBD	Sistema Gerenciador de Banco de Dados
SOC	Sistema de Organização do Conhecimento
THD	*Targeted Hypernym Discovery*
UMLS	*Unified Medical Language System*
URI	*Unified Resource Identifier*

SUMÁRIO

1 INTRODUÇÃO

O conhecimento é central para orientar o pensamento e as ações das pessoas. O processo de construção do conhecimento é dotado de informação que, em tempos atuais, é amplamente disseminada na *internet*. No contexto acadêmico e científico, essa disseminação é feita por intermédio de ferramentas como catálogos públicos *online*, bases de dados bibliográficas, bibliotecas digitais de teses e dissertações, entre outras. No caso das bibliotecas digitais de teses e dissertações, como o próprio nome sugere, elas armazenam e fornecem acesso a teses e dissertações em formato digital na *internet*. Logo, elas são de grande importância para estudantes, professores e pesquisadores, pois apresentam pesquisas de doutorado e mestrado completas e de fácil alcance, pelo fato de estarem disponíveis na *internet.*

Textos como dissertações e teses, por vezes, podem ser extensos e conter conceitos e conceitualizações que, para alguns leitores, podem ser de difícil assimilação, mesmo para aqueles que pertencem à mesma área de conhecimento da pesquisa. Logo, o conhecimento contido nesses textos pode ser representado, até mesmo de maneira não textual, por meio, por exemplo, de mapas conceituais. Os mapas conceituais, mostram os conceitos e as relações existentes entre eles no contexto em que foram representados.

A Biblioteconomia e Ciência da Informação (BCI) e a Ciência da Computação (CC) têm estratégias próprias de representação do conhecimento. No âmbito da BCI, os Sistemas de Organização do Conhecimento (SOC) organizam e representam domínios do conhecimento por meio de instrumentos para apoiar a recuperação da informação pelo usuário. Esses instrumentos têm como base conceitos organizados de acordo com o propósito do SOC. Essas bases conceituais, são caracterizadas nesta tese como estruturas classificatórias, tais como taxonomias, sistemas de classificação, tesauros e estruturas facetadas[1].

---

[1] Tradicionalmente, as estruturas classificatórias, como as estruturas facetadas, trazem em seus arranjos termos. Contudo, nessa tese, optou-se por tratar os *termos* das estruturas classificatórias como *conceitos*. Isto porque as relações semânticas nesse contexto acontecem entre conceitos e não entre termos. Entretanto, nessa tese, *termo* e *conceito* serão sinônimos e serão utilizados intercambiavelmente.

Por outro lado, a CC contribui oferecendo tecnologias para a representação dos SOC. Um dos benefícios da convergência dessas duas áreas é o aproveitamento da perspicácia da representação de um domínio do conhecimento pela BCI com a aplicação de algoritmos da CC para aprimorar a representação do conhecimento. Utiliza-se algumas vezes, nesses casos, o Processamento de Linguagem Natural (PLN), que permite a tradução da linguagem humana para uma representação formal manipulável pelo computador.

O PLN pode ser aplicado para resolver diversos tipos de problemas, um deles é a extração automática de informações, que, por sua vez, abriga estudos acerca da extração de relações, o que permite que relações semânticas entre conceitos possam ser explicitadas em estruturas como os SOC.

Acredita-se que as relações semânticas são fundamentais para a compreensão de representações do conhecimento. As relações semânticas entre os conceitos nessas representações definidas efetivamente facilitam para o usuário assimilar o propósito da associação entre os conceitos no contexto que lhe é apresentado.

1.1 Problema

A especificação de relações semânticas em estruturas classificatórias realizada manualmente pode demandar um esforço humano que tende a dificultar a análise dos relacionamentos entre os conceitos, tornando essa prática inviável em alguns contextos, principalmente em domínios complexos. Nesse sentido, pesquisas sobre extração[2] automática de relações semânticas buscam auxiliar a explicitação[3] de relações semânticas. No entanto, no cenário nacional da BCI, percebeu-se que os esforços para a determinação de relações semânticas, encontrados na bibliografia, foram realizados manualmente, como em Café e Mendes (2009), Sales (2010) e Sales; Sayão e Motta (2012), que apontaram as relações semânticas de

[2] A extração segundo o Dicionário de Português licenciado para Oxford University Press, é o “ato ou efeito de extrair”; sendo que extrair é “tirar (algo) de dentro de[…] extratar”.

[3] Explicitação, de acordo com o Dicionário de Português licenciado para Oxford University Press, é “ação de explicitar, de tornar explícito” sendo que explicitar é: “ […] tornar claro, sem margem para ambiguidades” “[o] que é claro, explicado sem ambiguidade […]”.

uma estrutura facetada sobre o domínio da área nuclear, e em Maculan (2015), em que a autora propõe enriquecer semanticamente um tesauro, detalhando a natureza das relações associativas. Nesse caso, o domínio foi a agropecuária brasileira.

Na bibliografia pesquisada notou-se que a maioria das pesquisas tratam da problemática da extração automática de relações semânticas no contexto do idioma inglês. Segundo Batista *et al.* (2013, p. 52), as abordagens de extração automática de relações “não são facilmente transponíveis para línguas ou domínios diferentes”. De fato, cada língua tem suas características; logo, a adaptação das pesquisas sobre extração de relações semânticas no inglês para o português pode não ser adequada.

Com base no que foi exposto e na bibliografia pesquisada, duas lacunas foram percebidas: (1) não existe, no Brasil, pesquisas sobre explicitação de relações semânticas na BCI cujas relações são extraídas automaticamente e; (2) não foram constatadas pesquisas sobre extração de relações semânticas em português brasileiro para explicitar relacionamentos em estruturas classificatórias.

Alicerçado nessas lacunas e considerando que o idioma é o português brasileiro, o seguinte problema de pesquisa foi levantado: Dado dois conceitos A e B, retirados de uma estrutura classificatória, uma relação semântica pode ser descoberta a partir da análise de A e B nas frases de um texto?

1.2 Pressupostos

O problema de pesquisa partiu dos seguintes pressupostos: (1) a estrutura sintática das frases pode apoiar a descoberta de relações semânticas e; (2) alguns algoritmos de extração de relações semânticas utilizam a proximidade entre os termos para indicar a existência de uma relação semântica entre eles. Esses pressupostos são refletidos a seguir.

Sintaticamente, uma frase possui essencialmente sujeito e predicado e, complementarmente, objeto. O sujeito e o objeto são sintagmas nominais e o predicado é o sintagma verbal, sendo que, os sintagmas nominais necessariamente possuem substantivos (LUFT, 2002). Considera-se nesta tese, que os conceitos da

estrutura classificatória são substantivos. Logo, eles compõem os sintagmas nominais das frases. Desse modo, ao encontrar dois conceitos da estrutura classificatória em uma frase determina-se que eles compõem os sintagmas nominais para o sujeito e o objeto. Assim, considerando que uma frase pode ser estruturada como sujeito-predicado-objeto, resta analisar o predicado. Naturalmente, o sujeito e o predicado são uma relação, conforme Luft (2002). Por sua vez, o objeto só existirá para complementar o predicado, logo o predicado e o objeto também são uma relação. Assim, o predicado é (ou) contém a relação semântica entre o sujeito e o objeto, que são, nesse contexto, dois conceitos da estrutura classificatória.

O segundo pressuposto baseia-se no entendimento que a proximidade espacial entre os conceitos é um indicativo que eles se relacionam. Nesse sentido, algoritmos de mineração de dados podem ser utilizados para apontar padrões de proximidade entre os termos, que nesse caso, são os conceitos da estrutura classificatória. Logo, se os termos (ou conceitos da estrutura classificatória) estão na mesma frase, eles estão próximos espacialmente, então, existe uma relação semântica entre eles.

1.3 Objetivos

Diante do exposto, esta seção apresenta os objetivos geral e específicos, dados a seguir.

1.3.1 Objetivo geral

Propor um modelo de extração de relações semânticas para a representação do conhecimento de documentos acadêmicos[4] no contexto do idioma português brasileiro.

[4] do tipo Dissertações e teses.

1.3.2 Objetivos específicos

Os objetivos específicos desta tese são:

- Explicitar as relações semânticas existentes em estruturas classificatórias a partir da extração de suas fontes de informações.
- Contribuir com estudos sobre a extração de relações semânticas em português brasileiro.
- Facilitar a representação do conhecimento contido em documentos acadêmicos em meio digital.
- Colaborar com os estudos acerca de relações semânticas no idioma português brasileiro âmbito da Biblioteconomia e Ciência da Informação.

1.4 Justificativa

As justificativas que fundamentam esta tese são: (1) a extração de relações semânticas em textos em linguagem natural é fundamental para explicitar as relações entre os conceitos das estruturas classificatórias; (2) partindo da ideia que os conceitos de um texto ou domínio foram determinados *a priori*, é importante que se identifique consistentemente quais as relações semânticas entre esses conceitos para o entendimento da representação pelo usuário e; (3) as pesquisas sobre extração de relações semânticas em português são importantes para o desenvolvimento de estratégias que atendam às características do idioma no processamento de linguagem natural. Esses três fundamentos são refletidos a seguir.

Uma contribuição que justifica esta tese é que a extração de relações semânticas pode melhorar a atividade de explicitação das relações semânticas em estruturas classificatórias. De acordo com Cooke (1992), a explicitação de relações semânticas é fundamental para elucidar o conhecimento, permitindo a coerência das ideias apresentadas. Acrescenta-se a declaração de Blanco e Moldovan (2013), de que a extração de relações, utilizada para explicitar relações semânticas, é um passo

preliminar para entender o significado do texto. Segundo eles, quanto mais relações semânticas forem extraídas de uma sentença, melhor será a representação do conhecimento codificado nessa sentença. Bach e Badaskar (2007) também indicam que a extração de relações semânticas em um texto em linguagem natural é crucial para aplicações de entendimento de linguagem natural. Essas aplicações permitem representações do conhecimento de um domínio mais próximas da realidade.

No contexto desta tese, a representação do conhecimento é precedida pela organização do conhecimento, que utiliza estruturas classificatórias em que os conceitos podem ser apresentados, *a priori*, hierarquicamente. Contudo, as relações hierárquicas podem ser detalhadas em subtipos de hierarquias, como hipônimo-hiperônimo e merônimo-holônimo. Esses subtipos, por sua vez, ainda podem ser singularizados. Além disso, outras relações semânticas tendem a ocorrer entre os conceitos, como as relações de equivalência e as associativas, que também apresentam subtipos específicos. Logo, carece detalhar para as estruturas classificatórias a natureza das relações semânticas com base no domínio do conhecimento, geralmente expresso em linguagem natural. Isso é corroborado por Bräscher (2014) ao afirmar que,

> A identificação de relacionamentos hierárquicos em classificações e taxonomias é feita, geralmente, pelo posicionamento dos termos na estrutura vertical, mas não há indicação do tipo específico do relacionamento, o que há é uma ideia de subordinação quando categorias são expandidas, mas os relacionamentos dentro delas podem ser diferentes na sua natureza (BRÄSCHER, 2014, p. 177, tradução nossa).[5]

Portanto, nesse caso, compreende-se que a interpretação da ligação entre os conceitos fica a cargo do usuário, que pode não entender o que o classificionista pretendeu transmitir pela estrutura classificatória. Isso destaca a relevância desta tese, cuja proposta permite ao usuário assimilar a essência do relacionamento entre os conceitos no domínio do conhecimento em determinado contexto.

Ao contemplar o contexto para estabelecer as relações semânticas entre os conceitos, deve-se considerar especialmente o idioma, pois ele está estreitamente relacionado ao significado dos conceitos e suas relações. Ao ponderar que cada

---

[5] "*The identification of the hierarchical relationship in classifications and taxonomies is made, generally, by the positioning of the terms in the vertical structure, but there is no indication of the specific type of relationship, there is an idea of subordination when categories are expanded, but the relationships within them may be different in nature*". (BRÄSCHER, 2014, p. 177)

idioma tem as suas características, constata-se que a formação dos elementos frasais sujeito-predicado-objeto, que geralmente denotam uma relação semântica entre conceitos, deve contemplar as características do idioma em questão, o que, no caso desta tese, é o português brasileiro. Da mesma forma, deve-se observar, além da sintaxe, a morfologia, principalmente no que diz respeito à estrutura das palavras, pois as conjugações dos verbos, que geralmente indicam as relações semânticas, respeitam convenções próprias de seus idiomas. Logo, a extração de relações semânticas em português brasileiro deve considerar as peculiaridades próprias deste idioma.

1.5 Estrutura da tese

Esta tese está organizada da seguinte forma: inicialmente realizou-se uma introdução com a intenção de contextualizar o leitor quanto ao teor da problemática da pesquisa. Dessa forma, no capítulo introdutório, o problema de pesquisa foi devidamente elaborado com base na literatura pesquisada sobre extração e explicitação de relações semânticas no contexto da BCI. A partir do problema, os objetivos gerais e específicos foram detalhados. Para encerrar o Capítulo 1, foram apresentadas justificativas que atestam a relevância da pesquisa.

No Capítulo 2 apresenta-se a metodologia de pesquisa, discorrendo sobre a caracterização da pesquisa com base no problema levantado, assim como no objetivo traçado. Além disso, detalha-se sobre a amostra utilizada e explica-se o modelo de processo de desenvolvimento de *software*.

No Capítulo 3 apresenta-se a fundamentação teórico-metodológica. Nesse capítulo, os principais conceitos são relatados para permitir o entendimento dos assuntos abordados na tese, entre eles, as relações semânticas são tratadas no contexto da BCI e propõe-se uma taxonomia que compila-as. Ainda no contexto das relações semânticas, as suas propriedades são especificadas para permitir melhor caracterização das relações. Esse capítulo é finalizado com a fundamentação sobre extração de relações semânticas.

No Capítulo 4 explana-se a revisão de literatura realizada para investigar o estado da arte da extração das relações semânticas nos últimos cinco anos.

Prosseguindo, no Capítulo 5 define-se o Modelo de Extração de Relações Semânticas, iniciando com uma contextualização do modelo e seguido pelo planejamento para a criação de uma aplicação *web* para dar suporte ao modelo, aplicação que foi nomeada como Semantizar. Tal planejamento inclui a modelagem de dados e a especificação das etapas do Modelo de Extração de Relações Semânticas implementadas.

No Capítulo 6 descreve-se o estudo de caso para aplicação do Semantizar. O capítulo envolve a validação dos dados, a determinação das relações semânticas – respeitando a fundamentação teórica – e a análise e interpretação dos dados.

No Capítulo 7 apresenta-se os resultados do estudo de caso, destacando as contribuições para os estudos sobre extração de relações semânticas e com respeito ao Semantizar.

Por fim, no Capítulo 8 apresenta-se as considerações finais, com destaque para as contribuições gerais da tese.

2 METODOLOGIA DE PESQUISA

De acordo com Gerhardt e Souza (2009), a metodologia estabelece regras e procedimentos para realizar uma pesquisa. Segundo as autoras, "metodologia é o estudo da organização, dos caminhos a serem percorridos, para se realizar uma pesquisa ou estudo, ou para se fazer ciência" (GERHARDT; SOUZA, 2009, p.12). Este capítulo expõe a metodologia adotada para a realização desta pesquisa de doutorado com base em seus objetivos geral e específicos. Apresenta-se a caracterização da pesquisa e a caracterização da amostra.

2.1 Caracterização da pesquisa

Na metodologia de pesquisa é necessária a caracterização da pesquisa para que os métodos e técnicas mais adequados ao problema sejam adotados. De acordo com Pradonov e Freitas (2013, p. 49), "[c]ada tipo [de pesquisa] possui, além do núcleo comum de procedimentos, suas peculiaridades próprias". Desse modo, três classificações são necessárias para o delineamento de uma investigação; são elas: a classificação quanto à natureza, quanto aos objetivos e quanto aos procedimentos.

Quanto à sua *natureza*, esta pesquisa está classificada como *aplicada*. Segundo Silveira e Córdova (2009, p. 35), a pesquisa aplicada "[o]bjetiva gerar conhecimentos para a aplicação prática, dirigidos à solução de problemas específicos". No caso desta tese, essa característica se aplica porque um modelo de extração de relações semânticas foi criado e implementado em uma aplicação *web*.

Quanto aos objetivos da pesquisa, esta tese tem a característica de ser uma pesquisa *exploratória*, que permitiu o conhecimento dos assuntos que intercederam a proposta da pesquisa. Segundo Gil,

> Estas pesquisas têm como objetivo proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses. Pode-se dizer que estas pesquisas [exploratórias] têm como objetivo principal o aprimoramento de ideias ou a descoberta de intuições. (GIL, 2002, p.41)

Quanto aos *procedimentos*, definiu-se a realização da *pesquisa bibliográfica* e do *estudo de caso*. A pesquisa bibliográfica, como o próprio nome sugere, busca em fontes diversas, tais como livros e artigos científicos, suportes para fundamentação teórica (PRADONOV; FREITAS, 2013). Nesta tese, a pesquisa bibliográfica foi realizada em dois momentos: (1) no levantamento bibliográfico da fundamentação teórica-metodológica e; (2) no levantamento de trabalhos publicados sobre extração de relações semânticas para a revisão de literatura.

Na pesquisa bibliográfica para a fundamentação teórica-metodológica, utilizou-se ferramentas *online*, como as bases de dados bibliográficos do portal de periódicos Capes[6] e o *site* de pesquisas Google Acadêmico[7]. As buscas foram feitas em português e inglês, visando recuperar tanto referências bibliográficas nacionais quanto internacionais. Tais buscas resultaram em um conjunto de artigos publicados em revistas científicas, anais de eventos, livros eletrônicos e capítulos de livros.

Na seleção das referências bibliográficas também considerou-se sugestões de especialistas das áreas de Organização do Conhecimento e Representação do Conhecimento da Biblioteconomia e Ciência da Informação (BCI). Além disso, buscou-se conhecer as recomendações das bases de dados bibliográficas suscitadas a partir das buscas *online*. Publicações sobre as temáticas que permeiam esta tese, também foram contempladas na pesquisa bibliográfica, considerando os autores mais citados em outros trabalhos. Não obstante, utilizou-se livros que fundamentam os assuntos tratados.

Na pesquisa bibliográfica para a revisão de literatura, buscou-se identificar os esforços realizados na comunidade acadêmica e científica sobre a explicitação de relações semânticas por meio de técnicas de extração de relações. Uma revisão de literatura, de acordo com Fink (2013, p. 3, tradução nossa[8]), "[…] é um método sistemático, explícito e reproduzível para identificação, avaliação e sintetização de trabalhos completos e registrados, produzidos por pesquisadores, acadêmicos e outros". Por meio da revisão de literatura é possível identificar diferentes pontos de vista sobre determinado assunto referente ao problema de pesquisa. Logo, ela

---

[6] Disponível em: <http://periodicos.capes.gov.br/>. Acesso em 06 jul. 2015.

[7] Disponível em: <https://scholar.google.com.br/>. Acesso em 06 jul. 2015.

[8] "*A research literature review is a systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researches, scholars, and practitioners.*" (FINK, 2013, p. 3).

fornece o estado da arte relativo a uma temática (FIGUEIREDO, 1990; DA SILVA & MENEZES, 2005; MUKUL & DEEPA, s. d.). Nessa perspectiva, utilizou-se a metodologia de revisão de literatura sugerida por Okoli e Schabram (2010) para apontar trabalhos correlatos sobre extração de relações semânticas. Essa metodologia possui oito etapas: (1) proposta de revisão de literatura; (2) protocolo e treinamento; (3) busca na literatura; (4) tela prática; (5) avaliação de qualidade; (6) extração de dados; (7) análise dos achados; (8) escrita da revisão.

Nas etapas iniciais da revisão de literatura, determinou-se, entre outras coisas, a temporalidade da pesquisa, limitada a artigos publicados entre o período de 2013 a 2017, e as bases de dados: *Library, Information Science & Technology Abstracts with Full Text* (EBSCO)[9], *Information Science & Technology Abstracts* - ISTA (EBSCO)[10], *Library and Information Science Abstracts* - LISA (ProQuest)[11], *Web Of Science*[12] e *Scopus*[13]*,* cujas buscas resultaram em 9.386 artigos. Esses artigos abrangeram também assuntos correlatos à extração de relações semânticas.

Seguindo as etapas da metodologia para realização de revisão de literatura de Okoli e Schabram (2010), os artigos identificados no início passaram por avaliações e análises de relevância para a pesquisa. Dessa forma, após as etapas 3, 4, 5 e 6, dos artigos recuperados selecionou-se nove artigos, conforme pode ser observado no gráfico da Figura 1. Esses artigos foram criteriosamente estudados e apontaram técnicas e estratégias para a extração de relações semânticas em diferentes contextos, como pode ser visto no Capítulo 4, que também detalha todas as etapas da revisão de literatura.

---

[9]Disponível em: <https://www.ebscohost.com/public/library-information-science-and-technology-abstracts-with-full-text> Acesso em 20 fev. 2017.

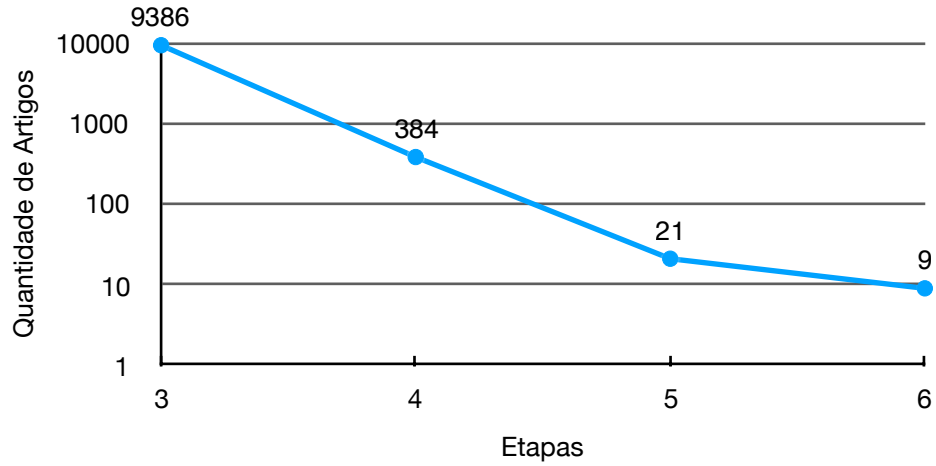
[10]Disponível em <https://www.ebscohost.com/corporate-research/information-science-technology-abstracts>. Acesso em 20 fev. 2017.

[11] Disponível em <http://www.proquest.com/products-services/lisa-set-c.html>. Acesso em 20 fev. 2017.

[12] Disponível em <http://webofknowledge.com>. Acesso em 20 fev. 2017.

[13] Disponível em <https://www.scopus.com>. Acesso em 20 fev. 2017.

Figura 1 – Gráfico com a quantidade de artigos selecionados nas referentes etapas da revisão de literatura



Fonte: Elaborada pela autora.

Gil (2002, p. 54) afirma que o estudo de caso “[c]onsiste no estudo profundo e exaustivo de um ou poucos objetos, de maneira que permita seu amplo e detalhado conhecimento". Para o autor, o estudo de caso permite uma visão global do problema ou identifica possíveis fatores que influenciam ou que são por ele influenciados na amostra.

Nesta tese, o estudo de caso foi utilizado para investigar a eficiência do Modelo de Extração de Relações Semânticas elaborado. A metodologia para a realização do estudo de caso foi organizada de acordo com a proposta de Processo de Experimentação de Wohlin *et al.* (2000), porém adaptada para esta tese. Dessa forma, quatro etapas foram utilizadas na condução do estudo de caso. A primeira, a definição, formalizou o objeto, o objetivo e o contexto do estudo de caso. A segunda etapa, o planejamento, legitimou a amostra a ser utilizada no estudo de caso e definiu os dados necessários para a realização das análises quantitativa e qualitativa do Modelo de Extração de Relações Semânticas. A etapa de operação consistiu na preparação da amostra para ser executada no sistema Semantizar, implementado para dar suporte computacional ao Modelo, e na validação dos dados coletados a partir da execução do sistema. Por fim, na quarta e última etapa, análise e interpretação, analisou-se quantitativamente e qualitativamente os dados validados na etapa anterior. Todas essas etapas estão apresentadas no Capítulo 6. Os resultados do estudo de caso são apresentados no Capítulo 7.

2.2 Caracterização da amostra da pesquisa

Um universo de pesquisa, segundo Gil (1989, p. 91), “[é] um conjunto definido de elementos que possuem determinadas características”. No caso desta tese, o universo de pesquisa utilizado na experimentação são estruturas classificatórias e seus respectivos documentos acadêmicos do tipo tese ou dissertação. Nesse universo, a *amostra* – que, segundo Gil (1989, p. 92), é um "subconjunto do universo, por meio do qual se estabelecem ou se estimam as características desse universo ou população” – é a estrutura facetada, o Mapa hipertextual MHTX de Lima (2004) e a respectiva tese por ele representada: *Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores*, de Naves (2000).

O MHTX é um modelo de navegação hipertextual em contexto criado por Lima (2004) para organizar Teses e Dissertações, visando apoiar a leitura e a recuperação desses documentos em Bibliotecas Digitais de Teses e Dissertações. No protótipo do MHTX, Lima (2004) criou três ferramentas de navegação: o sumário expandido, o mapa conceitual e a estrutura facetada. Na implementação, ela utilizou a tese supracitada de Naves (2000) para instanciar os instrumentos criados. Desses instrumentos, a estrutura facetada, mostrada na Figura 2, apresenta as características da amostra desejada para o Modelo de Extração de Relações Semânticas que será criado nesta tese.

Figura 2 – Estrutura Facetada do Modelo Hipertextual MHTX

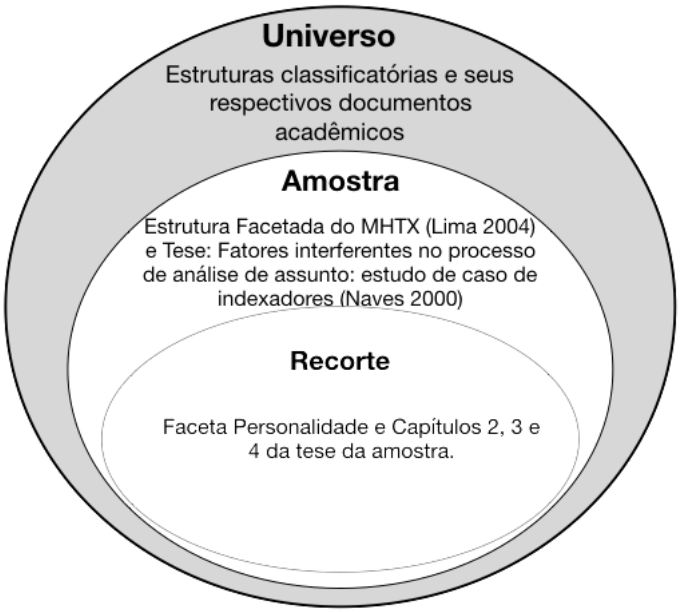
Fonte: Disponível em <http://www.gercinalima.com/mhtx/pages/prototipo-btdeci/teses/naves-mml/estrutura-facetada.php>. Acesso em 03 mai. 2014.

Para tornar mais eficiente a aplicabilidade da amostra, um *recorte* foi realizado na estrutura facetada, uma vez que a quantidade de conceitos não é o principal fator nesse momento. Dessa forma, os assuntos da faceta Personalidade serão utilizados no estudo de caso. Da mesma maneira, decidiu-se por um recorte da tese relativa à estrutura facetada utilizada. Logo, optou-se por considerar os capítulos 2, 3 e 4 da tese de Naves (2000). Essa escolha decorreu do fato de esses capítulos referirem-se à parte de definições conceituais da tese em questão, por isso eles parecem ser mais relevantes para o propósito ao qual serão aplicados. A Figura 3 mostra o extrato do sumário que consta esses capítulos. A Figura 4 pontua o universo, a amostra e o recorte mencionados.

Figura 3 – Recorte do Sumário da tese – Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores

Fonte: Naves, 2000.

Figura 4 – Delimitação do universo, amostra e recorte da pesquisa



Fonte: Elaborada pela autora.

2.3 Metodologia de desenvolvimento de *software*

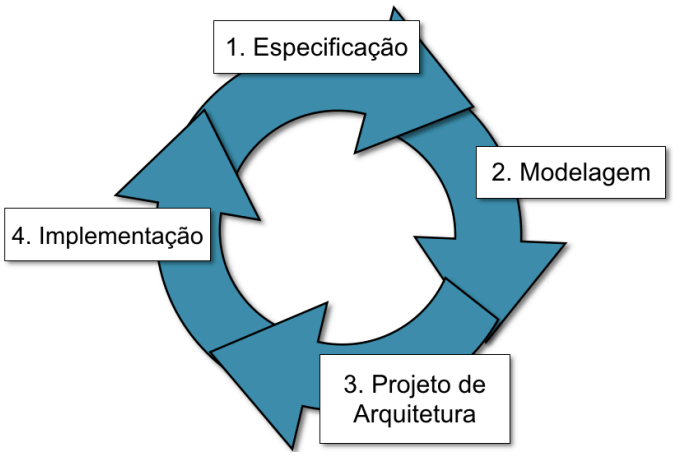
O Modelo de Extração de Relações Semânticas que se apresenta nesta tese foi idealizado de antemão para ser uma construção teórica de um modo de explicitar relacionamentos entre conceitos de documentos acadêmicos a fim de construir estruturas semânticas. Contudo, durante a elaboração do Modelo, compreendeu-se que sua implementação computacional apoiaria a investigação sobre a melhor maneira de extrair relações semânticas. Dessa forma, o Modelo se expandiu para um protótipo computacional.

A prototipação é um tipo de modelo de processo de *software* evolucionário que, como o próprio nome sugere, permite que o *software* evolua com o passar do tempo. Dessa forma, não é necessário ter a definição exata do *software* quando ele é iniciado (PRESSMAN; MAXIN, 2016).

De acordo com Pressman e Maxin (2016), o protótipo pode servir como “primeiro sistema”, conforme proposto nesta tese. “Embora alguns protótipos sejam construídos para serem 'descartáveis', outros são evolucionários, no sentido de que evoluem lentamente até se transformarem no sistema real” (PRESSMAN; MAXIN, 2016, p.46).

A prototipação, assim como outros modelos de processo de *software*, utiliza métodos para embasar as técnicas de desenvolvimento do *software*. Segundo Pressman e Maxin (2016, p. 16), “[o]s métodos envolvem uma ampla variedade de tarefas que incluem: comunicação, análise de requisitos, modelagem de projeto, construção de programa, testes e suporte.” Contudo, os autores pontuam que o desenvolvedor de *software* pode determinar os métodos que julgar necessários para a criação da aplicação. Nesse sentido, para o desenvolvimento do Modelo de Extração de Relações Semânticas, decidiu-se realizar as seguintes tarefas: (1) especificação, (2) modelagem de dados, (3) projeto arquitetural e (4) implementação do protótipo, conforme pode ser visto na Figura 5.

Figura 5 – Modelo de processo de desenvolvimento de *software*



Fonte: Adaptado pela autora com base em Pressman e Maxin (2016)

A *especificação* envolve uma descrição contextual do problema que o Modelo de Extração de Relações Semânticas se propõe resolver e uma descrição de como se pretende que o mesmo funcione. A *modelagem de dados* aborda a especificação das tabelas que compõem o banco de dados. Como se trata de um banco de dados relacional, a modelagem também contempla as relações entre as tabelas. O *projeto de arquitetura* determina e descreve os componentes necessários para a implementação computacional do Modelo como a interface, o *script*, o banco de dados e o servidor. Por fim, a *implementação* desenvolve computacionalmente o Modelo para funcionar como uma aplicação *Web*. O Capítulo 5 detalha sobre essas tarefas de desenvolvimento de *software* para o Modelo de Extração de Relações Semânticas que se propõe.

Este capítulo descreveu a metodologia empregada para a realização desta tese de doutorado com o detalhamento da caracterização da pesquisa e a delimitação do universo. O detalhamento dos procedimentos metodológicos para a realização da revisão de literatura, para o desenvolvimento do Modelo de Extração de Relações Semânticas e para o estudo de caso está descrito em seus respectivos capítulos.

Nesse contexto, o próximo capítulo aborda os assuntos que compõem a fundamentação teórica e metodológica.

3 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA

Este capítulo apresenta o embasamento teórico e metodológico para possibilitar a compreensão dos principais assuntos abordados nesta tese, que abrange duas áreas do conhecimento: Biblioteconomia e Ciência da Informação (BCI) e Ciência da Computação (CC). Aportes da Linguística também poderiam ser considerados, contudo, julgou-se que essas duas áreas têm o fundamento suficiente para o que se propôs investigar.

Nesse sentido, primeiramente aborda-se o conceito e a sua representação no contexto da Organização do Conhecimento (sub-área da BCI), destacando, dentro dessa temática, a classificação e as estruturas classificatórias. Em seguida, aborda-se a Representação do Conhecimento onde destacam-se os estudos acerca da Inteligência Artificial (IA).

Prosseguindo, as relações semânticas, objeto desta tese, são exploradas sobretudo sob o olhar da BCI, o que culminou em uma taxonomia que engloba seus vários tipos encontrados na literatura. Outrossim, descreve-se neste capítulo as propriedades das relações semânticas, fundamentais para o estabelecimento e interpretação coerente das restrições que envolvem os relacionamentos semânticos na Representação do Conhecimento. Por fim, encerra-se o capítulo com a apresentação de técnicas para extração de relações semânticas. Essas técnicas envolvem a IA, que, por sua vez, abrange o Processamento em Linguagem Natural (PLN).

3.1 O conceito e sua representação

O conhecimento é o elemento vital da sociedade moderna (SOERGEL, s. d.). O ato de conhecer permite às pessoas se inteirarem sobre os acontecimentos e tomarem decisões. A todo momento novos conhecimentos são produzidos. Contudo, segundo Soergel ([s. d.]), sem organização, esse conhecimento torna-se inutilizável, pois pessoas (e máquinas) devem conseguir alcançá-lo. Nesse sentido, a observância

a técnicas, métodos, processos e Sistemas de Organização do Conhecimento (SOC) é importante para conduzir a tarefa de representar o conhecimento para que os usuários possam encontrá-lo.

Broughton *et al*. (2005, p.140, tradução e grifos nossos)[14], são pragmáticos ao afirmar que "[o] termo 'organização do conhecimento' implica que o que está sendo organizado é o *conhecimento*". O conhecimento, nesse contexto, pode ser decomposto em unidades chamadas "conceitos".

Dahlberg (1978, p. 102) define "o conceito como uma compilação de enunciados verdadeiros sobre determinado objeto, fixada por um símbolo linguístico". Por essa definição pode-se dizer que, a partir da observação dos objetos, elaboram-se proposições verdadeiras sobre eles. Por exemplo, dado o objeto "copo", as seguintes proposições podem ser ditas: é de vidro, serve para beber líquidos e é um utensílio de cozinha. A reunião dessas características do objeto observado formam o conceito do objeto. Contudo, é necessário um símbolo que permita associar as características ao objeto. No exemplo dado, o símbolo linguístico desse conceito é a palavra "copo".

Os conceitos não ocorrem isoladamente, isso porque, primeiramente, eles pertencem a um domínio e, por conseguinte, eles se relacionam com outros conceitos desse domínio. Assim, conhecer o domínio é importante porque as características que formam um conceito podem variar de acordo com o contexto, pois diferentes domínios têm diferentes necessidades. Logo, o olhar sobre o objeto terá observações diferentes. No exemplo apresentado acima sobre o objeto "copo", as características levantadas atendem ao domínio de utensílios domésticos. Já no domínio de um supermercado, por exemplo, as características que interessam seriam preço e marca do copo.

Um conceito em um domínio relaciona-se com outros conceitos, formando uma estrutura semântica do conhecimento. Por meio das relações semânticas entre os conceitos, é possível compreender, entre outras coisas, o contexto do conceito em uma estrutura. Ainda exemplificando com o conceito "copo", pode-se ter a seguinte relação no domínio de utensílios domésticos em uma casa: *pessoa usa copo*. Já no contexto do supermercado, a relação seria: *pessoa compra copo*. Nesse exemplo, pode-se perceber que a relação entre "pessoa" e "copo" mudou de acordo com o

---

[14] *"The term 'knowledge organization' implies that what is being organized is* knowledge*."* (BROUGHTON *et al*., 2005, p. 140, grifos dos autores).

domínio em que esses dois conceitos estão inseridos e conforme o propósito da representação.

Estabelecer os conceitos e suas relações é importante, sobretudo para possibilitar estruturas de conhecimento organizadas de maneira compreensível (VICKERY; 2008). As estruturas de conhecimento, conforme tratadas nesta tese, têm como fundamento a classificação, que, segundo Bailey (1994), pode ser um processo e um resultado final. Enquanto processo, a classificação é definida pela ordenação de entidades (entende-se, de conceitos) em grupos ou classes, com base em suas similaridades. As similaridades são observadas considerando as características dos objetos e são definidas de acordo com os critérios do classificionista para atender às suas demandas. Logo, na classificação (agora, enquanto resultado final), cada classe resultante é diferente das outras, porém, internamente, elas são homogêneas, devido ao agrupamento por características similares. Nesta tese, a classificação é abordada como resultado final, cujo termo adotado para caracterizá-lo é a estrutura classificatória.

3.1.1 Estruturas classificatórias

As estruturas classificatórias, no contexto desta tese, são bases conceituais que quando arranjadas podem ser taxonomias e estruturas facetadas, que também podem ser chamadas de Sistemas de Organização do Conhecimento (SOC) ou instrumentos de representação do conhecimento. Neste subcapítulo, essas estruturas serão apresentadas sucintamente, pois não é um propósito desta pesquisa sua construção. Dessa forma, espera-se apenas contextualizar o leitor sobre tais estruturas.

Segundo Prieto-Díaz (2003, p. 460, tradução nossa[15]), “a taxonomia é uma estrutura de categorias e a classificação é o ato de arranjar entidades para as categorias dentro de uma taxonomia”. De acordo com Bailey (1994), assim como na classificação, a taxonomia pode se referir ao processo e ao resultado final. Enquanto resultado final, que é de interesse desta tese, a taxonomia é uma classificação de

[15] “*A taxonomy is a structure of categories and classification is the act of assigning entities to categories within a taxonomy*” (PRIETO-DÍAZ, 2003, p. 460).

entidades empíricas, ou seja, entidades definidas a partir da experiência e/ou observação de algo. Ainda de acordo com Bailey (1994), geralmente as taxonomias são hierárquicas e evolucionárias, ou seja, os conceitos são organizados hierarquicamente em classes e, se houver, subclasses. Além disso, a taxonomia pode ser atualizada com o passar do tempo, de acordo com a demanda e a evolução do conhecimento.

Na visão de Garshol (2004), o sentido de um tesauro expande o de uma taxonomia, pois ele descreve explicitamente uma estrutura semântica, diminuindo assim a ambiguidade terminológica. Logo, no tesauro, além dos conceitos serem relacionados hierarquicamente, as relações associativas e de equivalência também são contempladas por meio de notações específicas (*tags*) que assinalam seus tipos. Essas notações estão apresentadas no Quadro 1. A Figura 6 mostra um exemplo de um recorte de um tesauro.

Quadro 1 – Notações utilizadas em tesauros

Notação (*Tag*)	**Significado**
TG	Termo Geral: Representa o termo mais genérico ao qual o termo em questão pertence na hierarquia.
TE	Termo Específico: Representa, dentro da hierarquia, o termo genérico com acréscimo de características específicas do termo.
TR	Termo Relacionado: São termos que têm alguma relação com o termo em questão.
USE	Determina o termo sinônimo do termo em questão.
Use Para	Também determina o termo sinônimo do termo em questão, contudo, o termo assinalado com a *tag* USE é o termo preferido.
NE	Nota Explicativa: Utilizada para acrescentar alguma informação sobre o termo em questão.
CAT	Categoria: Utilizado para tesauros facetados, que possuem categorias. Logo essa *tag* representa a categoria a qual o termo em destaque pertence.
ING	Inglês: Utilizado em alguns tesauros para assinalar o termo em questão no idioma inglês.
ESP	Espanhol: Utilizado em alguns tesauros para assinalar o termo em questão no idioma inglês.

Fonte: Elaborado pela autora a partir da interpretação de Pinheiro e Ferrez (2014).

Figura 6 – Recorte do Tesauro Brasileiro da Ciência da Informação

classificação

ING: classification
ESP: clasificación (UP clasificación bibliográfica)
TG organização do conhecimento
TE classificação automática
classificação facetada
classificação hierárquico-enumerativa
TR arranjo sistemático
indexação
teoria da classificação
notação
representação do conhecimento
sistemas de classificação
NE: Usar para documentos sobre como classificar e como criar um sistema de classificação. Para esquemas específicos, use "sistemas de classificação".
CAT: 2.1 Organização do Conhecimento

Fonte: Pinheiro; Ferrez, 2014.

Por fim, apresenta-se as estruturas facetadas (também chamadas de estruturas de classificação facetada ou ainda esquemas de classificação facetada). Essas estruturas são referidas nesta tese como resultado da análise e classificação facetada. Para alguns autores, como Broughton (2008) e La Barre (2010), a análise facetada é uma metodologia para a construção da classificação facetada. Nas estruturas facetadas, resultantes da classificação facetada, as facetas correspondem ao agrupamento de conceitos que seguem uma mesma regra. Tais regras podem ser conceitos que têm as mesmas características e que combinam padrões similares a outros conceitos, formando classes básicas ou categorias (SOERGEL, s. d.).

O entendimento de categorias, no contexto da análise e classificação facetada, foi originalmente proposto por Ranganathan, que indicou o uso de cinco categorias fundamentais, conhecidas como PMEST (*Property* (Propriedade), M - *Matter* (Matéria), E - *Energy* (Energia), S - *Space* (Espaço) e T - *Time* (Tempo), para representar o conhecimento. Posteriormente, os membros do *Classification Research Group* (CRG) entenderam que essas cinco categorias não eram suficientes e ampliaram para treze o número de categorias: coisa, tipo, parte, propriedade, material, processo, operação, agente, paciente, produto, subproduto, espaço e tempo (BROUGHTON, 2008). Contudo, ainda de acordo com Café e Bratfisch (2012), algumas vezes, dependendo do domínio, é necessário ainda utilizar outras categorias além do PMEST ou das treze categorias do CRG para representar a informação.

As taxonomias, tesauros e estruturas facetadas abordadas nesta seção

representam o conhecimento de um domínio e são utilizadas em sistemas de informação que apoiam os usuários a compreenderem tal domínio e encontrarem a informação desejada. Desse modo, a representação do conhecimento, da forma que é abordada na próxima seção, expande a representação em estruturas classificatórias e embute tecnologia da informação para aperfeiçoar computacionalmente tais estruturas.

3.2 Representação do conhecimento

A Representação do Conhecimento (RC) busca criar abstrações que refletem o mundo real para atender a determinada finalidade. Dessa forma, ela representa conceitos (unidades do conhecimento) utilizando ou formando bases de conhecimento e, como no caso desta tese, bases conceituais como as estruturas classificatórias apresentadas, além de metadados e padrões de tecnologia da informação.

De acordo com Van Harmelen, Lifschitz e Porter (2008), a RC é o cerne do grande desafio da Inteligência Artificial (IA), qual seja: compreender a natureza da inteligência e da cognição tão bem que os computadores possam ser feitos para executar habilidades semelhantes às humanas. Em 1958, John McCarthy contemplou sistemas de IA que poderiam exercer o senso comum. A partir desse e de outros trabalhos, os pesquisadores ganharam a convicção de que a inteligência artificial poderia ser formalizada como raciocínio simbólico com representações explícitas do conhecimento e que o desafio central de pesquisa era descobrir como representar o conhecimento em computadores e utilizá-lo para resolver problemas (VAN HARMELEN; LIFSCHITZ; PORTER, 2008).

Pesquisas acerca da RC a apontam como instrumento para a abstração da realidade refletida em modelos que permitem o raciocínio inteligente sobre algo para a posterior tomada de decisão. Diante disso, para Davis; Shrobe e Szolovits (1993) a RC pode ser vista sob cinco perspectivas: (1) como um substituto de algo, quando ela permite o raciocínio sobre o mundo ao invés do agir sobre ele; (2) como um conjunto de compromissos ontológicos, em que a representação busca refletir sobre quais termos deve-se pensar sobre o mundo; (3) como uma teoria fragmentária de raciocínio inteligente, quando ela possibilita refletir sobre como as pessoas raciocinam

inteligentemente; (4) como um meio para computação eficiente, no momento em que a RC fornece um ambiente computacional em que a reflexão é alcançada; e (5) como um meio de expressão humana, quando ela retrata uma forma de se dizer coisas sobre o mundo.

As perspectivas de RC podem ser traduzidas em esquemas, tais como os apresentados por Chua; Storey e Chiang (2012), que sumarizam quatro tipos de esquemas de RC baseados em lógica, em hierarquia, em grafos e em *frames.* Desses modelos, pressupõe-se que os hierárquicos são os que mais empregam os sistemas de base de conhecimento.

De acordo com Dolk e Konsynski (1984), os esquemas baseados em lógica são comuns em ambientes de apoio à decisão amparados por sistemas baseados em regras. As regras são expressas sob a forma `if` <condição> `then` <consequência>. As inferências são feitas examinando as condições e consequências recursivamente até que um objetivo ou conclusão seja alcançado(a).

Já nos esquemas de RC hierárquicos (que, na abordagem da BCI, seriam as estruturas classificatórias), todos os conceitos são arranjados de modo que alguns sejam refinamentos de outros. Nesse caso, o conhecimento é gerado por meio de navegação pelos relacionamentos entre os conceitos. Nessa perspectiva, as redes semânticas – que também são um tipo de SOC, de acordo com Zeng (2008) – podem ser vistas como um tipo de esquema de RC hierárquico (CHUA; STOREY; CHIANG, 2012). De acordo com Grigorova e Nikolov (2007), uma rede semântica consiste em nós e *links* (arcos). Os nós representam objetos e eventos; os arcos representam relacionamentos entre eles. Segundo os autores, uma rede semântica pode ser estruturada sob diferentes pontos de vista: tipo-subtipo, todo-parte, antes-depois e assim por diante.

Nos esquemas de RC baseados em grafos, o conhecimento é representado como conceitos relacionados por meio de grafos, que são estruturas com nós e arestas, em que os nós representam os conceitos e as arestas o relacionamento entre eles. No caso dos grafos, o conhecimento é inferido pela distância entre os nós (CHUA; STOREY; CHIANG, 2012).

Por fim, os esquemas de RC baseados em *frames*. Segundo Marinov (2008), um *frame* é uma coleção de atributos (usualmente chamados de *slots*) e valores associados (e possíveis restrições nos valores) que descrevem alguma entidade do

mundo a partir de um senso absoluto ou de um ponto de vista em particular. Conforme o autor, em muitos aspectos, um *frame* representa um registro em uma base de dados. Contudo, os *frames* são mais estratificados que os registros, e, ao contrário de um registro que contém campos, um *frame* contém *slots*, facetas e dados.

Nesta tese, a estrutura resultante do modelo proposto, contempla a representação de conceitos e relações. Essa representação poderá, futuramente, apoiar a criação de redes semânticas e de ontologias. Logo, na classificação dos modelos de RC de Chua, Storey e Chiang (2012) apresentados, os esquemas RC hierárquicos poderão ser contemplados nesta pesquisa. Nesse sentido, entende-se que enriquecer uma estrutura classificatória com relações semânticas explicitadas é um passo importante para a representação do conhecimento desses esquemas hierárquicos.

3.3 Relações semânticas

Nesta tese, as relações semânticas são abordadas no contexto de associação entre dois conceitos[16]. De acordo com Khoo e Na (2006) e Green, Bean e Myaeng (2013), os conceitos podem ser vistos como blocos de conhecimento e as relações como uma ligação que conecta e mantém esses blocos juntos dentro das estruturas de conhecimento na mente das pessoas[17]. Logo, as atividades cotidianas dependem

---

[16] Alguns autores, como Green (2001) e Szostak (2012), utilizam a expressão “relações semânticas entre *coisas*”. Campos (2004) e Hjørland (2007) utilizam “relações semânticas entre *objetos*”. Já Storey (1993), Green (2001) e Khoo e Na (2006) utilizam “relações semânticas entre *entidades*”. Nesta tese, essas expressões são empregadas intercambiavelmente, salvo quando o uso delas for estritamente necessário para o contexto empregado. Da mesma forma, as palavras “relação” e “relacionamento” também são utilizadas intercambiavelmente, embora possuam significados diferentes. Segundo o dicionário Michaelis, um dos significados de relação é a “[l]igação que existe entre pessoas, coisas ou fatos”. E, um dos significados de relacionamento, por sua vez, é o “[a]to ou efeito de relacionar(-se)”. A maioria dos autores citados nesta tese utilizam a expressão “*relação* semântica (CHAFFIN; HERRMANN, 1984; ;MURPHY, 2003; KHOO e NA, 2006; HJØRLAND, 2007; PETERS e WELLER, 2008; STOCK, 2010; CAFÉ e MEDEIROS, 2011; BRÄSCHER, 2014). Já Storey (1993) e Bean; Green (2001) utilizam a expressão “*relacionamento* semântico”.

[17] Muitas pesquisas em semântica são baseadas na aceitação de que os conceitos são de alguma forma conectados na mente das pessoas. Contudo, para Hjørland (2007), na Organização do Conhecimento, as relações semânticas são principalmente produtos de modelos ontológicos científicos. Segundo ele, as relações entre os elementos químicos, por exemplo, não são conectadas em nossos cérebros, mas são descobertas por pesquisadores químicos. Logo, os criadores de Sistemas de Organização do Conhecimento precisam identificar as relações semânticas na literatura referente ao assunto.

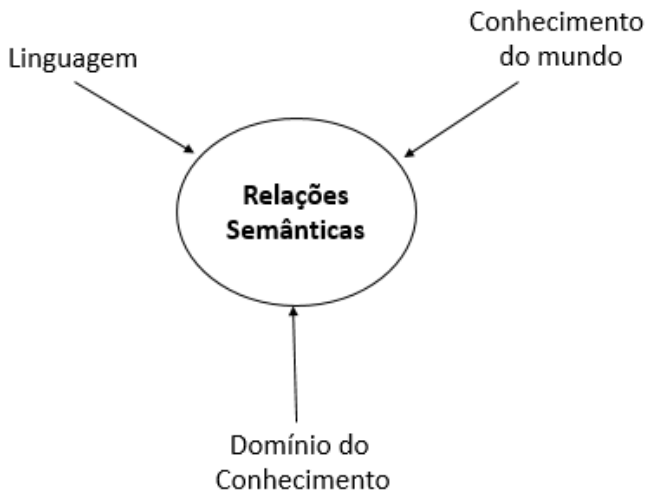
da exatidão e da riqueza dessas estruturas de conhecimento e da sua rede de relações.

Khoo e Na (2006) ressaltam que é importante compreender que os conceitos e as relações semânticas são bilateralmente dependentes, ou seja, se por um lado os conceitos são obrigatoriamente parte integrante de uma relação, por outro, eles precisam das relações semânticas para tornarem-se coerentes.

Nas relações entre conceitos, o número de conceitos participantes pode variar; isso se chama aridade e define o n-ário de uma relação, com n podendo ser de 1 a infinito. Embora a maioria das relações sejam binárias, também podem existir relações unárias, ternárias, quaternárias, e assim sucessivamente. Exemplos de relações unárias (que envolvem apenas um conceito) são as relações de casos: ação-destinatário, que ocorre entre a ação e a coisa que recebe essa ação (por exemplo: supervisionar-empregado); ação-instrumento, que acontece entre a ação e o instrumento que é usado nessa ação (por exemplo: cortar-faca). Essas relações foram propostas por Storey (1993). Além delas, existem as relações ação-produto, como no exemplo escrever-publicação, e ação-paciente, como em ensinar-estudante. Essas relações foram sugeridas por Peters e Weller (2008). Já as relações binárias (que envolvem dois conceitos) são as mais comuns e são o foco desta tese. Nesse caso, dado o exemplo “consultor *realiza* consultoria”, pode-se afirmar que a relação *realiza* está ligando dois conceitos: consultor e consultoria. Por fim, uma relação ternária (que envolve três conceitos) é exemplificada por Stock (2010) utilizando a relação *curar*. Segundo ele, essa relação abarca três elementos: pessoa-doença-medicamento. Contudo, de acordo com o autor, essa relação pode ser desmembrada nas seguintes relações binárias: pessoa-doença, doença-medicação e medicação-pessoa.

Para a definição das relações semânticas, deve-se considerar que elas são influenciadas pela linguagem e cultura, pelo conhecimento do mundo e pelo domínio do conhecimento, como ilustrado na Figura 7. Esses fatores são refletidos a seguir.

Figura 7 – Fatores que influenciam a definição e a compreensão das relações semânticas



Fonte: Elaborada pela autora.

Segundo Khoo e Na (2006), os aspectos linguísticos são fundamentais para a compreensão de estruturas de conhecimento que envolvem conceitos e suas relações. Para os autores, é difícil analisar o significado dos conceitos e suas relações à parte da linguagem, pois cada linguagem tem as suas características e está atrelada ao fator cultural.

O outro fator que afeta as relações semânticas é o conhecimento inerente do ser humano sobre as coisas, ou seja, o conhecimento do mundo. Para Hjørland (2007), sem esse tipo de conhecimento não se pode estabelecer uma relação semântica entre as coisas, como, por exemplo, sem o conhecimento do fato de que Copenhagen é a capital da Dinamarca, é impossível determinar a relação entre Copenhagen e Dinamarca.

Por fim, ainda de acordo com Hjørland (2007), as relações semânticas refletem metas específicas de seus respectivos domínios[18]. Para o autor, o mesmo objeto em diferentes domínios pode ser descrito destacando características diferentes, refletindo interesses humanos distintos. O autor exemplifica utilizando um produto químico analisado nos domínios da Farmacologia e da Química. De acordo com ele, uma base de dados com informações sobre química enfatizará descrições estruturais

[18] Hjørland é o pai do paradigma domínio-analítico. De acordo com o autor, nessa abordagem é necessário relacionar as necessidades do usuário à estrutura e organização do conhecimento, aos padrões de cooperação, às formas de linguagem e comunicação, aos sistemas de informação e às regras da sociedade. Nesse sentido, é importante compreender que diferentes domínios têm diferentes necessidades (HJØRLAND, 1997).

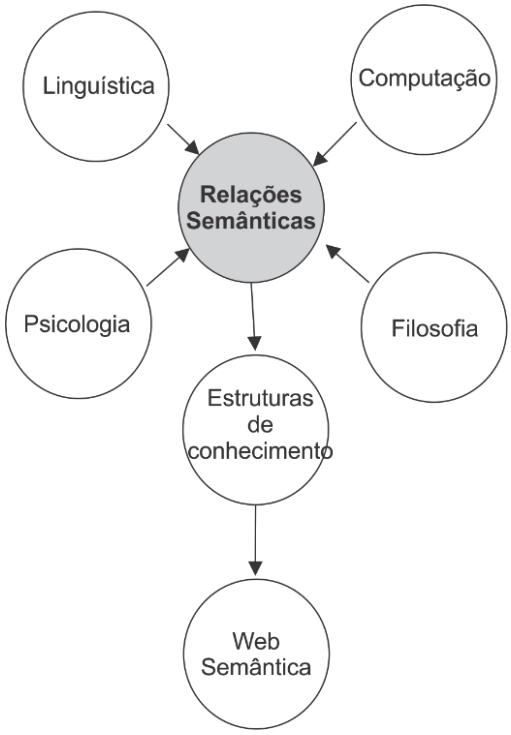
sobre o produto, enquanto uma base de dados farmacológica enfatizará os efeitos desse mesmo produto. Nesse caso, por mais que esses domínios sejam intimamente ligados e compartilhem muitos conceitos e relações, suas especificações contêm descritores e relações semânticas que refletem metas próprias, específicas de seus contextos.

3.3.1 Relações semânticas na Biblioteconomia e Ciência da Informação

As relações semânticas são de interesse de diversas áreas do conhecimento, tais como Psicologia e os estudos relacionados à memória semântica (CHAFFIN; HERRMANN, 1984), à Ciência da Computação (CC) e aos Sistemas de Informação (SI) (STOREY, 1993), à Linguística (MURPHY, 2003), entre outras. As relações semânticas permitem a criação de estruturas de representação do conhecimento que apoiam, por exemplo, a *Web* semântica (Figura 8), o que significa que, além de inferências humanas, inferências computacionais podem ser feitas, o que permite o avanço das estruturas do conhecimento.

Na Biblioteconomia e Ciência a Informação (BCI), Maculan (2015), em seu Apêndice A, apresenta uma proposta de sistematização de relacionamentos semânticos em tesauros que abrangem desde Vickery (1960) à ISO 25964 de 2011. Contudo, como as abordagens apresentadas pela autora foram especificamente de um tipo de SOC (tesauro), elas não foram abordadas nesta seção, com excessão de Dahlberg (1978), cuja classificação das relações semânticas pode ser utilizada em quaisquer representações de relações entre conceitos. Assim, como Dahlberg (1978), os demais autores apresentados nessa sub-seção abordam as relações semânticas em um contexto mais geral da Organização do Conhecimento (OC) e da Representação do Conhecimento (RC).

Figura 8 – Contribuições para/das relações semânticas de outras áreas do conhecimento



Fonte: Elaborada pela autora.

Para Green (2001), os relacionamentos envolvidos, sobretudo na OC, são numerosos e geralmente complexos. Essa magnitude e complexidade atrapalham seu uso consistente pelos profissionais da informação, assim como pelos usuários finais. Segundo a autora, muitas vezes não existe consenso sobre como tratar certos relacionamentos e os usuários finais não entendem o que está sendo comunicado pelo relacionamento e pelo padrão de notações de relacionamento utilizado.

Hjørland (2007) também afirma que o número de relações semânticas pode ser infinito e que diferentes domínios desenvolvem novos tipos de relações. E mais, segundo o autor, as contribuições sobre relações semânticas são muito diferentes e difíceis de apresentar de forma coerente porque elas não estão relacionadas umas com as outras ou sistematicamente relacionadas com visões mais amplas.

Contribuições acerca de relações semânticas sustentadas por autores da BCI foram importantes para compor o cenário sobre como as relações semânticas são abordadas nessa área. Nesse sentido, em 2001, Bean e Green editaram um livro com aportes de diferentes autores que tratam de relacionamentos semânticos em várias perspectivas da OC, tais como: relacionamentos bibliográficos,

relacionamentos em tesauros, em listas de cabeçalho de assunto, entre outros. Green (2008) complementa que os relacionamentos são o coração da OC.

Hjørland (2007) apresenta questões referentes ao *status* da pesquisa sobre relações semânticas na BCI. Entre as suas contribuições, o autor aborda a importância das garantias literária e do usuário para o estabelecimento das relações semânticas, sobretudo na OC. Nesse sentido, a abordagem de análise de domínio é bastante tradicional na identificação de relações semânticas, com base principalmente na garantia literária.

Café e Brascher (2011) apontam o uso de teorias linguísticas, como a teoria da valência, a gramática de casos, os gráficos conceituais e a teoria da gramática funcional de Simon Dik, para fundamentar os aspectos semânticos para base e estudos da representação dos conceitos na OC.

Bräscher (2014) discute as relações semânticas na construção de sistemas conceituais, considerando a importância dos relacionamentos em SOC semânticos. A autora explora a abordagem composicional como possibilidade de definir relações sintagmáticas. Essa abordagem, de acordo com a autora, baseia-se no princípio de que o significado de uma palavra pode ser analisado em unidades menores, chamadas características ou primitivas semânticas. Em sua análise, ela considera a lógica de predicados, a teoria da valência e a gramática de casos. Brächer conclui que estudos de relacionamentos sintagmáticos devem ser encorajados para contribuir para a construção de um referencial teórico para o desenvolvimento de SOC.

As relações semânticas são importantes também na recuperação da informação, tanto na estratégia de pergunta (*querying)* quanto na estratégia de navegação (*browsing)*. Nesse sentido, Khoo e Na (2006) acreditam que o processamento de linguagem natural e as relações semânticas, apontam o caminho a seguir para a Recuperação da Informação (RI). Segundo Green (2001), na estratégia de *browsing*, os assuntos[19] relacionados podem formar uma “teia” para o usuário navegar. Isso é importante para apoiar o usuário quando o documento (tanto buscado quanto recuperado) tratar de vários assuntos, quando o usuário falhar ao estabelecer o assunto e quando o número de assuntos requisitados por ele for muito grande. Por outro lado, de acordo com HJØRLAND (2007) e Khoo e Na (2006), na

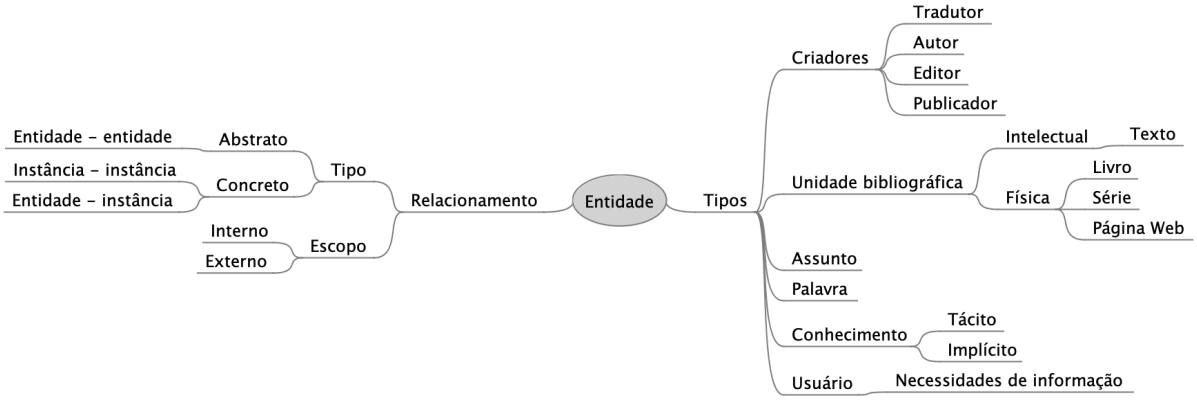
[19] Green (2001) usa “assunto”, já Khoo e Na (2006) utilizam “termo/conceito”. Nesse caso, optou-se por respeitar a terminologia dos autores em suas citações, mesmo as indiretas.

estratégia de *querying,* as relações semânticas podem ser empregadas para melhorar a revocação e a precisão. A melhoria da revocação, segundo Khoo e Na (2006), aumenta o número de documentos relevantes recuperados. Isso pode ser feito por meio da expansão de consulta, utilizando termos alternativos, especialmente com o uso de relações paradigmáticas, como os sinônimos. Hjørland (2007) acrescenta, nesse caso, a utilização de termos genéricos. Já a melhoria da precisão, na estratégia de *querying,* reduz a proporção de documentos não relevantes recuperados. Nesse caso, existe uma rigorosidade para que os conceitos e os relacionamentos inseridos na *query* sejam os mesmos nos documentos recuperados, ou seja, os documentos recuperados devem conter não somente os termos/conceitos especificados na consulta, como também expressar as mesmas relações entre os conceitos, como requeridos na *query*. Diante disso, normalmente as relações sintagmáticas são usadas (KHOO; NA, 2006). Para Hjørland (2007), a diferenciação de homônimos e a especificação dos termos relacionados podem contribuir para a melhoria da precisão na recuperação da informação.

As relações semânticas ainda podem acontecer em outras perspectivas além da associação entre dois conceitos. Para Green (2001), no que se refere à OC, os relacionamentos são entre coisas, formalmente representadas por entidades. Essas entidades podem ser os criadores de documentos físicos, unidades bibliográficas, assuntos, palavras, conhecimentos e usuários. Segundo a autora, ao combinar entidades (o que ela chama de entidades simples) por meio de relacionamentos, formam-se entidades mais complexas.

Para Green (2001), os relacionamentos entre entidades podem ser externos (entidade-entidade) e internos (entre uma entidade e seus elementos básicos). Outrossim, a autora distingue relacionamentos abstratos, ou seja, entre entidades (exemplo: pessoa *nasceu_em* cidade) de relacionamentos concretos, isto é, que envolvem instâncias (por exemplo: João nasceu_em Recife). A Figura 9 apresenta um mapa mental sobre entidades e relacionamentos, de acordo com a abordagem de Green (2001).

Figura 9 – Mapa conceitual sobre entidade e relacionamentos



Fonte: Elaborada pela autora a partir de Green (2001).

Visto que existem diferentes tipos de entidades, como pode ser verificado na Figura 9, é importante destacar que, para cada um destes tipos, diferentes abordagens de relacionamentos semânticos podem ocorrer. Assim, Green (2001) agrupa os relacionamentos em OC em quatro áreas: (1) relacionamentos entre unidades de conhecimento registrado, que são baseadas em descrições bibliográficas destas unidades; (2) relacionamentos intratextuais e intertextuais, incluindo relacionamentos baseados na estrutura do texto, relacionamentos de citação e *links* hipertextuais; (3) relacionamentos de assuntos em tesauros e outras estruturas classificatórias; e (4) relações de relevância.

Nesta tese, o foco são os relacionamentos de assuntos em estruturas classificatórias, item 3. Nesse caso, segundo Green (2001), quando se trata de relacionamento de assuntos, é importante considerar os relacionamentos entre os conceitos. Os conceitos, nessa concepção, são expressos por palavras ou por esquemas notacionais.

Os relacionamentos entre conceitos em estruturas classificatórias são determinados a partir da análise dos conceitos. Isto envolve determinar: as características do conceito, a posição que o conceito ocupa no domínio e a relação do conceito com outros conceitos do sistema conceitual (BRÄSCHER, 2014). Para Dalhberg (1978) e Cabré Castellví (1999), sempre que existir características idênticas entre os conceitos, deduz-se que eles se relacionam.

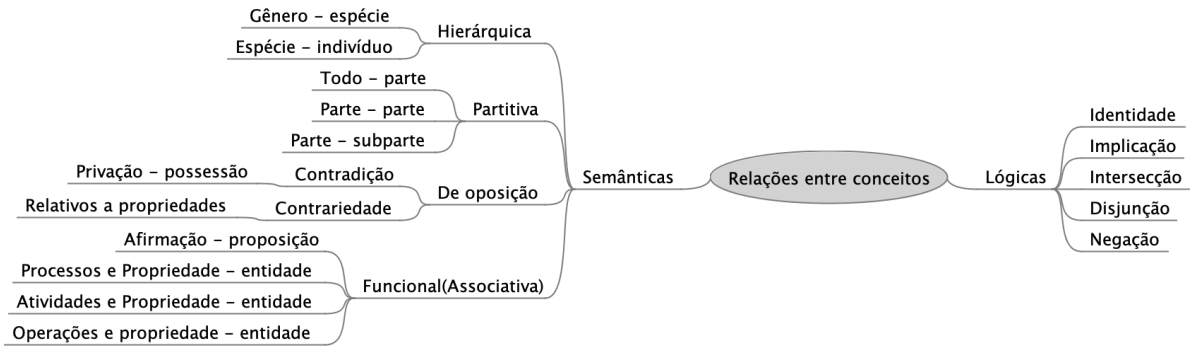
Nesta seção explorou-se pontos de vista sobre relações semânticas em contextos da BCI. A próxima seção apresenta as classificações de relações semânticas apresentadas tanto por autores da BCI quanto por autores de outras áreas.

3.3.1.1 Classificações de relações semânticas encontradas na literatura

Alguns autores apresentam em seus trabalhos as relações semânticas em forma de taxonomia, explicitando a classificação das relações ou apresentando as relações semânticas de maneira que elas possam ser representadas por taxonomias.

Sob um ponto de vista cronológico, a primeira referência sobre relações semânticas entre conceitos encontrada na literatura pesquisada remete a Dahlberg (1978). Em sua Teoria do Conceito, amplamente utilizada na BCI, ela classifica os relacionamentos em lógicos e semânticos, conforme apresentado na Figura 10.

Figura 10 – Relações entre os conceitos de Dahlberg



Fonte: Elaborada pela autora a partir de Maculan (2015) e Dahlberg (1978a, 1978b e 1978c)

As relações lógicas (ver Quadro 2) são baseadas nas características comuns entre os conceitos. De acordo com a autora, esse tipo de relação é muito importante, pois, a partir dele, é possível comparar os conceitos, organizá-los e relacioná-los semanticamente. Já as relações semânticas sugeridas por Dahlberg (1978) são: hierárquica, partitiva, de oposição e funcional. No Quadro 3 há uma breve explicação dessas relações e suas intersecções com as relações lógicas, bem como exemplos.

Quadro 2 – Relações lógicas entre os conceitos

Relação	Características		Explicação
	Conceito A	Conceito B	
Identidade	X, X, X	X, X, X	As características dos conceitos são as mesmas;
Implicação	X, X	X, X, X	O conceito A está contido no conceito B;
Intersecção	X, X, O	X, O, O	Os dois conceitos coincidem algum elemento;
Disjunção	X, X, X	O, O, O	Os conceitos se excluem mutuamente. Nenhuma característica em comum;
Negação	X, X, O	O, X, O	O conceito A inclui uma característica cuja negação se encontra em B.

Fonte: Elaborado pela autora a partir de Dahlberg (1978a).

Quadro 3 – Relações Semânticas sugeridas por Dahlberg

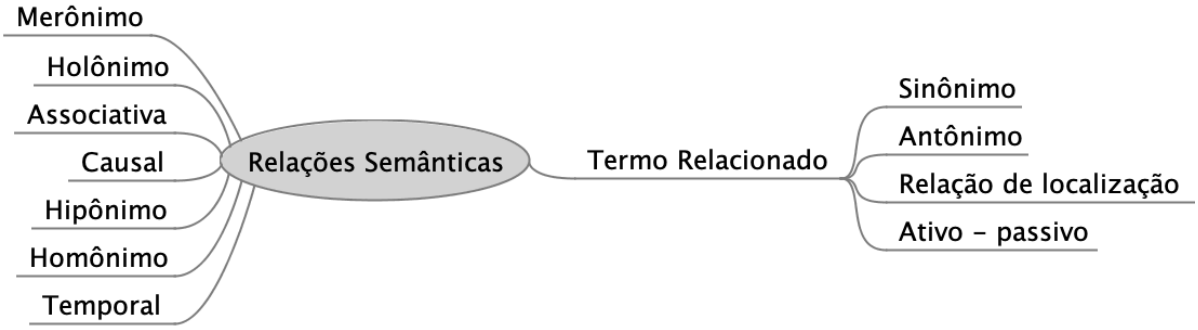
Tipo de relação	Intersecção com relação lógica	Breve explicação	Exemplo
Hierárquica	Implicação	Neste tipo de relação, os conceitos possuem as mesmas características, porém há aqueles que possuem características específicas.	Árvore árvore frutífera macieira pereira pessegueiro árvore de nozes amendoeira aveleira
Partitiva		Diz respeito ao todo e às suas partes.	Árvore raízes, tronco, galhos, folhas, flores, frutos
De oposição	Negação	Os conceitos se contradizem ou contrapõem um outro conceito.	numérico - não numérico presente - ausente branco – preto
Funcional	Intersecção	São aplicadas a conceitos que expressam processos.	Medição - objeto medido - fins da medição - instrumento de medição - graus de medição

Fonte: Elaborado pela autora a partir de Dahlberg (1978a).

Broughton *et al*. (2005) apresentam um conjunto de relações semânticas que consideram importantes em SOC (ver Figura 11), sobretudo relações que podem ser observadas em tesauros. Nesse caso, os autores acrescentam às relações semânticas apresentadas por Dahlberg (1978) os relacionamentos associativos, causais, homônimos e temporais. Eles também criam uma classe de relações chamada de Termos Relacionados que inclui antônimos, sinônimos, relações de

localização e a relação ativo/passivo. O Quadro 4 apresenta um resumo destas relações.

Figura 11 – Relações Semânticas de Broughton



Fonte: Elaborada pela autora a partir de Broughton *et al*. (2005).

Quadro 4 – Classificação de relações semânticas propostas por Broughton *et al*.

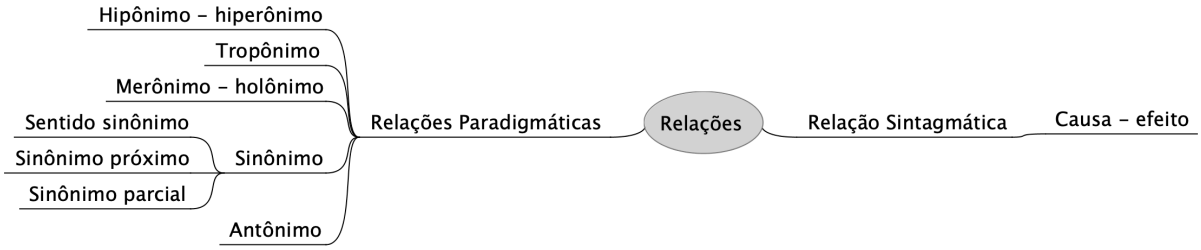
Relação		Definição
Hipônimo		Relações gênero-espécie e tipo-de.
Merônimo		Relação partitiva. Uma relação entre o todo e suas partes (A é parte de B). Um merônimo é o nome de uma parte constituinte, substância de, membro de algo.
Holônimo		Oposto de merônimo. B tem A como parte.
Associativa		A é mentalmente associado com B por alguém. Geralmente as relações associativas são relações não especificadas em alguns instrumentos, como a taxonomia.
Causal		A é a causa de B.
Homônimo		Dois conceitos A e B são expressos pelo mesmo símbolo.
Temporal		Uma relação semântica a qual um conceito indica um tempo ou período de um evento designado para outro conceito. Ex.: Segunda Guerra Mundial (1939-1945) (BROUGHTON *et al.*, 2005, p. 143).
Termo relacionado	Sinônimo	A denota o mesmo de B; A é equivalente a B.
	Antônimo	Relação de oposição.
	Relação de localização	Uma relação semântica na qual o conceito indica a localização de algo indicado por outro conceito. A está localizado em B.
	Ativo - passivo	Um conceito expressa a execução de uma operação ou processo afetando o outro conceito.

Fonte: Elaborado pela autora a partir de Broughton *et al*. (2005).

Por sua vez, Khoo e Na (2006) discutem a natureza das relações semânticas, principalmente daquelas expressas em textos, e de suas aplicações na BCI sob o

olhar da Linguística e da Psicologia. Como pode ser visto na Figura 12, os autores evidenciam cinco relações paradigmáticas (também chamadas relações semânticas ou relações lexicais) e uma relação sintagmática[20]. É importante ressaltar que a relação sintagmática de causa-efeito, conforme pode ser visto na Figura 12, foi considerada anteriormente por Broughton *et al.* (2005) como uma relação paradigmática. Em prosseguimento, Khoo e Na nomeiam a relação chamada hierárquica por Dahlberg (1978) e hipônimo por Broughton *et al*. (2005) de relação hipônimo-hiperônimo. De acordo com os autores, hipônimo denota o termo mais específico e o hiperônimo o termo mais genérico. Geralmente esse relacionamento é assinalado com as expressões "é-um" e "tipo-de". Já o tropônimo, incluído por Khoo e Na, é atribuído a relações entre verbos (essas relações não serão tratadas nesta tese). Outra mudança com respeito às relações semânticas apresentadas anteriormente é a classificação dos sinônimos em: sentido sinônimo (compartilham o mesmo sentido), sinônimo próximo (os significados são próximos) e sinônimo parcial (compartilham o mesmo sentido, mas são diferentes em alguns aspectos).

Figura 12 – Relações semânticas apontadas por Khoo e Na



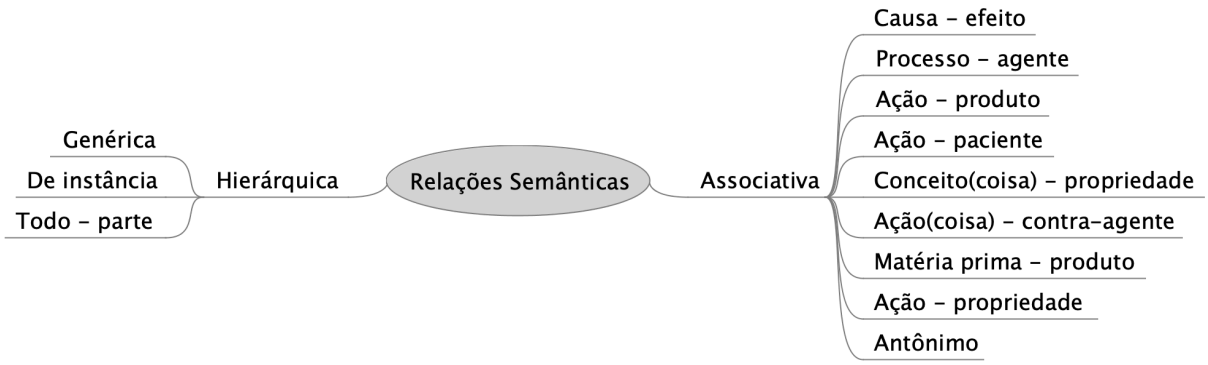
Fonte: Elaborada pela autorae a partir de Khoo e Na (2006).

Zeng (2008) apresenta um conjunto de relações semânticas que são utilizadas no contexto dos SOC. A autora agrupa as relações em duas classes: hierárquica e associativa. Enquanto Dahlberg (1978) separa as relações hierárquica e partitiva, Zeng aponta que a relação partitiva é uma relação hierárquica e que, além de partitiva, a relação hierárquica também pode ser de gênero-espécie (ou genérica) e de instância. A relação de instância, conforme apontado por Green (2001), é uma relação concreta entre uma entidade e um indivíduo. Como pode ser observado na Figura 13, a relação causa-efeito é considerada por Zeng (2008) como uma relação

[20] Buscou-se, nesta tese, respeitar a colocação dos autores quanto à denominação das relações sintagmáticas e paradigmáticas. De acordo com a ISO 25964-1 (2011), as relações sintagmáticas são aquelas que ocorrem em um contexto em particular. Já as paradigmáticas (hierárquicas, associativas e equivalentes) são válidas para quase todos os contextos.

semântica do tipo associativa, diferente de Khoo e Na (2006), que a classificaram como uma relação sintagmática. Outrossim, os antônimos foram classificados como um tipo de relação associativa. Outras relações associativas apontadas por Zeng são exemplificadas no Quadro 5.

Figura 13 – Relações semânticas conforme Zeng



Fonte: Elaborada pela autora a partir de Zeng (2008).

Quadro 5 – Exemplo de relacionamentos associativos

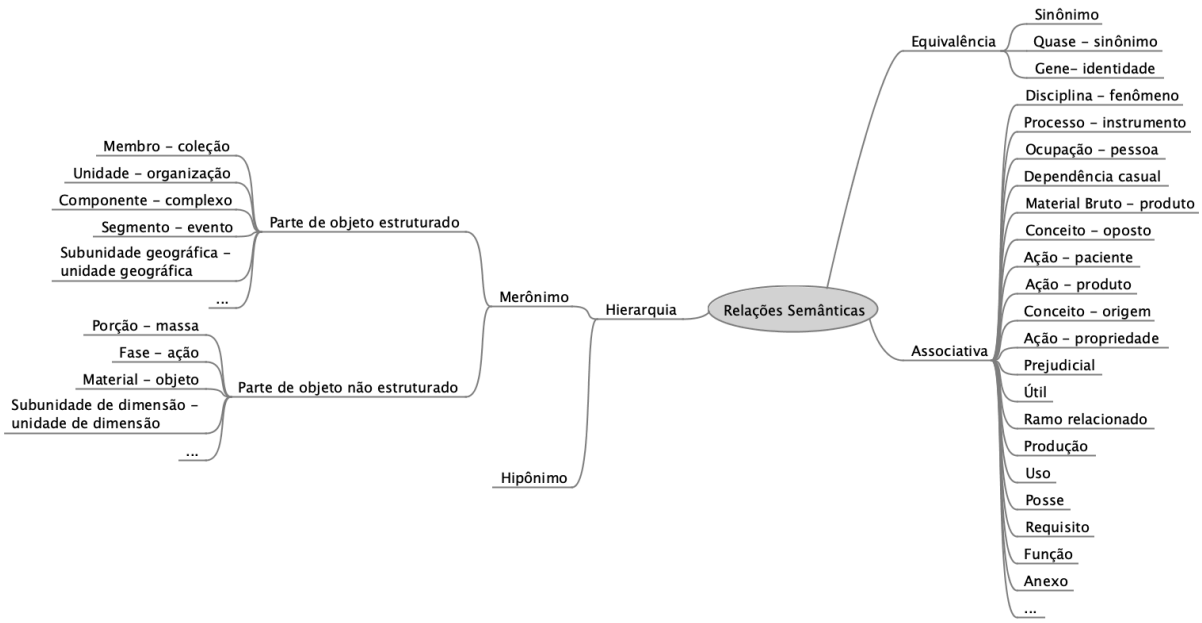
Relações	Exemplos
Causa - efeito	Acidente - ferimento
Processo - agente	Medição de velocidade - medidor de velocidade
Ação - produto	Escrever - publicação
Ação - paciente	Ensinar - estudante
Conceito ou coisa - propriedade	Liga de aço - resistência à corrosão
Coisa ou ação - contra-agente	Peste - pesticida
Matéria prima - produto	Uva - vinho
Ação - propriedade	Comunicação - habilidades de comunicação
Antônimo	Solteiro - casado

Fonte: Zeng (2008, tradução nossa).

Peters e Weller (2008) abordam tanto as relações paradigmáticas quanto as sintagmáticas em SOC, sobretudo no contexto da folksonomia e da ontologia. As autoras acreditam na necessidade de repensar as relações semânticas com respeito à sua classificação e generalização para o uso na RC e na RI, pois, assim como para

Hjørland (2007), para Peters e Weller (2008) algumas vezes falta consenso sobre os relacionamentos semânticos. Tal como Zeng (2008), Peters e Weller (2008) também agrupam os relacionamentos semânticos em hierárquicos e associativos. Além disso, elas criam uma classe de relações de equivalência para reunir os sinônimos, os quase sinônimos e uma relação que elas chamam de gene-identidade, baseada na ontologia *Gene Ontology*. Como pode ser visto na Figura 14, as autoras dividem as relações partitivas em relações em que as partes se referem a partes de um objeto estruturado e a partes de um objeto não estruturado. Já as relações associativas agrupam todas as outras relações que não são hierárquicas nem de equivalência. Uma observação sobre a taxonomia de Peters e Weller (2008) é que, nas relações de merônimo (partitivas) e nas associativas, as autoras não esgotam as possibilidades de incluir novos tipos nas classificações das relações.

Figura 14 – Taxonomia de relações semânticas de Peters e Weller

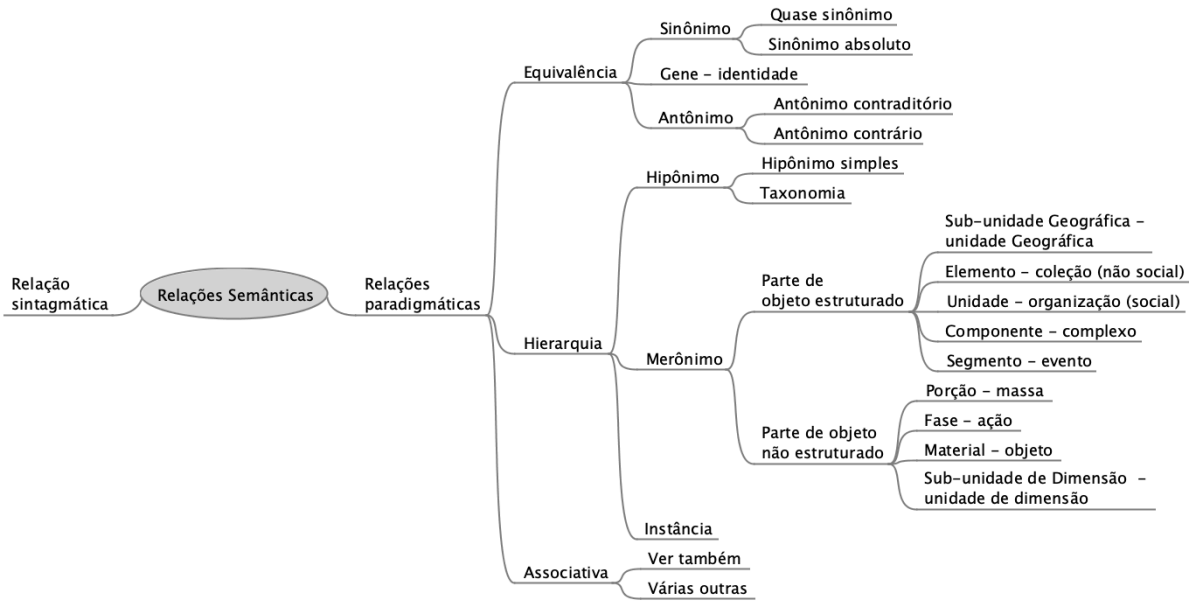


Fonte: Elaborada pela autora a partir de Peters e Weller (2008, p. 105).

Stock (2010), além de versar sobre as relações semânticas no contexto da BCI, também faz um estudo minucioso dos conceitos, trazendo várias perspectivas que abrangem a sua teoria do conceito. No que tange às relações semânticas, o autor associa-as às propriedades de transitividade, simetria e reflexividade (que serão explicadas na seção 3.3.3). Entre as várias conclusões que o autor chega sobre conceitos e relações semânticas na BCI, ele pondera sobre a importância de

considerar a teoria do conceito e a teoria das relações para a criação de vocabulários controlados, para o uso na *web* social (via *tagging* e folksonomia) e para a *web* semântica (por meio das ontologias). Na Figura 15, em que se apresenta a taxonomia das relações mais comuns na BCI, ainda segundo Stock (2010), as relações sintagmáticas são classificadas como um tipo de relação semântica, diferente de Khoo e Na (2006), por exemplo. Segundo o autor, a relação sintagmática é a única que ocorre na folksonomia. Ele também classifica as relações paradigmáticas em três grandes grupos: equivalência, hierarquia e associativa. Contudo, como pode ser visto, Stock aponta outras relações e cria outras subdivisões para essas classes. Diferente de Zeng (2008), que classifica os antônimos como um relacionamento associativo, e de Broughton *et al*. (2005) que os classificam como termos relacionados, Stock (2010) classifica os antônimos como uma relação de equivalência. Mais ainda, ele separa os antônimos em dois tipos: antônimo contraditório e antônimo contrário. Segundo o autor, no caso dos antônimos contraditórios existem dois extremos contrários, sem nada no meio. Uma mulher está grávida ou não está grávida, por exemplo. Já os antônimos contrários permitem outros valores entre os extremos, por exemplo: entre amor e ódio pode residir a indiferença.

Figura 15 – Relações semânticas conforme Stock

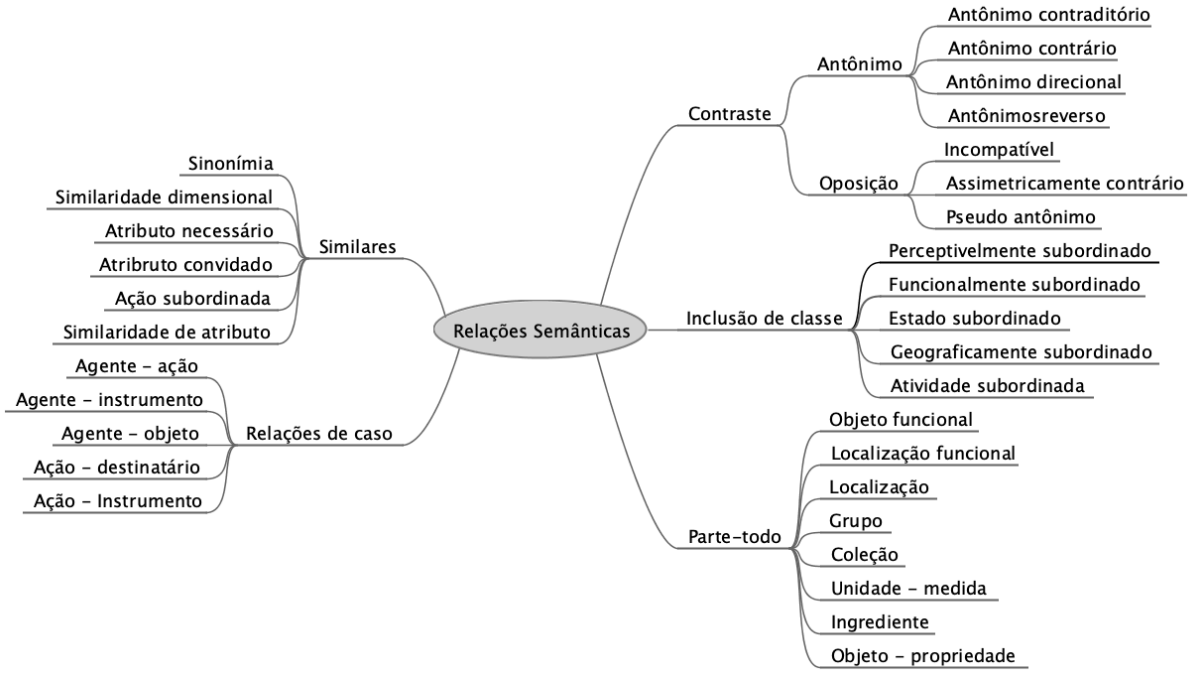


Fonte: Elaborada pela autora a partir de Stock (2010).

Nas relações hierárquicas, Stock (2010), assim como Zeng (2008), menciona as relações de instância. Nas relações de hipônimo, ele separa dois grupos: hipônimo simples e taxonômico. Nesse caso, o hipônimo simples é o relacionamento "*é-um*" e o taxonômico é o relacionamento "*é-um-tipo-de*", tais como as relações entre os *táxons* na Biologia, que indicam o parentesco da evolução dos organismos vivos. Do mesmo modo como em Peters e Weller (2008), Stock separa os merônimos em dois grupos de objetos estruturados e não estruturados. Por fim, as relações associativas são relações do tipo "ver também", utilizadas em tesauros e todas as demais relações que não são nem hierárquicas e nem de equivalência.

Em outros contextos à parte da BCI, duas taxonomias muito importantes foram apresentadas por Chaffin e Herrmann (1984) e por Storey (1993). Chaffin e Herrmann, autores ligados à Psicologia, organizaram uma taxonomia de trinta e uma relações semânticas, sugeridas anteriormente por outros autores, em cinco famílias (contraste, inclusão de classe, similares, relações de caso e parte-todo), conforme apresentado na Figura 16. Resumidamente, as relações similares reúnem relações onde os significados dos termos se sobrepõem, as relações de caso envolvem atribuições, as relações de contraste denotam conceitos que se opõem, as relações de inclusão de classe são relações "*tipo de*" e as relações todo-parte abarcam o conceito e seus constituintes.

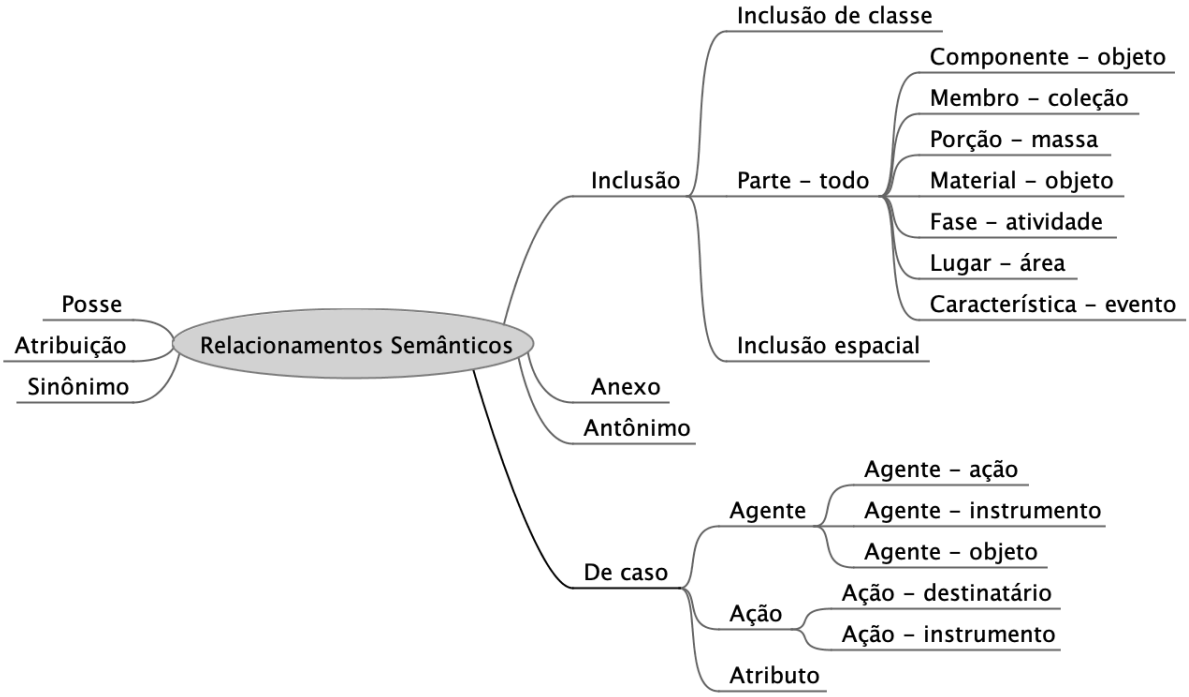
Figura 16 – Taxonomia das relações semânticas segundo Chaffin e Herrmann



Fonte: Elaborada pela autora a partir de Chaffin e Herrmann (1984).

Storey (1993), autora da área de Computação e Sistemas de Informação, fez uma taxonomia (ver Figura 17) considerando as relações apresentadas anteriormente por Chaffin e Herrmann (1984) e outros autores para uso em bancos de dados. Ela acrescentou à taxonomia de Chaffin e Herrmann as relações de posse, atribuição, anexo e inclusão espacial (subclasse dos relacionamentos de inclusão). Em síntese, os relacionamentos de posse são aqueles que indicam propriedade, os relacionamentos de atribuição referem-se a um objeto e seu atributo, os relacionamentos de anexo são aqueles onde um objeto é conectado a outro e o relacionamento de inclusão espacial descreve situações onde um objeto está em um local mas não é parte daquele local, tal como no exemplo “cliente está na recepção”.

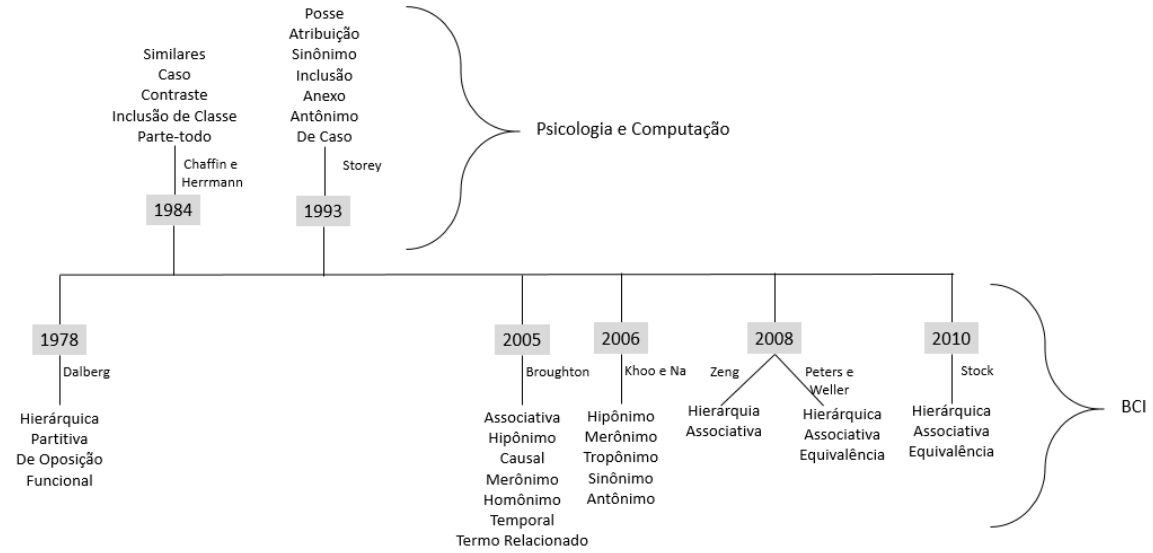
Figura 17 – Taxonomia dos relacionamentos semânticos segundo Storey



Fonte: Elaborada pela autora a partir de Storey (1993).

Como foi possível perceber, existem muitos tipos de relações semânticas para a representação do conhecimento. Observa-se que cada autor classifica-as conforme seu ponto de vista a partir de relações já verificadas em seus sistemas, estudos e pesquisas. A Figura 18 mostra uma linha do tempo com os autores apresentados e as respectivas classes de relações apontadas por eles.

Figura 18 – Linha do tempo das relações semânticas



Fonte: Elaborada pela autora.

Nesta seção explorou-se as classificações de relações semânticas sob o ponto de vista de autores encontrados na bibliografia. A próxima seção apresenta uma proposta de taxonomia para a BCI de maneira a compilar as classificações de relações semânticas apresentadas pelos autores.

3.3.2 Proposta de uma Taxonomia de Relações Semânticas para a Biblioteconomia e Ciência da Informação

Com base nas taxonomias de relações semânticas sugeridas pelos autores Dahlberg (1978), Chaffin e Herrmann (1984), Storey (1993), Broughton *et al.* (2005), Khoo e Na (2006), Zeng (2008), Peters e Weller (2008) e Stock (2010), propõe-se uma taxonomia que agrega todas as propostas apresentadas na literatura pesquisada. Esta taxonomia foi publicada nos anais do XVIII Encontro Nacional de

Pesquisa em Ciência da Informação (Enancib) 2017[21] e na revista Tendências da Pesquisa Brasileira em Ciência da Informação[22].

A estrutura da taxonomia que se propõe subdivide-se em relações hierárquicas, de equivalência e associativas, conforme apresentado na Figura 19. As seções que seguem descrevem esses tipos mais detalhadamente.

Figura 19 – Relações semânticas

Fonte: Elaborada pela autora.

3.3.2.1 Relações hierárquicas

De acordo com Stock (2010), a hierarquia é a relação mais importante em um SOC. Entende-se que todo o processo de classificação da organização e representação do conhecimento inicia-se com a relação hierárquica, em que os conceitos são agrupados em classes. Como corrobora Campos,

> A relação hierárquica é uma das relações principais em qualquer estrutura classificatória. Ela é a que forma a espinha dorsal de uma estrutura. [...] [E]la é imprescindível, sendo a partir dela que se estabelece o primeiro elemento de uma definição. (CAMPOS, 2004, p. 29)

À parte de reflexões filosóficas e com base na literatura pesquisada, as relações hierárquicas propostas nesta taxonomia são agrupadas em dois conjuntos: hipônimo-hiperônimo e merônimo-holônimo. Essa estrutura de divisão foi baseada em Zeng (2008), Peters e Weller (2008) e Stock (2010). As relações de hipônimo-hiperônimo são detalhadas em: hipônimo simples, instância, inclusão de espacial e

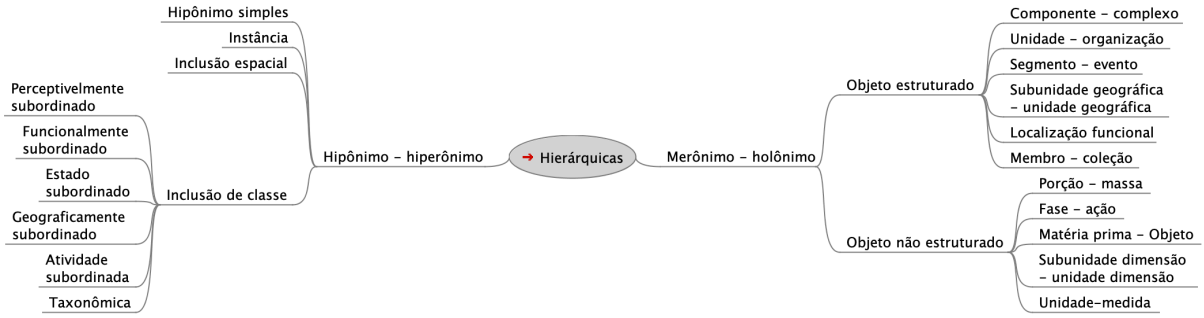
---

[21] MAIA, L. S.; LIMA, G. A. ; MACULAN, B. C. M. S. Taxonomia dos tipos de relações semânticas para a Organização e Representação do Conhecimento: Uma proposta a partir da literatura. *XVIII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB),* 2017. Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/view/334>. Acesso em: 13 jan. 2019.

[22] MAIA, L. S.; LIMA, G. A. ; MACULAN, B. C. M. S. Taxonomia dos tipos de relações semânticas para a Organização e Representação do Conhecimento: Uma proposta a partir da literatura. *Tendências da Pesquisa Brasileira em Ciência da Informação*, v. 10, p. 1, 2017. Disponível em: <http://www.brapci.inf.br/index.php/res/v/104795>. Acesso em: 13 jan. 2019.

inclusão de classe. Por outro lado, as relações de merônimo-holônimo são subdividas observando a estrutura do objeto analisado, ou seja, relações em que o objeto é estruturado e quando o objeto não é estruturado. A proposta da taxonomia dessas relações está apresentada na Figura 20.

Figura 20 – Relações hierárquicas



Fonte: Elaborada pela autora.

Das relações semânticas apresentadas nesta tese, as propostas por Broughton *et al*. (2005) e Khoo e Na (2006) não subdividem as relações hierárquicas. Já nas relações propostas por Zeng (2008), as relações hierárquicas são especificadas como relações genéricas, de instância e todo-parte. Essas últimas são detalhadas por Peters e Weller (2008) na estrutura das relações hierárquicas, mas as autoras não apontam as relações de instância.

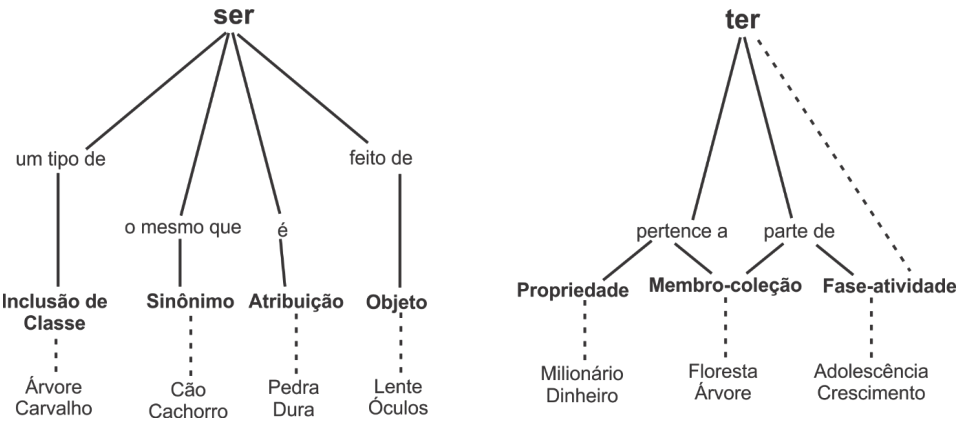
Stock (2010), além de detalhar as relações de merônimo (todo-parte) e apontar as relações de instância, detalha as relações hierárquicas em hipônimo simples e taxonomia.

Storey (1993), por sua vez, não especifica as relações hierárquicas tal como os autores apresentados acima, talvez porque as pesquisas apresentadas anteriormente estejam no contexto da BCI e Storey apresente as relações semânticas na visão da computação. Contudo, Storey apresenta a relação de inclusão como uma relação hierárquica. Essa relação é detalhada em inclusão de classe, inclusão espacial e merônimo. De acordo com a autora, as relações de inclusão são aquelas em que um tipo de entidade compreende ou contém outro tipo de entidade. Já a inclusão de classe são os relacionamentos subtipo/supertipo, geralmente expressos como “é um”. Logo “A é um B” significa que A é referido como

um tipo de entidade específica e B um tipo de entidade genérica. Além disso, a inclusão de classe também pode ser expressa como “tipo de”.

Storey (1993), agrega o uso dos verbos “ser” e “ter” para indicar relações de indicar algumas relações semânticas. Nesse sentido, a expressão “é um" é usada nos relacionamentos de inclusão de classe, sinônimo, atribuição e matéria-prima (*stuff*). Já o verbo ter pode ser usado para expressar “pertence a" e alguns relacionamentos parte-todo. A Figura 21 mostra os usos do verbo “ser” e do verbo “ter”, conforme Storey.

Figura 21 – Usos dos verbos *ser* e *ter* nos relacionamentos semânticos



Fonte: Storey (1993, p. 477, tradução nossa)

De acordo Arnold e Rahm (2014), a expressão “é um" apresenta variações e pode se tornar complexa, como no exemplo "X é uma variação de Y". Além disso, expressões “é um" podem aparecer em textos em linguagem natural por meio de advérbios – tais como comumente, geralmente, tipicamente – e expressões – como classe de, forma de ou pedaço de (coletivos e partitivos). Outrossim, essa expressão pode ocorrer no singular ou no plural ("é um" ou "são um") e vir com diferentes determinadores, como o artigo definido o/a. Outrossim, os autores perceberam que as relações de hipônimo-hiperônimo não necessariamente se restringem à expressão padrão “é um”, podendo abarcar as seguintes expressões: “é qualquer forma de”, “é uma classe de”, “é comumente uma variedade de”, “descreve um”, “é definido como um” e “é usado para qualquer tipo de”.

3.3.2.1.1 Relação Hipônimo-Hiperônimo

A relação hipônimo-hiperônimo[23] seguiu a denominação dada por Khoo e Na (2006), pelo entendimento de que essa designação é mais formal. Nessa relação, o hiperônimo é o termo mais genérico e hipônimo é o termo mais específico. Nesse sentido, o hipônimo herda características do hiperônimo (STOCK, 2010).

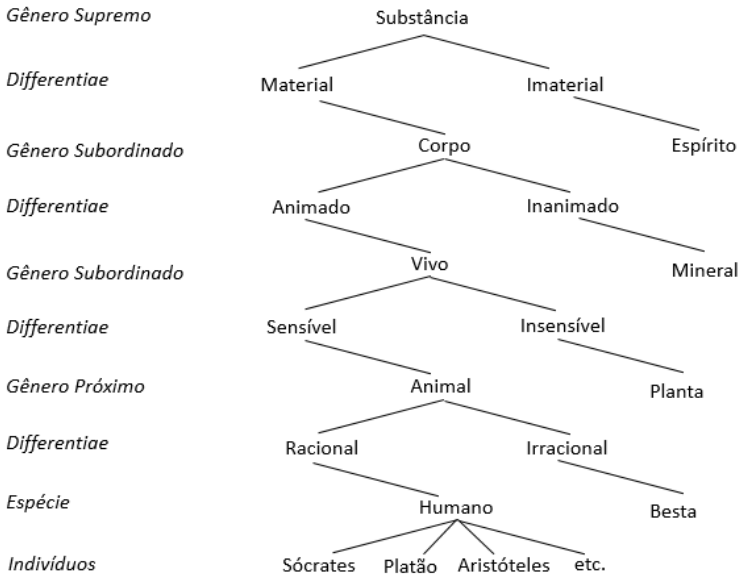
Em Dahlberg (1978), entende-se que a sua definição de relações hierárquicas refere-se especificamente à relação hipônimo-hiperônimo. De acordo com a autora, na hierarquia (ou seja, no hipônimo-hiperônimo), os membros de uma classe têm características comuns às classes gerais e algumas características específicas, ou seja, os membros das classes mais específicas herdam características das classes gerais. No nível dos conceitos, as características são os atributos ou elementos dos conceitos.

Por sua vez, em Mazzocchi (2017), entende-se que a relação hipônimo-merônimo corresponde à relação gênero-espécie. Segundo o autor, essa relação liga um gênero com suas espécies ou uma classe com suas subclasses. Para ele, uma importante propriedade dessas relações é a herança de características: qualquer atributo do gênero (hiperônimo) também deve ser atribuído à espécie (hipônimo). Ainda de acordo com Mazzocchi, essa relação foi historicamente retratada na Árvore de Porfírio (Figura 22). Porfírio, um filósofo neoplatônico do século III, seguiu o método para realizar a divisão correta dos gêneros supremos (ou categorias) para as espécies individuais. Logo, a árvore ramificada reflete a ordem hierárquica, que é obtida por uma série de bifurcações que começa com o gênero supremo da substância, em que características diferenciais essenciais são adicionadas a uma espécie. Segundo o autor, esse é o elemento principal para distinguir uma espécie de todas as outras espécies que compartilham o mesmo gênero. Por fim, para Mazzocchi, a relação gênero-espécie é vista como verdadeiramente definitiva e independente do contexto.

---

[23] Nesta tese, essa relação algumas vezes será chamada apenas de hipônimo ou apenas hiperônimo, mas com o sentido de hipônimo-hiperônimo.

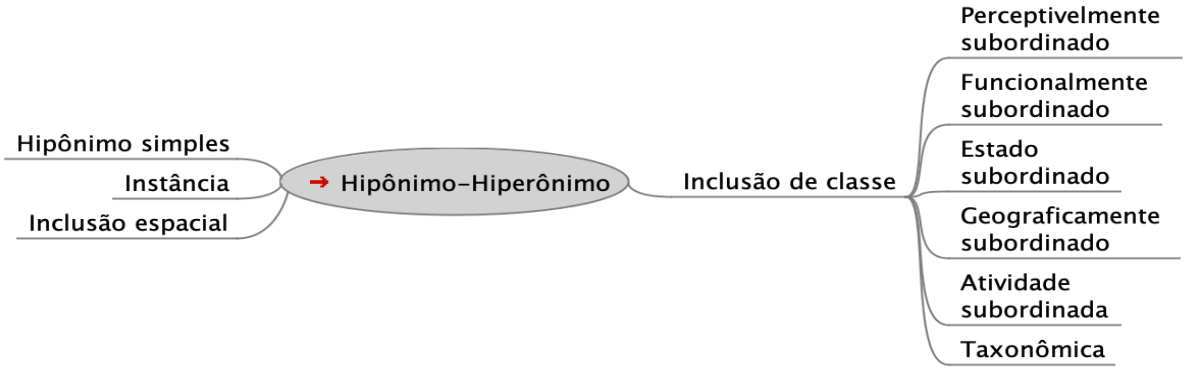
Figura 22 – Árvore de Porfírio



Fonte: Mazzocchi (2017, p. 372, tradução nossa).

Além da nomeação gênero-espécie, outros nomes foram encontrados na bibliografia para a relação hipônimo-hiperônimo, tais como: relação “é um”, “tipo de”, relação genérica, relação taxonômica, classe/subclasse, relações de inclusão, inclusão de classe, subordinado hierarquicamente e superordinado/subordinado (HJØRLAND, 2007; KHOO & NA, 2006). Algumas dessas denominações foram definidas nessa taxonomia como um tipo de relação de hipônimo-hiperônimo, conforme destacado na Figura 23.

Figura 23 – Relação hipônimo-hiperônimo



Fonte: Elaborada pela autora.

A relação de hipônimo simples foi proposta por Cruse (2002) e por Stock (2010). De acordo com os autores, essa relação é denotada pelo padrão "é um", por

exemplo: “A rainha *é uma* mulher”. Nesse caso, conforme Stock (2010) observa, ficaria estranho dizer que “A rainha *é um tipo de* mulher”.

A relação de instância, na bibliografia pesquisada, foi proposta por Zeng (2008), Stock (2010) e também está descrita na ISO 25964-1 (2011). Em todas essas propostas, o relacionamento de instância é um tipo de relação hierárquica. Essa relação pode ser rotulada como "*é um*" ou "*é parte de*". A relação de instância é definida por Green (2001) como uma relação concreta. Logo, uma instância é entendida como um objeto nomeado de uma entidade, ou um exemplar de uma entidade, como, por exemplo: João *é uma* pessoa. As relações de inclusão de classe (geograficamente subordinada) e a relação de merônimo-holônimo (subunidade geográfica-unidade geográfica), podem ser vistas como uma relação de instância. Contudo, decidiu-se criar e aplicar a seguinte regra: as relações supracitadas de inclusão de classe e merônimo têm prioridade sobre a relação de instância. Assim, em uma relação de instância cujo(s) conceito(s) remete(m) a espaços geográficos, deverá ser verificado se a relação é uma das relações acima. Se sim, elas são classificadas com as respectivas denominações; se não, elas são classificadas como uma relação de instância. O Quadro 6 mostra exemplos de entidades e suas respectivas instâncias.

Quadro 6 – Exemplos de entidades e instâncias

Entidade	Instância
Pessoa	João
Usuário	Maria
Livro	*Relationships in the Organization of Knowledge*
Artigo	*Semantic Relations in Knowledge Organization Systems*
País	Brasil
Cidade	Belo Horizonte

Fonte: Elaborado pela autora.

Nas classificações de relações semânticas apresentadas pelos autores da BCI, nenhuma delas contemplou as relações de inclusão de classe e de inclusão espacial. Contudo, Cruse (2002) pontua que um hipônimo-hiperônimo nunca ocorre isoladamente. Segundo o autor, a relação de hipônimo envolve inclusão. Nesse sentido, uma classe de maçãs, por exemplo, está incluída na classe de frutas, o que

permite concluir que o significado de fruta está incluído no significado de maçã. Logo, decidiu-se incorporar às relações de hipônimo-hiperônimo as relações de inclusão de classe e inclusão espacial.

Segundo Winston, Chaffin e Herrmann (1987), as relações de inclusão de classe podem ser rotuladas por "*é um*" ou "*é um tipo de*". Chaffin e Herrmann (1984) classificam essas relações em: (1) perceptivelmente subordinado, objetos caracterizados principalmente por suas propriedades físicas visíveis (por exemplo: cavalo *é um* animal ou cavalo *é um tipo de* animal); (2) funcionalmente subordinado, que envolve objetos caracterizados por suas funções (por exemplo: carro *é um* veículo ou carro *é um tipo de* veículo); (3) estado subordinado, que envolve o estado das coisas (por exemplo: medo *é uma* emoção ou medo *é um tipo* de emoção); (4) atividade subordinada, que envolve atividades (xadrez *é um* jogo ou xadrez *é um tipo de* jogo); e (5) geograficamente subordinado, que envolve lugares (América *é um* continente)[24].

Nas relações de inclusão de classe da taxonomia proposta nesta tese, inseriu-se as relações taxonômicas**.** A regra para classificar um par de conceitos como uma relação taxonômica dentro das relações de inclusão é a seguinte: se a relação no par de conceitos denotar inclusão de classe e se, contudo, ela não puder ser classificada nos subtipos: perceptivelmente subordinado, funcionalmente subordinado, estado subordinado, atividade subordinada ou geograficamente subordinado. Nas classificações das relações semânticas da bibliografia pesquisada, Stock (2010) menciona a relação taxonômica. Ele tem como base a definição de Cruse (2002), que, por sua vez, refere-se à relação espécie-indivíduo proposta por Dahlberg (1978). Nesse caso, a espécie é um conjunto de indivíduos que apresentam características comuns (tais como morfológicas, anatômicas, bioquímicas e fisiológicas), semelhanças entre si e pertencem a uma mesma família ou gênero. Cruse (2002) ressalta que na relação taxonômica é importante considerar a natureza das subcategorias resultantes e suas relações umas com as outras à luz da proposta da taxonomização.

Por fim, a inclusão espacial descreve situações onde um objeto está situado em outro, mas não é parte desse em que está situado (KUCZORA & COSBY, 1989; STOREY, 1993; WINSTON, CHAFFIN & HERRMANN, 1987). Storey (1993)

---

[24] Nesse caso, esse exemplo também pode ser classificado como uma relação de instância.

exemplifica com a seguinte sentença: “O cliente *está na* recepção”. Nesse contexto, o cliente se encontra na recepção mas ele não é parte dela.

3.3.2.1.2 Relações de Merônimo-Holônimo

Na bibliografia, vários nomes podem ser encontrados para a relação merônimo-holônimo, tais como: parte-todo, todo-parte, partição, partitiva, mereológica, meronímia e partonímia (BEGHTOL, 2001). A nomeação adotada na taxonomia proposta segue a sugestão de Khoo e Na (2006), pois se considera que o nome merônimo-holônimo é mais objetivo para descrever a relação.

De acordo com Beghtol (2001), existe um consenso sobre entendimento da relação de merônimo-holônimo em que conceitos complexos são subdivididos em pequenas unidades. Segundo Weller e Stock (2008), essa subdivisão figura uma estrutura hierárquica em que o conceito que representa a totalidade (holônimo) é visto como uma classe superior e os conceitos que denotam as partes desse todo (merônimos) são classes inferiores.

A relação merônimo-holônimo geralmente é expressa com o rótulo “*é parte de*”. Contudo, de acordo com Arnold e Rahm (2014), algumas vezes as preposições *em* e *de* podem indicar relações de merônimo-holônimo, como, por exemplo, em "A CPU é o hardware *d*o computador", que permite entender que “ A CPU *é parte do* computador”. Além disso, verbos como “consistir” podem rotular uma relação de merônimo-holônimo. Por exemplo, "Um computador *consiste de* pelo menos um elemento de processamento", ou seja: elemento de processamento *é parte do* computador.

A organização das relações de merônimo-holônimo na taxonomia das relações semânticas proposta nesta tese tem como base o agrupamento realizado por Peters e Weller (2008); Weller e Stock (2008) e Stock (2010) (ver Figura 24). Eles se fundamentam na abordagem de Gerstl e Pribbenow (1996), que consideram o aspecto construtivo do objeto. Para eles, "[c]ada relação parte-todo representa um modo diferente de particionar o todo em partes” (GERSTL; PRIBBENOW, 1996, p.

306, tradução nossa)[25]. Assim, eles agrupam as formas de particionamento de objetos em duas categorias: (1) Partições baseadas na estrutura composicional do todo e (2) partições dirigidas às características internas ou critérios externos.

Figura 24 – Relações de Merônimo-Holônimo

Fonte: Elaborada pela autora a partir de Peters e Weller (2008), Weller e Stock (2008) e Stock (2010).

No primeiro caso (partições baseadas na estrutura composicional do todo), observa-se que a estrutura interna pode ser naturalmente decomposta em partes de estruturas dependentes e permanentes (GERSTL; PRIBBENOW, 1996). Essas partições foram nomeadas por Peters e Weller (2008), Weller e Stock (2008) e Stock (2010) como parte de um objeto estruturado.

Por outro lado, as partições cujas partes dizem respeito a objetos não estruturados são dirigidas às características internas ou aos critérios externos. As partes resultantes são construções temporárias e podem não pertencer ao domínio do conhecimento (GERSTL & PRIBBENOW, 1996). Nesse caso, Peters e Weller (2008); Weller e Stock (2008) e Stock (2010) denominaram esse tipo de partição como parte de um objeto não estruturado.

Começando pelas relações de objeto estruturado, a relação componente-complexo – assim nomeada por Peters e Weller (2008), Weller e Stock (2008) e Stock (2010) –, conhecida como propriedade-objeto ou objeto funcional por Chaffin e Herrmann (1984), componente-objeto integral ou componente-objeto por Winston, Chaffin e Herrmann (1987), é uma relação cujo objeto pode ser segmentado em seus componentes. Exemplo: telhado *é parte da* casa (onde casa é o complexo e o telhado é um dos componentes de uma casa). Outros exemplos são: pedal *é parte da* bicicleta, motor *é parte do* carro, nariz *é parte do* rosto, asa *é parte do* avião (STOCK, 2010).

[25] "*Each part-whole relation represents a different way of partitioning a whole into parts.*" (GERSTL; PRIBBENOW, 1996, p. 306)

A relação unidade-organização compreende grupos sociais, que pode ser uma pessoa ou um conjunto de pessoas que pertençam a um grupo organizado. Por exemplo: Departamento *é parte da* Universidade, pesquisador *é parte do* departamento (WELLER; STOCK, 2008; STOCK, 2010).

Já na relação segmento-evento, um evento é composto e pode ser dividido em seções. Os eventos diferem de objetos (como na relação componente-complexo), uma vez que um evento pode ter partes que ocorrem em diferentes momentos no tempo, enquanto as partes de um objeto normalmente ocorrem ao mesmo tempo. É exemplo de relação segmento-evento: cena do trapézio *é parte do* espetáculo do circo (STOREY, 1993). Nesse contexto, o espetáculo do circo é o evento maior, composto de atos (segmentos), como as cenas do palhaço, da bailarina e do trapézio.

A relação entre subunidade geográfica e unidade geográfica**,** assim intitulada por Peters e Weller (2008), Weller e Stock (2008) e Stock (2010), também é conhecida como lugar-área por Winston, Chaffin e Herrmann (1987), e como localização por Chaffin e Herrmann (1984) e Broughton *et al.* (2005). Ela abarca relações onde espaços geográficos podem ter divisões, administrativas ou não. Exemplos dessa relação são: Minas Gerais *é parte do* Brasil, Hamburgo *é parte da* Alemanha; China *é parte da* Ásia.

Por outro lado, a localização funcional diz respeito a locais que podem ser desmembrados de um local maior. Exemplo: sala de jantar *é parte da* casa, geladeira *é parte da* cozinha, recepção *é parte da* biblioteca, caixa *é parte da* loja (CHAFFIN; HERRMANN, 1984).

Por fim, na relação membro-coleção ou elemento-coleção (não social), os membros em uma coleção diferem dos componentes, pois não é necessário que os membros executem uma função ou possuam um arranjo estrutural em particular em relação ao outro e sua totalidade (WINSTON; CHAFFIN; HERRMANN, 1987). Segundo Storey (1993), essa relação também pode ser expressa por "*pertence a*". Exemplo: o navio *pertence à* frota, a árvore *pertence à* floresta, a carta *pertence ao* baralho, o jurado *pertence ao* júri (WINSTON; CHAFFIN; HERRMANN, 1987; STOCK, 2010).

Das relações que dizem respeito aos objetos não estruturados ou à parte de objetos não estruturados, a relação porção-massa refere-se a uma totalidade

dividida em porções aleatórias (STOCK, 2010). De acordo com Winston; Chaffin e Herrmann (1987), nesse tipo de relação, as partes são similares umas com as outras e ao todo. São exemplos: Fatia-torta, grão-sal.

A relação fase-ação caracteriza-se pela divisão de uma atividade contínua em fases simples (STOCK, 2010). De acordo com Winston; Chaffin e Herrmann (1987), ela pode ser usada para se referir a estágios, fases, períodos discretos ou subatividades. Exemplos: pagamento-compra, namoro-adolescência, ovulação-ciclo menstrual.

Na relação matéria-prima-objeto, a matéria-prima não pode ser fisicamente separada do objeto. Exemplos: aço-carro, hidrogênio-água, alumínio-bicicleta (WINSTON; CHAFFIN; HERRMANN, 1987). Nesse tipo de relação, pode ser utilizada a expressão “é usado(a) para fazer” para rotular a relação. Como, por exemplo: alumínio *é usado para fazer* bicicleta.

Na relação entre subunidade de dimensão e unidade de dimensão, uma unidade homogênea pode ser dividida em subunidades. Por exemplo: um litro de vinho *é parte de um* barril de vinho (STOCK, 2010).

Por fim, a relação unidade-medida, definida por Chaffin; Herrmann (1984) denota relações de medidas decompostas em unidades, como, por exemplo, minuto-hora, centavo-dólar.

De acordo com Winston; Chaffin e Herrmann (1987), as relações de merônimo-holônimo diferem-se de três maneiras principais: funcionalidade, homogeneidade e separabilidade. No primeiro caso, as partes funcionais são restritas por sua função, localização espacial ou temporal. Por exemplo, a alça de uma xícara só pode ser colocada num número limitado de posições para poder funcionar como alça. Já a homogeneidade refere-se às partes homogêneas, que são do mesmo tipo de coisa que os seus conjuntos (por exemplo: fatia-torta), e às partes não-homogêneas, que são diferentes dos seus conjuntos (por exemplo: árvore-floresta). Por fim, na separabilidade, as partes separáveis podem, em princípio, serem separadas do conjunto (por exemplo: alça-copo), enquanto as partes inseparáveis não podem (por exemplo: aço-bicicleta).

Entende-se que algumas dessas características são visíveis nas relações semânticas, outras não. Winston; Chaffin e Herrmann (1987) determinam essas propriedades para algumas das relações apresentadas (Quadro 7). Devido ao

escopo do trabalho, decidiu-se não estabelecer as propriedades para as demais relações de merônimo-holônimo apresentadas pois, para isso, precisa-se de discussões que atestem com veemência a implicação das propriedades para as relações, o que não é o caso deste trabalho.

Quadro 7 – Relações de Merônimo-Holônimo e suas propriedades

Relação	Propriedades
Lugar - área	Não funcional, homogêneo e inseparável
Membro - coleção	Não funcional, homogêneo e separável
Porção - massa	Não funcional, homogêneo e separável
Fase - ação	Funcional, heterogêneo e inseparável
Matéria-prima - objeto	Não funcional, heterogêneo e inseparável
Componente - complexo	Funcional, heterogêneo e separável

Fonte: Winston, Chaffin e Herrmann (1987, p. 421, tradução nossa).

3.3.2.2 Relações de equivalência

As relações de equivalência dizem respeito às palavras cujos significados ou sentidos (conceitos por trás das palavras) são os mesmos ou são considerados equivalentes em um dado contexto. De acordo com o dicionário Aurélio[26], equivalência significa: “2 – Do mesmo valor. 3 – Que tem valor igual (a outro). 4 – Que pode substituir outro produzindo os mesmos efeitos ou tendo igual virtude, igual significado, etc.”. De acordo com Dodebei (2002, p. 91) as relações de equivalência permitem “controlar os três conjuntos de dispersões semânticas, característicos da língua natural; léxicas, simbólicas e sintáticas”.

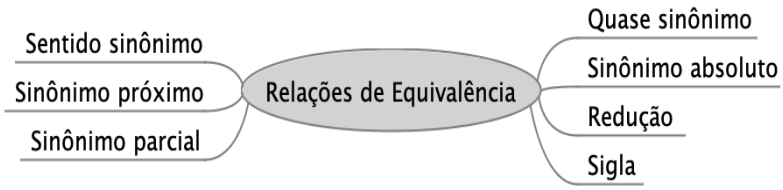
A respeito das classificações de relações semânticas apresentadas nesta pesquisa, pode-se dizer que somente a partir de Peters e Weller (2008) as relações de equivalência foram indicadas como uma relação semântica; porém, no contexto dos tesauros, a relação semântica foi indicada em 1960, por Vickery.

A Figura 25 apresenta as relações de equivalência proposta na taxonomia das relações semânticas. Na classificação proposta, as relações de equivalência remetem aos sinônimos, conforme determinado por Peters e Weller (2008) e Stock

[26] Disponível em: ‹https://dicionariodoaurelio.com/equivalente›. Acesso em: 11 mai. 2017.

(2010). Já a indicação dos tipos de sinônimos baseou-se em Khoo e Na (2006) e Chaffin e Herrmann (1984).

Figura 25 – Relações de equivalência



Fonte: Elaborada pela autora.

Os sinônimos, segundo Murphy (2003), incluem palavras-conceitos que têm contextualmente as mesmas propriedades relevantes, mas que diferem na forma do signo. O autor ressalta que tais propriedades relevantes devem incluir pelo menos o sentido da palavra.

Os sinônimos são denotados pela relação de sinonímia. A sinonímia é uma relação semântica diferente das demais relações semânticas porque ela é uma relação entre palavras e não entre conceitos (MURPHY, 2003), como as relações hierárquicas e associativas.

Na bibliografia pesquisada, foram encontrados alguns tipos de sinônimos. A base para a determinação dos tipos de sinônimos foram os autores Khoo e Na (2006), Stock (2010) e Chaffin e Herrmann (1984).

De acordo com Khoo e Na (2006) e Stock (2010), o tipo de relação "sentido sinônimo" diz respeito às palavras que compartilham um ou mais significados, como, por exemplo, carro e automóvel. Já o tipo "sinônimos próximos" não determina um sentido idêntico, mas significados que se aproximam, por exemplo, cidade e município. O tipo de relação "sinônimos parciais" indica que duas palavras compartilham muitos significados, mas diferem-se em alguns aspectos. Por exemplo[27], na frase: "Uma grande *depressão* assola o Brasil", *depressão* é sinônimo de crise. Já na frase “O Grand Canyon *é uma enorme* depressão”, *depressão* não é sinônimo de crise e, sim, termo relativo à faixa baixa de terreno. A relação "quase sinônimo" ocorre quando duas palavras possuem conceitos mais ou menos similares em termos de extensão e intensão, por exemplo, discussão e briga[28]. Chaffin e

[27] Baseado no exemplo de: http://www.magialectora.com/sinonimos-totales-y-parciales. Acesso em 25 mai. 2017.

[28] Exemplo extraído de Seligmann-Silva (2007, p. 11).

Herrmann (1984) chamam essa relação de similaridade dimensional, em que ocorre um significado denotativo que não é suficiente para sinonímia, mas que ocorre em pontos adjacentes em uma dimensão comum, ou seja, existe um grau de dimensão que indica em que dois conceitos diferem. Exemplos: riso e sorriso; bruto e desagradável, atormentado e irritado. Por fim, o tipo "sinônimo absoluto"; nesse caso, o significado é o mesmo na sua totalidade, por isso, esse tipo de sinônimo é raro.

Decidiu-se acrescentar à proposta de relações de equivalência as reduções e as siglas. Nas reduções, as palavras lexicalmente complexas são diminuídas ao seu primeiro elemento. Exemplo: foto (fotografia), micro (microcomputador), pós (pós-graduação) (AZEREDO, 2000). Por sua vez, a sigla é um "conjunto de letras iniciais dos vocábulos e/ou números que representa um determinado nome" (NBR ABNT 14724, 2011, p.4). Exemplos de siglas são: SP, RG, ONU e Unicamp. De acordo com Bueno (2017), "quando a sigla forma uma palavra pronunciável, como Unicamp, torna-se um acrônimo".

Segundo Arnold e Rahm (2014), algumas vezes os sinônimos aparecem entre parênteses, como na frase "Um ônibus (coletivo) é um veículo de rodas". Ainda de acordo com os autores, algumas vezes podem ser utilizadas as expressões: *é um sinônimo de*, *é abreviatura para*, *é o nome mais curto para*, ou, ainda, a palavra *significa.*

3.3.2.3 Relações associativas

De acordo com Hjørland (2007), a relação associativa é definida como uma associação mental de conceitos. Nesse contexto, se *A* é mentalmente associado a *B,* há uma relação associativa entre eles.

Em uma estrutura de classificação, a relação associativa se aplica para termos que são irmãos, ou seja, que estão no mesmo nível de hierarquia, ou para termos em que exista uma conexão clara, tanto conceitual quanto linguística, mas o relacionamento não atenda aos critérios necessários para uma relação hierárquica (BROUGHTON, 2008). De fato, como consta na ISO 25964-1 (2011), a relação

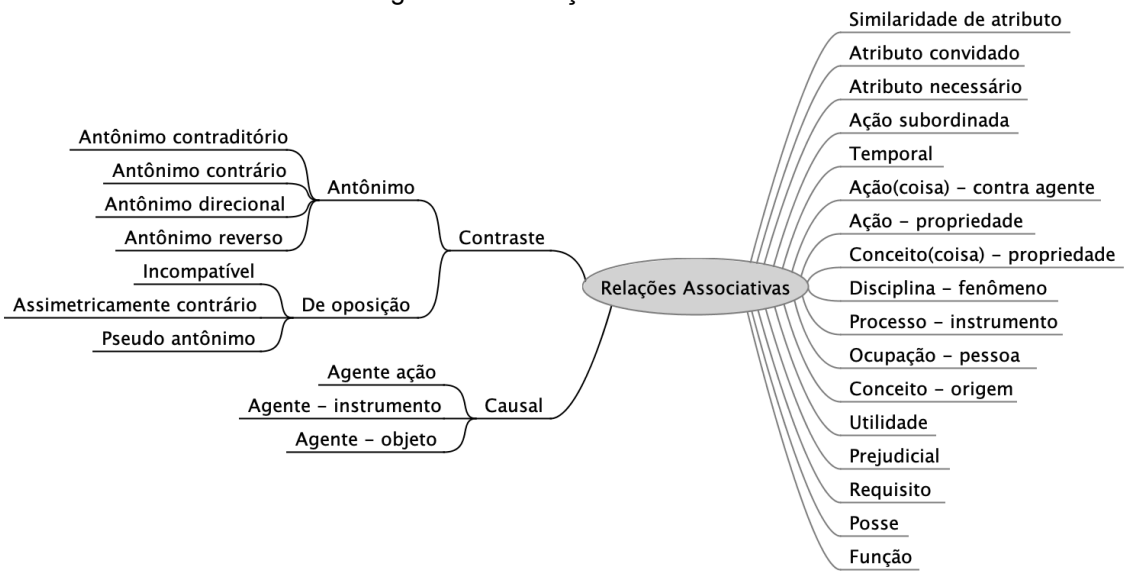
associativa ocorre entre um par de conceitos que não são relacionados hierarquicamente, mas compartilham uma forte conexão semântica.

Zeng (2008) expande essa afirmação e declara que uma relação associativa é aquela que não é nem hierárquica, nem de equivalência. De acordo com Green (2008), especificamente em tesauros, essas relações são usualmente especificadas com TR (Termo Relacionado).

Diferentemente dos outros tipos de relações, as relações associativas não têm um padrão tal como *é um* ou *é parte de* das relações hierárquicas. Nas relações associativas podem ocorrer quaisquer tipos de verbos (exceto os que denotam padrão para relações hierárquicas e de equivalência), desde que eles sejam transitivos, ou seja, aqueles que necessitam de um complemento para fazer sentido, tais como *causar, usar, realizar, pagar, comprar*, entre outros.

A Figura 26 mostra as relações associativas que são abordadas nesta seção. Nessa classe de relações estão as relações de contraste, que, segundo Chaffin e Herrmann (1984), são termos que se opõem, ou seja, que se contradizem. As relações de contraste subdividem-se em tipos de antônimos e de oposição.

Figura 26 – Relações associativas



Fonte: Elaborada pela autora.

De acordo com Hjørland (2007), especificamente em tesauros, os antônimos não são usualmente especificados, mas podem ser listados com outros termos como relação associativa. Da mesma forma, Zeng (2008) também propõe que os antônimos são um tipo de relação associativa.

Na concepção de Chaffin e Herrmann (1984), os antônimos podem ser: antônimos contraditórios**,** aqueles que são opostos dicotomicamente (vivo-morto); antônimos contrários, opostos simetricamente, por exemplo, quente e frio são temperaturas igualmente opostas; antônimos direcionais, que são opostos no tempo e espaço (antes-depois, em cima-embaixo); e, por fim, os antônimos reversos, que denotam ações opostas (vender-comprar).

As relações agrupadas como de oposição são aquelas que são contrárias, mas que não necessariamente são antônimas. Nesta tese, a relação de oposição foi classificada como associativa porque na definição de relação equivalente, de acordo com a ISO 25964-1 (2011, p. 4), "[esse] relacionamento [acontece] entre dois termos [...] onde ambos representam o mesmo conceito." Ao seguirmos essa definição, torna-se incoerente afirmar que a relação de oposição é uma relação de equivalência, um contraponto em relação a Tristão, Fachin e Alarcon (2004), que a consideram um quase sinônimo, posição que foi seguida por Maculan (2015) ao observar a falta de consenso sobre essa relação no que tange à sua classificação como equivalente ou associativa.

No grupo de relações de oposição da taxonomia proposta encontram-se as relações assimetricamente contrárias, ou seja, os conceitos são opostos em uma dimensão contínua, como por exemplo, os conceitos “seco” e “úmido”. Como se sabe, o oposto de seco é molhado e o estado úmido é um certo grau de molhado. Nesse caso, a simetria imperfeita é a razão pela qual o relacionamento não é geralmente considerado antônimo. As relações incompatíveis ocorrem quando um significado é oposto de parte de outro significado. Por exemplo, franqueza é incompatível com hipocrisia. A hipocrisia envolve desonestidade, enquanto a franqueza envolve honestidade e sinceridade. Há ainda os pseudoantônimos, que são assim chamados porque sua oposição é baseada em um significado conotativo de um termo; por exemplo, popular e tímido se opõem porque a popularidade conota a extroversão, que é denotativamente oposta à timidez (CHAFFIN, HERRMANN; 1984).

Outro grupo de relações associativas são as relações causais**.** Assim como em Broughton *et al.* (2005), optou-se por tratar as relações de caso de Chaffin e Herrmann (1984) e causa-efeito de Khoo e Na (2006) e Zeng (2008) como relações causais. Nesse tipo de relação, *A* é causa para *B* (BROUGHTON *et al.,* 2005).

Storey (1993) agrupa as relações causais em dois tipos: agentes e ações. No caso dessa proposta de taxonomia, optou-se por incluir apenas as relações que referem-se a agentes, isso porque as relações que envolvem ações dentro desse contexto são relações unárias e, no âmbito dessa taxonomia, somente as relações binárias estão sendo abordadas. Nesse sentido, as relações causais foram classificadas como agente-ação: ocorrem entre um agente e a ação que geralmente executa (por exemplo: consultor-consultoria, programador-programação). O contrário dessa relação, que não está classificada como relação causal, é a relação ocupação-pessoa**,** proposta por Peters e Weller (2008). Como exemplo, os autores sugerem a relação contabilidade-contador. As relações agente-instrumento ocorrem entre um agente e o instrumento que ele usa (por exemplo: programador-computador). E as agente-objeto ocorrem entre o agente e o objeto que ele usa ou faz (por exemplo: carpinteiro-madeira).

A relação similaridade de atributo, conforme definida por Chaffin e Herrmann (1984), denota termos em que os atributos proeminentes de um conceito se assemelham aos de outro. São exemplos: garfo-rastelo, quadro-filme, tapete-cobertor. Percebe-se que, nesses casos, existe uma equivalência no que diz respeito à função ou a alguma característica dos objetos. Chaffin e Herrmann classificaram essa relação no grupo das relações que denotam similaridades, tais como os sinônimos. Contudo, decidiu-se classificar essa relação como associativa por se entender que os conceitos têm “coisas em comum” mas eles não compartilham significados. O mesmo ocorreu para as relações atributo necessário e atributo convidado. Nesse último caso, Storey (1993) classificou como uma relação causal. Entende-se que a relação de atributo convidado envolve características que não necessariamente ocorrem na definição do termo, por exemplo: cama-confortável, professor-inteligente. Já a relação atributo necessário (ou atribuição necessária) envolve um termo e um atributo de definição do termo, como, por exemplo, limão-azedo (CHAFFIN; HERRMANN, 1984).

Chaffin e Herrmann (1984) classificam a relação ação subordinada como inclusão de classe; contudo, entende-se, pelos exemplos, que essa relação não diz respeito a uma unidade e uma classe, como na relação atividade subordinada. Em contrapartida, eles descrevem as duas relações como parecidas, a diferença está em que uma diz respeito à ação e outra à atividade. "Atividade subordinada e ação

subordinada envolvem atividades (ex.: xadrez-jogo) e ação (fritar-cozinhar), respectivamente" (CHAFFIN; HERRMANN, 1984, p. 136, tradução nossa)[29]. A explicitação da relação do exemplo fritar-cozinhar pode ser fritar *é parte de* cozinhar; contudo, devido ao fato de essa relação não se referir a um objeto, ela não foi classificada como merônimo-holônimo.

A relação temporal é uma relação semântica à qual um conceito indica um tempo ou período de um evento em relação ao outro conceito. Por exemplo: Segunda Guerra Mundial *ocorreu entre* (1939-1945) (BROUGHTON *et a*l., 2005, p. 143).

Conforme apontado por Peters e Weller (2008) e Stock (2010), as relações associativas podem ser infinitas. O Quadro 8 mostra exemplos de outras relações apontadas na literatura pesquisada. Como pode ser percebido nesse quadro, existem relações que envolvem ações e agentes; contudo, baseado em seus exemplos, elas não denotam causas e, por isso, não foram classificadas como relações causais. Além disso, constam no Quadro relações que denotam atributos (propriedades), porém elas não dizem respeito às relações de atributo necessário ou convidado.

Quadro 8 – Exemplos de outras relações associativas

Relação associativa	Exemplo
Ação(coisa) - contra agente	Peste - pesticida
Ação - propriedade	Medição de precisão - precisão
Conceito(coisa) - propriedade	Liga de aço - resistência à corrosão
Disciplina - fenômeno	Sismologia - terremoto
Processo - instrumento ou processo-agente	Medição de velocidade - velocímetro
Conceito - origem	Água - poço de água
Utilidade	Criação de trabalho - desenvolvimento da economia
Prejudicial	Fertilizante - diversidade de espécie
Requisito	Bateria - MP3 *Playe*r
Posse	MP3 *Player* - Sony Ericsson
Função	*Bluetooth* - computador

Fonte: Elaborado pela autora a partir de Peters e Weller (2008) e Zeng (2008).

[29] "*Activity subordinates and action subordinates involve activities (e.g., "game-chess") and actions ("cook-fry"), respectively.*" (CHAFFIN; HERRMANN, 1984, p. 136).

Algumas relações semânticas associativas não foram classificadas porque nas fontes onde elas se encontravam e na bibliografia pesquisada não havia especificações ou exemplos suficientes para o correto entendimento e a classificação delas, a saber, produção, uso e anexo, propostas por Peters e Weller (2008). Por sua vez, a relação matéria-prima-produto (*raw material-product*), proposta por Peters e Weller (2008) e Zeng (2008), foi classificada como uma relação associativa; contudo, pelos exemplos das autoras (pele-couro, uva-vinho), entende-se que essa é uma relação partitiva do tipo matéria-prima-objeto (*stuff-objeto*) proposta por Winston, Chaffin e Herrmann (1987), explicada anteriormente.

A relação "ver também", proposta por Stock (2010) como um tipo de relação associativa, foi entendida neste contexto como uma relação que engloba todas as relações associativas, por isso ela não foi classificada. O mesmo se aplica para a relação ramo-relacionado, proposta por Peters e Weller (2008).

A relação "dependência causal", proposta por Peters e Weller (2008), engloba as relações causais já mencionadas. Outras relações não classificadas foram as relações ativo-passivo, propostas por Hjørland (2007) e a relação ativa, que denota relação semântica entre dois conceitos, em que um deles expressa a execução de uma operação ou processo em outro. O inverso é a relação passiva. Como essas relações não foram exemplificadas pelos autores, o que permitiria um completo entendimento, decidiu-se não classificá-las na taxonomia proposta.

Os homônimos não foram classificados porque, segundo Stock (2010), não se trata de uma relação semântica, apesar de suas implicações para a Recuperação da Informação. Os homônimos são conceitos com a mesma grafia, ou seja dois conceitos, A e B, são expressos com o mesmo símbolo. Segundo Hjørland (2007), um exemplo de homônimo é a palavra “banco” que pode significar um banco de sentar ou uma instituição financeira. De acordo com Stock (2010), os homônimos cujas palavras têm o mesmo som são chamados homófonos (por exemplo, as palavras escritas em inglês *see* e *sea*); já os homógrafos são denotados por palavras iguais, porém com significados diferentes, como é o caso da palavra “banco”, exemplificada acima.

A proposta de taxonomia de relações semânticas para a Biblioteconomia e Ciência da Informação completa é enfim apresentada na Figura 27. No Quadro 9, ela é compilada com todos os exemplos.

Figura 27 – Proposta de taxonomia de relações semânticas para a Biblioteconomia e Ciência da Informação

Fonte: Elaborada pela autora.

Quadro 9 - Taxonomia das relações semânticas com exemplos

Taxonomia das relações semânticas
— Hierárquicas
--- Hipônimo-Hiperônimo
----Hipônimo Simples (A rainha *é uma* mulher)
---- Instância (João *é uma* pessoa)
---- Inclusão de Classe
----- Perceptivelmente Subordinado (cavalo *é um* animal)
----- Funcionalmente Subordinado (carro *é um tipo de* veículo)
----- Estado Subordinado (medo *é uma* emoção)
----- Atividade Subordinada (xadrez *é um tipo de* jogo)
----- Geograficamente Subordinado (América *é um* país)
-----Taxonômica (Macarrão *é um tipo de* massa)
---- Inclusão Espacial (O cliente *está na* recepção)
--- Merônimo-Holônimo
---- Objeto Estruturado
----- Componente-Complexo (Telhado *é parte da* casa)
----- Unidade-Organização (Departamento *é parte da* Universidade)
----- Segmento-Evento (O trapézio *é parte do* espetáculo do circo)
----- Subunidade Geográfica-Unidade Geográfica (Minas Gerais *é parte do* Brasil)
----- Localização Funcional (A sala de jantar *é parte da* casa)
----- Membro-Coleção (O navio *pertence à* frota)
---- Objeto Não-Estruturado
----- Porção-Massa (Fatia *é uma parte da* torta)
----- Fase-Ação (Pagamento *faz parte da* compra)
----- Matéria prima-Objeto (O aço *é usado para fazer* carro)
----- Subunidade de Dimensão-Unidade de Dimensão (Litro de vinho *é parte de um* barril de vinho)
----- Unidade-Medida (Minuto *é parte de* hora)
-- Equivalência
--- Sentido Sinônimo (O carro *é um* automóvel)
--- Sinônimos Próximos (Cidade *é um sinônimo próximo de* município)
--- Sinônimos Parciais (Depressão *é sinônimo de* crise)
--- Quase Sinônimos (Discussão *é quase uma* briga)
--- Sinônimo Absoluto
--- Reduções (Foto *é uma redução de* fotografia)
--- Sigla (SP *é a sigla de* São Paulo)

Taxonomia das relações semânticas
-- Associativas
--- Contraste
---- Antônimo
----- Antônimos Contraditórios (Vivo *é o contrário de* morto)
----- Antônimos Contrários (Quente *é o contrário de* frio)
----- Antônimos Direcionais (Antes *é o oposto de* depois)
----- Antônimos Reversos (Vender *é o oposto de* comprar)
---- De Oposição
----- Assimetricamente Contrárias (Seco *é assimetricamente contrário de* úmido)
----- Incompatíveis (franqueza *é incompatível com* hipocrisia)
----- Pseudo-Antônimos (Popular *é o oposto de* tímido)
--- Relações Causais
---- Agente-Ação (Consultor *realiza* consultoria)
---- Agente-Instrumento (Programador *usa* computador)
---- Agente-Objeto (Carpinteiro *usa* madeira)
--- Ocupação-Pessoa (Contabilidade *é a profissão do* contador)
--- Similaridade de Atributo (O garfo *tem algumas características do* rastelo)
--- Atributo Convidado (A cama *é* confortável)
--- Atributo Necessário (O limão *é* azedo)
--- Ação Subordinada (Fritar *é parte de* cozinhar)
--- Temporal (A Segunda Guerra Mundial *aconteceu entre os anos* 1939 a 1945)
--- Ação(Coisa)-Contra agente (A peste *é eliminada com o* pesticida)
--- Ação-Propriedade (A medição de precisão *determina a* precisão)
--- Conceito(Coisa)-Propriedade (A liga de aço *tem* resistência à corrosão)
--- Disciplina-Fenômeno (A sismologia *estuda os* terremotos)
--- Processo-instrumento (A medição de velocidade *é feita pelo* velocímetro)
--- Conceito-origem (A água *vem do* poço de água)
--- Utilidade (Criação de trabalho *gera* desenvolvimento da economia)
--- Prejudicial (O fertilizante *afeta a* diversidade de espécies)
--- Requisito (A bateria *é um requisito para o* MP3 *Player*)
--- Posse (O MP3 *player é da* Sony Ericsson)
--- Função (O bluetooth *é uma funcionalidade do* computador).
— Agente Subordinado (autores *são intermediados por* indexadores)

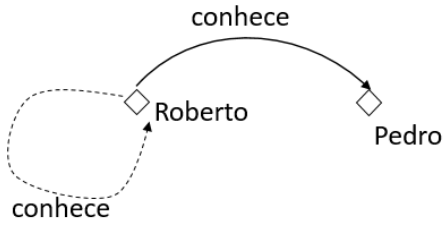
Fonte: Elaborado pela autora.

3.3.3 Propriedades das relações semânticas

O estabelecimento das relações semânticas requer indicar as propriedades de reflexividade, simetria e transitividade, que facilitam a realização de inferências por pessoas e computadores, o que permite inteligibilidade na representação do conhecimento.

A reflexividade diz respeito ao comportamento do conceito em consideração a si mesmo, de acordo com a relação semântica. Um exemplo de relação reflexiva é mostrado na Figura 28. Nesse exemplo, Roberto *conhece* Pedro e Roberto *conhece* si mesmo.

Figura 28 – Exemplo de relação reflexiva

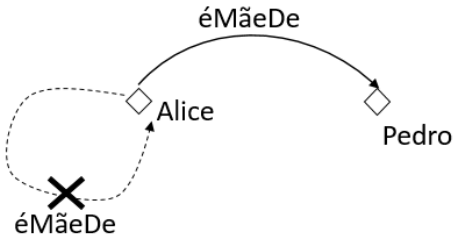


Fonte: Traduzida e adaptada de <https://pt.slideshare.net/adrianomelo/protg-lgica-de-descries-na-pratica> Acesso em 02 de jul. 2017.

A relação *conhece,* apresentada no exemplo da Figura 28, é associativa; contudo, não se deve dizer que todas as relações associativas são reflexivas, pois elas devem ser analisadas caso a caso. Os sinônimos e as relações de instância são tipos de relações em que pode-se afirmar que são reflexivas, segundo Stock (2010).

Já na relação irreflexiva, o conceito não se relaciona consigo mesmo. Isso é exemplificado na Figura 29, em que é possível observar que Alice *é mãe de* Pedro, mas Alice não pode ser mãe dela mesma. Com exceção dos sinônimos e das relações de instâncias, todas as outras relações de hierarquia e equivalência são irreflexivas, de acordo com Stock (2010). No caso das relações associativas, como não há um padrão de verbos, elas devem ser analisadas em particular.

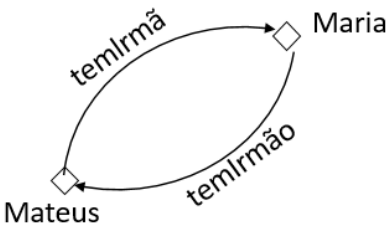
Figura 29 – Exemplo de relação irreflexiva



Fonte: Traduzida e adaptada de <https://pt.slideshare.net/adrianomelo/protg-lgica-de-descries-na-pratica> Acesso em 02 de jul. 2017.

A propriedade de simetria ocorre quando uma relação entre A e B também existe na direção oposta, ou seja, entre B e A. Como pode ser visto no exemplo apresentado na Figura 30, a relação *tem irmã(o)* é simétrica, pois se Mateus é irmão de Maria, Maria é irmã de Mateus. Outros exemplos, de acordo com Green (2001), são: *é primo(a) de* e *é esposo(a) de.* Segundo Stock (2010), os antônimos e sinônimos também são simétricos.

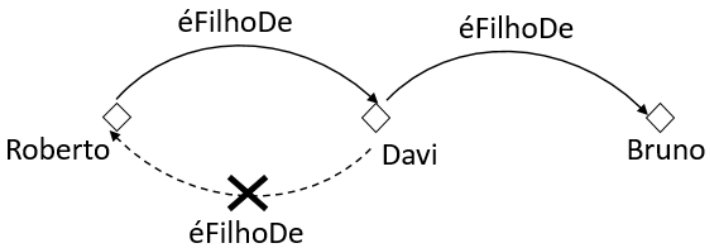
Figura 30 - Exemplo de relação simétrica



Fonte: Traduzida e adaptada de <http://mba.eci.ufmg.br/onto_owl/#_Toc205529350>. Acesso em 02 de jul. 2017.

O oposto da relação simétrica é a assimétrica. Nesse caso, A se relaciona com B, mas B não se relaciona com A. De acordo com Green (2001), muitos relacionamentos são assimétricos. Segundo a autora, não é usual que uma entidade A suporte um certo relacionamento com uma entidade B e a entidade B suporte o mesmo relacionamento com A. Por exemplo: se Maria gosta de João, não necessariamente João gosta de Maria. Do mesmo modo, como pode ser visto na Figura 31, Roberto *é filho de* Davi, mas Davi não pode ser filho de Roberto.

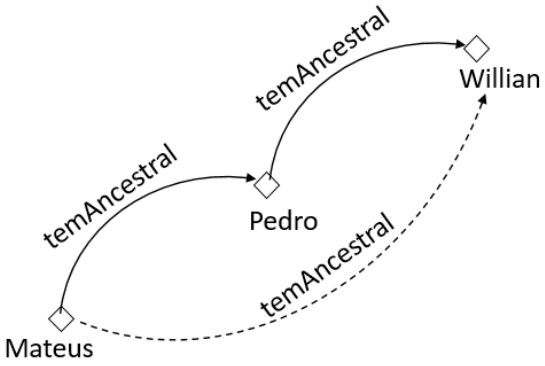
Figura 31 – Exemplo de relação assimétrica



Fonte: Traduzida e adaptada de <https://pt.slideshare.net/adrianomelo/protg-lgica-de-descries-na-pratica>. Acesso em 02 de jul. 2017.

Por fim, a transitividade. Segundo Stock (2010), quando existe uma relação entre dois conceitos A e B, entre B e C, e ainda entre A e C, trata-se da transitividade. A Figura 32 mostra um exemplo de relacionamento transitivo. Nesse exemplo, Mateus *tem ancestral* Pedro e Pedro *tem ancestral* Willian, ou seja, se Mateus é ancestral de Pedro e Pedro é ancestral de Willian, logo, por inferência, Mateus é ancestral de Willian. De acordo Stock (2010), os sinônimos e as relações de hipônimo simples e taxonômicas são transitivas.

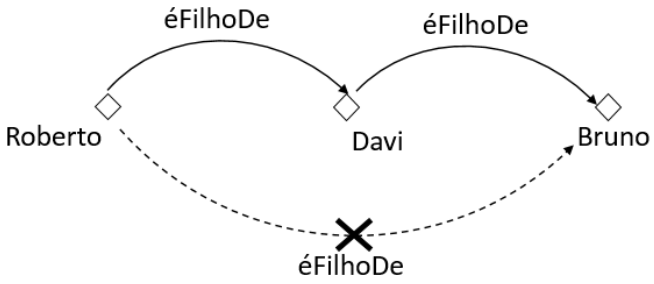
Figura 32 – Exemplo de relações transitivas



Fonte: Traduzida e adaptada de <http://mba.eci.ufmg.br/onto_owl/#_Toc205529350>. Acesso em 02 de jul. 2017.

O inverso da relação transitiva é a intransitiva. Como pode ser visto na Figura 33, Roberto *é filho de* Davi e Davi *é filho de* Bruno, contudo não se pode dizer que Roberto *é filho de* Bruno, pois, ao interpretar essas proposições, sabe-se que Roberto *é neto* de Bruno. Segundo Stock (2010), os antônimos e as relações de instância são intransitivos. De acordo com Khoo e Na (2006), a relação merônimo-holônimo é intransitiva. Ele exemplifica com a seguinte questão: a maçaneta *é parte da* porta e a porta *é parte da* casa; pode-se dizer que a maçaneta *é parte da* casa?

Figura 33 – Exemplo de relação intransitiva

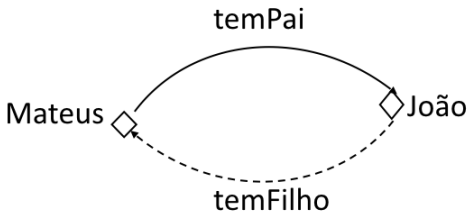


Fonte: Elaborada pela autora.

Entende-se que a transitividade é uma propriedade aplicada a relações ternárias, o que não não se aplica a esta pesquisa, que tem o propósito de explorar as relações binárias, ou seja, entre dois conceitos[30].

Decidiu-se tratar nesta seção as relações inversas. O inverso de uma relação semântica não é necessariamente uma propriedade, como as apresentadas até o momento; contudo, ao determinar as relações inversas, consegue-se tornar a representação semântica rastreável. Essa rastreabilidade é importante sobretudo para conferir se a relação semântica procede de fato. No exemplo da Figura 34, na relação Mateus *tem pai* João, há a relação inversa: João *tem filho* Mateus.

Figura 34 – Exemplo de relação inversa



Fonte: Traduzida e adaptada de < http://mba.eci.ufmg.br/onto_owl/>. Acesso em 14 de ago. 2017.

[30] Contudo, um estudo mais detalhado deverá ser feito no futuro para determinar a transitividade entre dois pares de relações semânticas consecutivos em alguns contextos, quando um desses conceitos faz parte das duas relações semânticas, interligando-as.

3.4 Extração de relações

A temática que envolve a Extração de Relações (ER) teve início a partir da MUC (*Message Understanding Conference* – Conferência de Entendimento de Mensagem). Com o propósito de promover e avaliar pesquisas em Extração de Informação (EI), a MUC foi uma iniciativa do NOSC (*Naval Ocean System Center* – Centro de Sistema Oceânico Naval), uma organização americana militar. Essa Associação tinha como propósito avaliar e fomentar pesquisas em análise automatizada de mensagens textuais militares (GRISHMAN; SUNDHEIM, 1996). As edições da MUC iniciaram-se em 1987, contudo, somente em 1997, na sétima e última edição da conferência, o problema da ER foi abordado pelos participantes (ZELENKO, AONE & RICHARDELLA, 2003).

A ER é parte da EI e uma tarefa estabelecida em Processamento de Linguagem Natural (PLN) (KONSTANTINOVA, 2014). A linguagem natural é o termo usado para distinguir linguagens humanas de linguagens formais ou computacionais. Portanto, o PLN é o estudo científico de linguagens naturais na perspectiva computacional. Logo, ele envolve a Ciência da Computação (CC) e a Linguística (KUMAR, 2011).

Com início nos anos 1940, as pesquisas que envolvem o PLN trabalham com dois tipos de sistemas: os que convertem informações de bases de dados computacionais em uma linguagem legível para humanos, que são os sistemas de geração de linguagem natural, e os sistemas de entendimento de linguagem natural, que convertem exemplos da linguagem humana em representações formais para manipulação de programas de computador (KUMAR, 2011). Nesse último caso, uma das tecnologias de PLN é a EI.

A EI busca extrair informação relevante a partir textos em linguagem natural. Sistemas que utilizam EI são capazes de incorporar informações relacionadas, distribuídas em diferentes fontes, e apresentá-las de maneira intuitiva para os usuários (RODRIGUES & TEIXEIRA, 2015).

Geralmente um processo de EI tem como entrada textos e, algumas vezes, falas. A partir disso, ele produz dados não ambíguos e gera saídas que podem ser mostradas para os usuários ou podem ser armazenadas em bases de dados ou

planilhas para análise posterior. Ainda, os dados resultantes de um processo de EI podem ser usados para indexação em aplicações de Recuperação da Informação (RI) (CUNNINGHAM, 2005).

A EI extrai automaticamente informações estruturadas, tais como entidades, atributos que descrevem entidades e relacionamentos entre entidades, a partir de fontes não estruturadas. As entidades são tipicamente sintagmas nominais, que denotam nomes de pessoas, locais, organizações, etc. (SARAWAGI, 2008). Nesse caso, no que se refere à identificação de entidades em textos, a EI conta com a tarefa de Reconhecimento de Entidade Nomeada (REN).

O REN foi abordado pela primeira vez na MUC-6. Essa tarefa da EI envolve identificar os nomes de pessoas, organizações, tempo, moeda e expressões de porcentagem em textos por meio de identificadores (GRISHMAN & SUNDHEIM, 1996). A Figura 35 apresenta um exemplo de REN em que primeiro é mostrado um trecho de um texto extraído do Wikipédia[31] e em seguida as entidades que foram anotadas, com seus respectivos identificadores, nesse mesmo texto. Nesse caso, foram reconhecidas as seguintes entidades: pessoa, local, data, profissão e organização.

Figura 35 – Exemplo de reconhecimento de entidade em um texto e um exemplo de relações extraídas

Timothy John Berners-Lee KBE, OM, FRS (TimBL ou TBL) (Londres, 8 de junho de 1955) é um físico britânico, cientista da computação e professor do MIT.
Texto de uma fonte não estruturada.
<PESSOA> Timothy John Berners-Lee **</PESSOA>** KBE, OM, FRS (TimBL ou TBL) (**<LOCAL>**Londres **</LOCAL>**, **<DATA>** 8 de junho de 1955 **</DATA>**) é um **<PROFISSÃO>** físico **</PROFISSÃO>** britânico , **<PROFISSÃO>** cientista da computação **</PROFISSÃO>** e **<PROFISSÃO>** professor **</PROFISSÃO>** do **<ORGANIZAÇÃO>** MIT **</ORGANIZAÇÃO>.**
Exemplos de Entidades Anotadas.

Fonte: Elaborada pela autora com base em Sarawagi (2008, p. 270).

[31] Disponível em <https://pt.wikipedia.org/wiki/Tim_Berners-Lee>. Acesso em 02 jul. 2017.

Uma vez que as entidades são reconhecidas no REN, é importante estabelecer a relação entre elas. Para as entidades anotadas na Figura 35 podem existir, por exemplo, relações entre as entidades PESSOA-LOCAL (Timothy John Bernes-Lee <NASCEU_EM> Londres) e PESSOA-ORGANIZAÇÃO (Timothy John Bernes-Lee <TRABALHA_EM> MIT).

De acordo com Sarawagi (2008), o problema de extração de relações entre dois conceitos pode ser colocado de três formas diferentes: (1) as entidades são identificadas no texto não-estruturado e então, para cada par de entidades, precisa-se descobrir o tipo de relação que existe entre elas; (2) existe um tipo de relação *r* e um nome de entidade *e*, e o objetivo é extrair as entidades com as quais *e* tem relação *r*; (3) dado um tipo de relacionamento fixo *r*, o objetivo é extrair todas as instâncias de pares de entidades *e* que têm relação *r* em um *corpus* não estruturado, onde não é possível assumir que os pares de entidades estão marcados.

Com o foco no primeiro caso, em que as entidades são identificadas no texto não-estruturado, para cada par de entidades precisa-se descobrir o tipo de relação que existe entre elas. Para isso, Sarawagi (2008) enumera alguns recursos úteis (tanto quando se trata de entidades quanto de conceitos) para Extração de Relações (ER). São eles: *Surface Tokens* (*Tokens*[32] *de Superfície*), *Part-Of-Speech tags* – *POS tags* (Marcação de Classe Gramatical), *Syntactic Parse Tree Structure* (Estrutura de Árvore de Análise Sintática) e, *Dependency Graph* (Grafo de Dependência).

Os *Tokens* de Superfície são *tokens* ao redor e entre duas entidades que podem sugerir relacionamentos entre elas (SARAWAGI, 2008). Por exemplo, dada a sentença: “UFMG está localizada em Belo Horizonte.” Ao detectar que UFMG é uma instância para a entidade ORGANIZAÇÃO e Belo Horizonte é uma instância da entidade LOCAL, entre elas existe o *token* de trigrama “está localizada em” que é uma sugestão da existência de um relacionamento entre as entidades ORGANIZAÇÃO e LOCAL, dado que, nesse trigrama existe a presença do radical “localiz” para o verbo localizar, que é um dos verbos recomendados para estas entidades.

As POS *tags* são marcações que denotam a classe gramatical das palavras em uma sentença. Portanto, elas indicam se determinada palavra é um verbo,

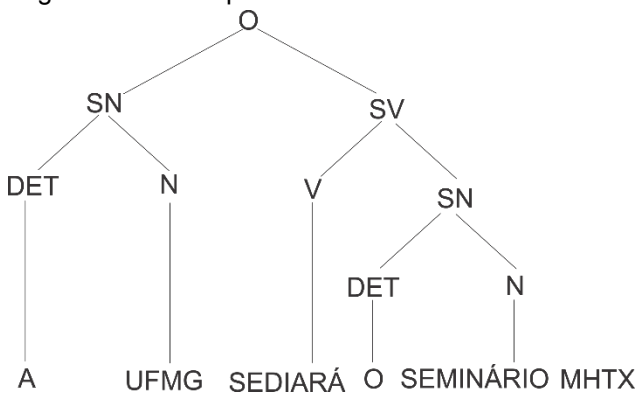
---

[32] Os *tokens* são aqui delimitados como uma ou mais palavras (símbolos) que definem objetos e abstrações.

pronome, adjetivo, advérbio, etc. Nesse caso, quando existe a indicação de um verbo, há um forte indício de uma relação, ou seja a presença de um verbo em uma frase é fundamental para definir uma relação entre entidades (SARAWAGI, 2008). Por exemplo, na sentença: “A UFMG sediará o Seminário MHTX.” Nesse caso, utilizando POS *tags*, a sentença seria marcada da seguinte forma: “A/AD UFMG/SP sediará/VB o/AD Seminário MHTX/SC.” Onde AD é a sigla para Artigo Definido; SP – Substantivo Próprio; VB – Verbo e; SC – Substantivo Composto. Como a palavra “sediará” está marcada como verbo e está entre dois substantivos (UFMG e Seminário MHTX), pode-se afirmar que sediará é a relação entre UFMG e Seminário MHTX.

Enquanto as POS *tags* determinam a classe gramatical das palavras (análise morfológica), a *Syntactic Parse Tree Structure* determina a função sintática das palavras de acordo com seus constituintes, ou seja, de acordo com seu papel dentro da oração, de um ponto de vista estrutural, isto é, de sujeito, objeto, predicado verbal, predicado nominal, adjunto adverbial, etc. A Figura 36 representa uma árvore de análise sintática da oração “A UFMG sediará o Seminário MHTX”, em que O indica oração, SN – sintagma nominal, SV – sintagma verbal, DET – determinante, N – nome e V – verbo. Da mesma forma como na utilização das POS *tags* o verbo indica o relacionamento.

Figura 36 – Exemplo de árvore de análise sintática

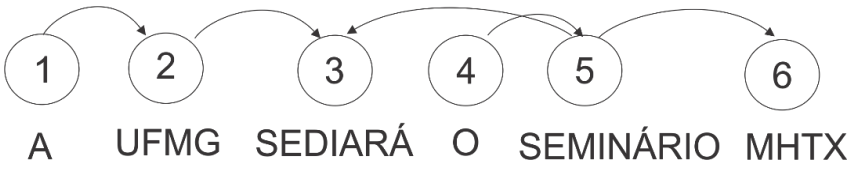


Fonte: Elaborada pela autora.

Por fim, apresenta-se o *Dependecy Graph*, que, *d*e acordo com Sarawagi (2008), é tipo de grafo direcionado que representa as dependências dos objetos uns com os outros. Nesse sentido, um grafo de dependência liga cada palavra às

palavras que dependem dela. Ele é frequentemente utilizado por ser tão adequado quanto uma árvore de análise sintática. No exemplo da Figura 37, cada nó representa uma palavra e as arestas representam a dependência entre as palavras. A palavra “sediará” é ligada e depende de ambos, do nó que representa UFMG e Seminário, o que pode indicar a relação entre UFMG e Seminário.

Figura 37 – Exemplo de grafo de dependência



Fonte: Elaborada pela autora.

Todos esses recursos apresentados podem auxiliar na indicação de relações semânticas, pois, de acordo com Sarawagi (2008), eles fornecem “pistas” da existência de um relacionamento em um fragmento de texto.

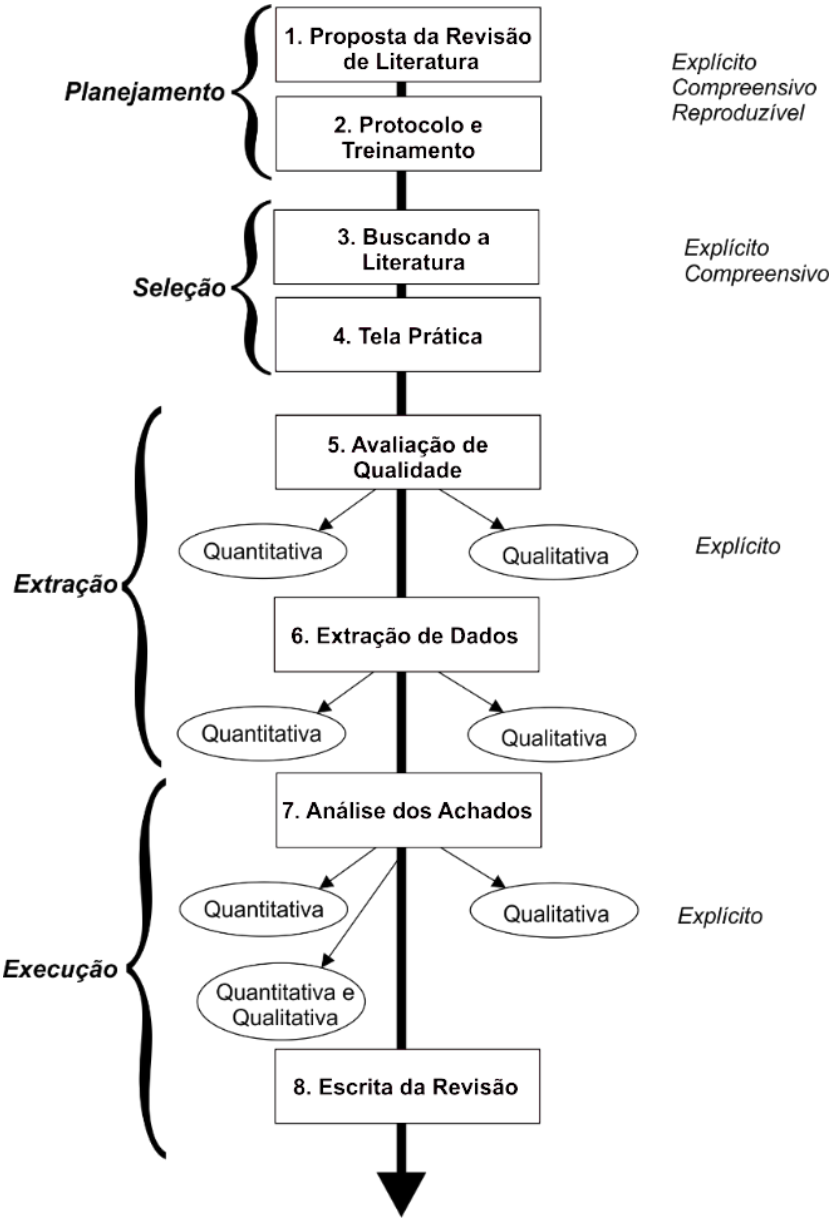
Esse capítulo apresentou a fundamentação teórico-metodológica no qual abordou os principais assuntos que tratam essa tese. O próximo capítulo dispõe sobre os trabalhos correlatos sobre extração de relações semânticas nos últimos cinco anos.

4 TRABALHOS CORRELATOS

Neste capítulo descreve-se os procedimentos realizados para a revisão de literatura que buscou identificar e relatar os estudos sobre extração de relações como maneira de explicitar relações semânticas, desenvolvidos nos últimos cinco anos (de 2013 a 2017, inclusive).

A revisão de literatura proposta neste capítulo utilizou a metodologia de Okoli e Schabram (2010), que sugerem oito etapas para a realização da pesquisa, conforme mostra a Figura 38.

Figura 38 – Etapas para a realização de uma revisão de literatura sistemática



Fonte: Okoli & Schabram (2010, p. 9, tradução nossa).

Na etapa 1, intitulada *Proposta da Revisão de Literatura*, busca-se identificar a intenção de realização dessa revisão. Okoli e Schabram (2010) enumeram seis razões que podem motivar uma revisão de literatura, são elas: (a) analisar o progresso de um fluxo específico de pesquisa; (b) fazer recomendações para pesquisas futuras; (c) revisar a aplicação de um modelo teórico da literatura; (d) revisar a aplicação de uma abordagem metodológica na literatura; (e) desenvolver um modelo ou *framework*; e (f) responder uma questão de pesquisa específica. Essa última pode ser considerada a razão desta revisão de literatura, que buscou, por meio de diversas publicações, responder a uma questão central: *Quais pesquisas foram realizadas nos últimos cinco anos sobre a extração de relações semânticas?*

A etapa 2, intitulada *Protocolo e Treinamento*, consiste em planejar a condução da revisão de literatura, o que equivale a especificar os passos e procedimentos para guiar a revisão (OKOLI; SCHABRAM, 2010) e que inclui definir as expressões de busca, o intervalo de tempo das publicações e a seleção das fontes de pesquisa. As fontes de pesquisa, segundo Fink (2010), podem ser: (a) bases de dados bibliográficos ou de artigos públicos *online*; (b) bases de dados bibliográficos privadas; (c) bases de dados especializadas; (d) busca manual de referências em artigos; e (e) orientações de especialistas. Nesta tese, optou-se por utilizar as bases de dados bibliográficos ou de artigos públicos *online* como principal fonte de dados. Para a escolha dessas bases, considerou-se as 29 bases da área e subárea Ciências Sociais – Ciência da Informação, do Portal de Periódicos Capes. Dessas 29 bases, cinco foram selecionadas de acordo com a relevância delas para a área, são elas: *Library, Information Science & Technology Abstracts with Full Text* (EBSCO), *Information Science & Technology Abstracts* – ISTA (EBSCO), *Library and Information Science Abstracts* – LISA (ProQuest), *Web Of Science* e *Scopus*.

Ainda nessa segunda etapa, prosseguindo com o planejamento, as expressões de busca com operadores *booleanos AND* e *OR* foram elaboradas a partir da questão de pesquisa (ver Quadro 10). Essas expressões também incorporaram os assuntos *typed link* e *links* automáticos, pois julgou-se que eles eram assuntos próximos à questão levantada. Todas as expressões foram testadas a fim de validar sua coerência para a pesquisa.

Quadro 10 – Expressões de busca utilizadas na revisão de literatura

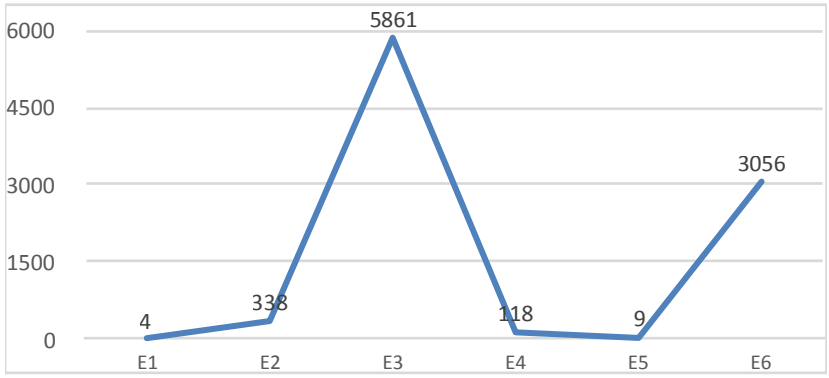
Sigla	Expressão de busca
E1	(Typed AND (Link OR Node) AND Automatic link) AND ((Semantic Relation OR Semantic Relationship) OR (Explanation OR Eliciting OR Explicitness))
E2	(Typed AND (Link OR Node) AND Automatic link) OR ((Semantic Relation OR Semantic Relationship) AND (Explanation OR Eliciting OR Explicitness))
E3	(Typed AND (Link OR Node) AND Automatic link) OR ((Semantic Relation OR Semantic Relationship) OR (Explanation OR Eliciting OR Explicitness))
E4	(Typed AND (Link OR Node) OR Automatic link) AND ((Semantic Relation OR Semantic Relationship) OR (Explanation OR Eliciting OR Explicitness)
E5	(Typed AND (Link OR Node) OR Automatic link) AND ((Semantic Relation OR Semantic Relationship) AND (Explanation OR Eliciting OR Explicitness))
E6	(Typed AND (Link OR Node) OR Automatic link) OR ((Semantic Relation OR Semantic Relationship) AND (Explanation OR Eliciting OR Explicitness))

Fonte: Elaborado pela autora.

Após a definição das expressões de busca, delimitou-se o intervalo de tempo pretendido das publicações. De acordo com Morehouse (s. d.), na área de Ciências Sociais as publicações dos últimos cinco anos mostram-se atuais, visto que a área não apresenta mudanças tão rápidas quanto outras ciências. Desse modo, restringiu-se a busca aos cinco anos compreendidos entre 2013 a 2017, na subárea Biblioteconomia e Ciência da Informação. Como esta pesquisa é interdisciplinar, buscou-se publicações também na Ciência da Computação e nas subáreas Sistemas de Informação e Inteligência Artificial (IA). Contudo, a periodicidade permaneceu a mesma devido ao fato de a área principal ser as Ciências Sociais.

Na etapa 3, *Buscando a Literatura*, acontece a execução da busca, em que as expressões apresentadas no Quadro 10 são executadas nos mecanismos de consulta das bases de dados bibliográficos selecionadas com foco em publicações escritas nos idiomas inglês e português. No total, essa etapa resultou em 9.386 (nove mil trezentos e oitenta e seis) artigos, excluindo os duplicados. O gráfico da Figura 39 mostra o resultado total da busca considerando cada expressão. Em uma análise quantitativa realizada superficialmente, as expressões de busca E3 e E6 obtiveram mais revocação. Em relação às bases de dados, como pode ser visto na Figura 40, a *Web of Science* e a Scopus geraram mais resultados nessa etapa. As bases EBSCO-ISTA e LISA não retornaram resultados

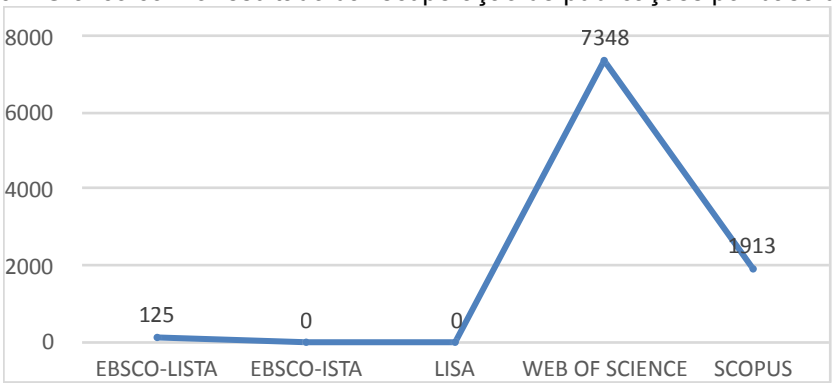
Figura 39 – Gráfico com o resultado da recuperação de publicações por expressão de busca



Fonte: Elaborada pela autora.

A etapa subsequente à de busca foi a *Tela Prática* (etapa 4). Nela, Okoli e Schabram (2010) sugerem a definição de critérios para inclusão dos artigos. Nesse sentido, definiu-se que a análise de assunto a partir do título seria o principal critério para a seleção dos documentos pertinentes para atender os objetivos da revisão de literatura. Nesse caso, os termos contidos no título deveriam explicitar o assunto referente ao problema de pesquisa. Assim, foram incluídos trezentos e oitenta e quatro artigos pelo título, o que correspondeu a aproximadamente 4,09% do total de artigos recuperados na busca.

Figura 40 – Gráfico com o resultado da recuperação de publicações por base de dados



Fonte: Elaborada pela autora.

Na etapa 5, de *Avaliação de Qualidade*, os artigos selecionados para serem incluídos na revisão de literatura foram separados de acordo com o assunto abordado. Dessa forma, os artigos foram discriminados, conforme o título, nas seguintes categorias: ontologia, classificação de textos, DBPedia, extração de informação, *Linked Data*, modelagem conceitual, modelagem facetada, RDF (*Resource*

Description Framework), redes semânticas, texto-hipertexto, grafos de conhecimento, grafos semânticos, *Web* semântica e, por fim, relações semânticas (englobando explicitação, extração, similaridade e outros). Por conseguinte, os artigos referentes à explicitação de relações semânticas (36 artigos) e à extração de relações semânticas (37 artigos) foram selecionados para prosseguir na revisão de literatura.

Os 73 artigos escolhidos foram submetidos a uma avaliação que considerou o número de citações no Google Acadêmico. Essa ferramenta foi escolhida pelo fato de indexar artigos de diversas fontes, pois os artigos selecionados são decorrentes de três bases de dados bibliográficas diferentes. Devido ao fato dos artigos serem muito recentes, eles obtiveram uma média de 4,5 citações no Google Acadêmico. Assim, decidiu-se realizar uma análise qualitativa a partir do conteúdo do resumo dos artigos com mais de cinco citações. O resultado dessa análise selecionou 21 artigos para a próxima fase, a de extração de dados.

Na sequência, na etapa 6 da metodologia de Okoli e Schabram (2010), intitulada *Extração de Dados***,** avaliou-se as palavras-chave e os objetivos de cada artigo, chegando a um total de nove artigos pertinentes para a pesquisa. Nesses artigos, colheu-se os seguintes dados: abordagem, tipos de relação, contexto do problema, domínio do conhecimento e idioma. Esses dados foram analisados e estão relatados a seguir.

4.1 Trabalhos sobre extração de relações semânticas

Esta seção se constitui das duas últimas etapas da metodologia utilizada para a revisão de literatura, conforme proposto por Okoli e Schabram (2010), quais sejam, *Análise dos Achados* e *Escrita da Revisão***.** Nela são apresentados os trabalhos selecionados apontando a metodologia/abordagem utilizada na pesquisa acerca da extração relações semânticas e os resultados obtidos.

A literatura indicou algoritmos com diferentes métodos utilizando diferentes recursos de extração de relações, o que possibilitou o conhecimento sobre as aplicações dos recursos de extração de relações, vislumbrando a possibilidade de alguns deles serem utilizados na elaboração do modelo de extração de relações

semânticas desta tese. A apresentação dos artigos segue uma ordem cronológica, iniciando-se da publicação mais antiga para a mais recente.

Em um esforço entre pesquisadores da Universidade Federal do Rio de Janeiro e da *National University of Ireland*, Carvalho, Freitas e Silva publicam o "*Graphia: Extracting Contextual Relation Graphs from Text*" em 2013. O *Graphia*[33] é uma aplicação *web* que tem como entrada textos em linguagem natural e produz como saída grafos. Cada nó do grafo representa uma entidade (nomeada ou não) e as arestas indicam as relações entre os nós. Para a representação contextual, o *Graphia* introduz um modelo de representação e interpretação de grafo de dados estruturados centrado na entidade, *Structured Data Graph* (SDG). Esse modelo representa triplas do tipo sujeito-predicado-objeto, em que o predicado expressa a relação entre o sujeito e o objeto. Resumidamente, o *Graphia* funciona da seguinte forma. Primeiramente ele (1) realiza a análise sintática do texto utilizando árvores de análise sintática e, na sequência, (2) compatibiliza as entidades nomeadas com URIs DBPedia[34] por meio de um componente de resolução de entidade nomeada. Quando isso não é possível, o sistema emprega POS *tags* nas estruturas de árvore de análise sintática geradas no início. A saída desse componente é o texto original com um conjunto de termos de entidade nomeadas anotadas com URIs. Depois, o *Graphia* (3) aciona o módulo de resolução de co-referência pessoal e normalização que determina os pronomes pessoais, possessivos e reflexivos. As instâncias dos pronomes pessoais são substituídas pelas entidades correspondentes. Os pronomes possessivos e reflexivos são anotados com as entidades correspondentes que mais tarde definirão os *links* de co-referência. O processo também considera o gênero e o número do pronome. A saída dessa fase são árvores de análise sintática com entidades nomeadas anotadas e substituições de co-referência para os pronomes anotados com as entidades nomeadas. Na sequência, a aplicação (4) extrai grafos das árvores de análise sintática resultantes da etapa anterior e gera árvores triplas para cada sentença por meio de regras baseadas em padrões sintáticos e (5) executa o componente de construção do grafo. Esse componente recebe as árvores triplas do componente anterior e produz a serialização final do grafo, criando URIs locais para

---

[33] Disponível em <http://graphia.dcc.ufrj.br/>. Acesso em 16 mar. 2017.

[34] *Uniform Resource Identifier (URI)* é um identificador único para um recurso na Web. Desse modo, URIs DBpedia são identificadores de recursos do DBpedia. O DBpedia, por sua vez é a base de dados do Wikipedia. Disponível em <http://wiki.dbpedia.org/about>. Acesso em 20 de mar. 2017.

cada recurso que não foi definido anteriormente no DBpedia. Como resultado, o *Graphia* fornece um modelo de representação flexível e extensível que combina informações linguísticas e dados abertos vinculados. Os autores não especificam o tipo de relação semântica que o *Grafia* extrai, porém entende-se que ele pode ser utilizado para quaisquer tipos de relações. Outrossim, ele pode ser empregado em qualquer domínio do conhecimento. Segundo os autores, o projeto ainda está ativo e uma versão para o português está atualmente em desenvolvimento. A versão original faz extração apenas para textos em inglês e recentemente tem tido apenas manutenção e otimizações pontuais. Desde sua concepção, o *Graphia* mantém um caráter experimental com a disponibilização de um sistema de demonstração na *Web*. Pessoalmente, os autores relataram que esperam adicionar a entrada em português no sistema de demonstração em breve.

O artigo intitulado “Extracção de Relações Semânticas de Textos em Português Explorando a DBpédia e a Wikipédia”, publicado por Batista, Forte, Silva, Martins e Silva (2013), apresenta contribuições para a extração de relações semânticas no português europeu no contexto da Wikipédia. O método apresentado por eles emprega aprendizagem de máquina. Logo, o primeiro momento envolve a criação e a revisão manual do conjunto de treinamento em que (a) recolhe-se do DBPedia todas as relações expressas correspondentes às entidades pessoas, locais, ou organizacionais e para cada uma das relações recolhidas mantém-se a informação sobre as duas entidades relacionadas e o tipo que denota a classe da relação semântica; na sequência, (b) analisa-se o texto dos artigos da Wikipédia correspondentes para cada relação entre entidades, tal como extraída anteriormente; (c) segmenta-se o texto dos artigos da Wikipédia em suas frases constituintes; (d) filtra-se as frases de modo a manter somente aquelas em que co-ocorrem as duas entidades envolvidas na relação; e (e) mantém-se as frases que resultam da etapa anterior como exemplares de um determinado tipo de relação semântica. Após a definição desse conjunto de dados de treinamento, o método é executado em dois estágios. No primeiro, realiza-se a análise do conjunto de frases envolvido na indexação dos exemplares de treinamento. Esse estágio é constituído das seguintes atividades: (a) extração de conjuntos de tetragramas de caracteres, preposições, verbos e padrões relacionais das *sub-strings* que ocorrem antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação; (b) extração de assinaturas *min-*

hash[35] a partir dos conjuntos gerados anteriormente; e (c) divisão das assinaturas em bandas e indexação dos exemplares de relações que elas representam em diferentes tabelas de dispersão com base nos valores presentes nas bandas das assinaturas. No segundo estágio, ocorre a classificação e a verificação da relação semântica. Nesse momento, repetem-se as atividades (a) e (b) do estágio anterior e, em seguida, são executadas as seguintes atividades: (c) (i) considera-se as relações de exemplo com pelo menos uma banda idêntica no índice construído na fase de indexação como candidatas e a sua semelhança com a relação a classificar é estimada usando as assinaturas *min-hash* completas; (ii) mantém-se os exemplos mais semelhantes numa lista de prioridades, de onde posteriormente se podem extrair os *kNN*[36] (*k-nearest neighbor*) exemplares mais semelhantes; (iii) analisa-se os *kNN* exemplares mais semelhantes e a classe semântica da relação é atribuída com base numa votação ponderada pelo valor de similaridade entre as classes presentes nos *kNN* exemplares mais semelhantes. Segundo os autores, testes com um conjunto de dados da Wikipédia comprovam a adequabilidade do método proposto, sendo que o mesmo é, por exemplo, capaz de extrair 10 tipos diferentes de relações semânticas, 8 deles correspondendo a tipos de relações assimétricos, com um *F-Score*[37] médio de 55,6%. Cabe ressaltar que a classificação dos tipos de relações semânticas apresentada pelos autores é diferente da apresentada nesta tese, que categoriza três tipos: hierárquica, equivalente e associativa.

Blanco e Moldovan (2013), em um artigo intitulado "*Composition of Semantic Relations: Theoretical Framework and Case Study*", apresentam o *Composition of Semantic Relations* (CSR), um algoritmo que, de acordo com os autores, pode ser aplicado para todos os tipos de relações e ser utilizado em qualquer domínio do conhecimento no idioma inglês. O CSR revela relações a partir de relações previamente extraídas. Assim, os autores tiveram que definir previamente um inventário de relações para constituir os valores de domínio e abrangência; o domínio estabelece os valores de entrada e a abrangência determina todos os valores que

---

[35] *Min-hash* é uma técnica para calcular a similaridade entre dois conjuntos. Disponível em <https://en.wikipedia.org/wiki/MinHash>. Acesso em 02 de jun. 2017.

[36] O kNN é um algoritmo utilizado para classificação de dados. A decisão do vizinho mais próximo atribui a um ponto de amostra não classificado a classificação do mais próximo de um conjunto de pontos previamente classificados (COVER; HART, 1967).

[37] O *F-Score* é uma pontuação composta que combina medidas de precisão e revocação. Quanto mais próximo ao numeral 1, melhor o valor (MANI; MAYBURY, 1999).

podem ser gerados a partir dessa entrada. Além das restrições de domínio e abrangência, o algoritmo CSR utiliza primitivas semânticas de acordo com a abordagem de Huhns e Stephens (1989), que especificam propriedades que determinam uma relação semântica entre um elemento do domínio e da abrangência da relação que está sendo descrita. Nesse caso, os valores para as primitivas de relações semânticas foram definidos como X = {+, 0, -} em que o sinal + indica que a relação semântica é pertinente, - que não é pertinente e 0 que não é aplicável. A partir da aplicação das primitivas semânticas no domínio e na abrangência, o CSR gera inferências para compor relações. Essas inferências auxiliam na automatização da descoberta de relações semânticas. Os autoreas realizaram um estudo de caso aplicando o CSR e utilizando um conjunto de 26 relações, o que possibilitou identificar 216 axiomas, dos quais um subconjunto de 8 axiomas foi avaliado com precisão de 0,86 usando uma heurística. Eles acreditam que o resultado foi útil para uma abordagem não supervisionada, contudo, eles apontam que o modelo tem limitações e nem sempre está correto, isso porque, para eles, as relações e a álgebra para a composição de primitivas são definidas manualmente, o que pode ocasionar erros. Ainda assim, a pesquisa de Blanco e Moldovan (2013) foi importante por apresentar a aplicação das restrições de domínio e abrangência para a descoberta de relações semânticas.

Li *et al.* (2013), no artigo "*A Relation Extraction Method of Chinese Named Entities Based on Location and Semantic Features*", utilizam o reconhecimento de entidade nomeada para a extração de relações e empregam propriedades de localização de entidades em um método *LaSE (Location and Semantic Extraction)*. O *LaSE* foi desenvolvido para a língua chinesa. De acordo com os autores, ele é escalável, semissupervisionado e independente de domínio. Na sua descrição não está especificado o tipo de relação semântica que o método extrai, contudo, entende-se que ele pode descobrir todas as variedades de relações semânticas. O algoritmo *LaSE* tem como entrada sementes[38] e um *corpus.* A partir dessas entradas, ele executa três etapas. A primeira minera padrões de extração de relações reais fazendo a correspondência das sementes com o *corpus*. Eles descrevem um padrão de extração de relações como uma tupla de quatro elementos {*lV*, *mV*, *rV*, *R*}, em que *lV,*

[38] As sementes são um conjunto de dados de treinamento, nesse caso relações semânticas, definidas manualmente, usadas para a aprendizagem de máquina nos métodos semissupervisionados e supervisionados.

mV e *rV* que referem-se a vetores que associam pesos com palavras e *R* é uma relação de entidade nomeada. Os vetores *lV, mV* e *rV* são obtidos a partir da análise das posições das entidades nomeadas na sentença. Essa análise possibilita a verificação de qual palavra está à esquerda (*lV - left vector*), no meio (*mV - middle vector*) e à direita (*rV - right vector*) da entidade nomeada. Na segunda etapa, os padrões de extração de relações descobertos são generalizados para redução de seu número e para produzir padrões mais gerais e úteis. Por fim, na terceira etapa extrai-se os resultados das instâncias de relação usando uma função de correspondência entre os candidatos à instância e as POS t*ags* cujo rótulo seja de relação (realizadas anteriormente durante a mineração dos padrões de extração de relação) que ocorrem na sentença analisada. Como resultado, os experimentos realizados pelos autores mostram que o *LaSE* tem um *F-Score* de 0,879 que é, segundo eles, pelo menos 0,113 melhor que os métodos de extração existentes que usam características de localização ou características semânticas. A combinação de padrões de extração de relações, POS *tags* e vetores que indicam os pesos das palavras em suas localizações fazem do *LaSe* um algoritmo eficiente, sobretudo por se tratar de aprendizagem semi-supervisionada.

Em prosseguimento, Xu *et al.* (2013), em uma publicação intitulada "*Discovering Missing Semantic Relations between Entities in Wikipedia*", desenvolvem uma abordagem descrita em três passos para descobrir relações semânticas entre entidades na Wikipédia. O primeiro passo é a identificação de menções de entidade nos *infoboxes* (quadros de informação resumida na Wikipédia). Esse passo extrai referências de entidades nos *infoboxes* a partir de um dicionário de menções que inclui todas as menções de entidade na Wikipédia. Esse dicionário registra as possíveis entidades que cada menção deve se referir e é representado por $D = (M, E)$, em que $M = \{m_1, m_2, ..., m_k\}$ é o conjunto de todas as menções da Wikipédia e $E = \{E_{m1}, E_{m2}, ..., E_{mk}\}$ é o conjunto de todas as entidades correspondentes para as menções em *M*. A abordagem extrai menções em *infoboxes* pela correspondência de todas as *n*-gramas de valores de atributos com menções *M* no dicionário *D*. O resultado é um conjunto de menções que foram encontradas através dos *n*-gramas. O segundo passo é o cálculo de características, o algoritmo calcula sete características do par de menção-entidade, realizando previamente uma consulta ao algoritmo para verificar a possibilidade de duas entidades relacionarem-se semanticamente. As sete

características são: (a) ocorrência da entidade; (b) probabilidade de *link*; (c) relacionamento com o contexto do *infobox*; (d) relacionamento com o contexto do artigo; (e) relacionamento com o contexto do resumo; (f) relacionamento com o contexto do atributo abrangência; e (g) relacionamento com o contexto do atributo domínio. Todas essas características envolvem cálculos matemáticos que estimam o quanto elas estão satisfeitas entre as menções dos *infoboxes* com as entidades. Por fim, o terceiro passo pauta-se no aprendizado para predizer novas entidades. Nele, um modelo de aprendizagem de máquina é empregado para predizer novos *links* de entidades. Para isso, a abordagem calcula o peso da soma das características (calculadas anteriormente) entre as menções e as entidades. Os autores avaliam sua abordagem na Wikipédia em inglês e os resultados apontam que a abordagem pode efetivamente encontrar relações entre entidades que, segundo eles, o *f-score* supera métodos semelhantes. Pode-se dizer que essa abordagem é importante sobretudo devido ao contexto de uma ferramenta colaborativa, como a Wikipédia. Os autores não especificam o tipo de relação que tratam; contudo, interpreta-se que o modelo pode ser aplicado para todos os tipos de relação. Outrossim, o modelo pode ser utilizado em qualquer domínio do conhecimento.

Zhu *et. al* (2013) apresentam o artigo "*Detecting Concept Relations in Clinical Text: Insights from a state-of-the-art model*", no qual exploram as aplicações das relações semânticas entre os conceitos do domínio da área médica que envolvem problemas de saúde, exames e tratamentos em textos clínicos. Esse método apresenta cinco etapas. Na primeira etapa, que prevê o treinamento para o aprendizado supervisionado, são usados um conjunto de dados de treinamento composto de textos clínicos de diversas instituições da área da saúde; posteriormente, há uma classificação de acordo com termos da área médica. Para a classificação desses dados e para a categorização das relações, os autores utilizam entropia máxima (isso depois de testarem máquina de vetor de suporte, regressão logística e problema do vizinho mais próximo – kNN). A segunda etapa é de detecção do conceitos nos textos, detecção que é tratada como um problema de reconhecimento de entidade nomeada. Zhu *et. al* (2013) exploram aspectos relacionados às características de palavra/frase/conceito, como o número de conceitos em uma sentença, características rígidas (que incorporam um modelo estatístico de regras rígidas) e características relacionadas a *n-gramas* e a pontuações. Entre os recursos

empregados nessa etapa está a utilização de um algoritmo de *cluster* dentro de uma árvore binária, em que cada nó não-folha converge para as palavras semanticamente similares, criando um *sub-cluster* de palavras. A terceira etapa sugerida pelos autores é a determinação do domínio semântico. Nela os autores exploram a efetividade dos domínios manualmente criados, ou seja, os domínios de conhecimento explícito, como o *Unified Medical Language System* (UMLS – Sistema de Linguagem Médica Unificada), criado e mantido pela Biblioteca Nacional de Medicina dos Estados Unidos, e os automaticamente adquiridos de um grande volume de textos do domínio, como o MEDLINE, uma base de dados bibliográfica de informação acerca das ciências da vida e biomedicina. A quarta etapa é a de sintaxe, em que é empregada a análise sintática para grafos de dependência utilizando o cTAKES (*clinical Text Analysis and Knowledge Extraction System*), uma arquitetura de gerenciamento de informação não-estruturada para a realização de marcações POS para as palavras que aparecem entre dois conceitos e para as palavras da árvore localizadas antes e depois de cada conceito. Por fim, a quinta e última etapa é a de composição de núcleos: árvores de *kernel* são integradas para encontrar o melhor dos candidatos à relação. Na avaliação utilizando *f-score*, foi obtido o resultado 0,742, o que os permitiu concluir que as estruturas sintáticas complexas podem melhorar a qualidade de modelagem para a tarefa semântica, mesmo quando a semântica de domínio já foi cuidadosamente utilizada. Ainda assim, o rigor de um sistema de apoio à decisão do domínio médico exige um cuidadoso estudo de estratégias para melhor definir as relações semânticas, por isso o modelo apresentado pode ser considerado complexo.

Em "*Discovering Semantic Relations from Unstructured Data for Ontology Enrichment Asssociation rules based approach*", Paiva *et al*. (2014) desenvolvem um método baseado em mineração de regras de associação para o refinamento de relações de uma ontologia de domínio na área de construção civil em um processo que envolve quatro etapas. A primeira é a análise de documento, atividade que divide o documento em sentenças e das sentenças são extraídos os *tokens*; então, os termos coincidentes existentes na *stop list* são removidos e, em seguida, os *n-gramas* são colocados em um vetor estatístico. Na segunda etapa da abordagem ocorre a aplicação do algoritmo *FP-Growth*, que descobre a frequência de conjuntos de termos que aparecem no vetor estatístico. Na terceira etapa regras de associação são aplicadas para indicar, por exemplo, se um conceito A aparece no texto, qual a

probabilidade de ocorrer o conceito B? As regras de associação utilizadas são: (a) confiança, denota a probabilidade do conceito B dado o conceito A; (b) interesse, indica a independência entre dois conceitos; (c) convicção, ajuda o usuário a utilizar as duas primeiras regras e mede a implicação de um conceito dado outro; (d) suporte, calcula a frequência dos conceitos; (e) influência, identifica a diferença entre o real e o esperado, indicando a independência dos conceitos; (f) *laplace*, estima a confiança e indica, por exemplo, se o suporte do conceito A diminui se a relevância de tal conceito também diminui. Com a aplicação dessas regras, encontra-se a equivalência dos conceitos na ontologia. Para tal, são realizados cálculos matemáticos que indicam a similaridade dos conceitos. A quarta e última etapa é a de mapeamento dos conjuntos de itens. Os autores utilizam um banco de dados de regras de associação que armazena as regras definidas pelo módulo de mapeamento de conjuntos de itens frequentes. Em seguida, em uma interface são mostrados os itens, os itens candidatos e a porcentagem de frequência dos pares de itens. Por fim, utilizando um *corpus* de 40 artigos científicos publicados na área de construção civil, Paiva *et al*. (2014) afirmam que os resultados indicam que a mineração de regras de associação pode ser um instrumento interessante para explorar relações semânticas em fontes não estruturadas. Além disso, o mapeamento de conjuntos de itens frequentes é útil para a redução do número de regras e para a identificação do nível de similaridade entre os conjuntos de itens frequentes e de termos equivalentes na ontologia. Assim, o suporte estatístico do método possibilita confiança no estabelecimento das relações semânticas.

Arnold e Rahm (2014), em "*Extracting Semantic Concept Relations from Wikipedia*", publicam uma estratégia para criar automaticamente repositórios extraindo relações semânticas de artigos da Wikipédia. Para isso, eles criam uma abordagem de cinco etapas. A primeira é a extração de artigos da Wikipédia, com foco em cada nome e seção de resumo do artigo. Caso o artigo não tenha esses dois elementos (nome e resumo), os autores utilizam a estratégia de extrair os 750 primeiros caracteres do artigo pois, segundo eles, geralmente esses caracteres contêm informação relevante necessária para a sua abordagem. A segunda etapa da abordagem é o pré-processamento dos artigos extraídos; nessa fase, os autores retiram a primeira sentença de cada artigo da Wikipédia, considerada como a sentença de definição. Ainda nessa etapa os autores assinalam os termos da

sentença utilizando POS *tags* e simplificam a mesma sentença utilizando um algoritmo para esse fim. A terceira etapa identifica padrões de relações semânticas. Assim, caso exista um número *n* de padrões ($n \geq 1$), eles dividem a sentença naqueles padrões e assim obtém ($n+1$) fragmentos de sentença; caso contrário, passa-se para o próximo artigo da Wikipédia. Em seguida, aplica-se um modelo matemático de máquina de estado finito para cada palavra da sentença para descobrir os padrões de relações semânticas das sentenças. Os padrões de relações semânticas cobrem relações de hipônimo-hiperônimo, merônimo-holônimo e equivalência. A quarta etapa analisa os fragmentos de sentenças. Para cada fragmento de sentença, buscam-se os conceitos relevantes ligados pelos padrões de relações semânticas. Logo após, a sentença é analisada para identificar os conceitos ou lista de conceitos que participam da relação semântica. Os substantivos diretamente à esquerda e à direita do padrão de relação semântica representam conceitos relevantes. Por fim, na última etapa, há a determinação das relações semânticas. Selecionados os termos e padrões, as respectivas relações semânticas são construídas e exportadas para um repositório, tudo automaticamente. Em uma avaliação, utilizando como referência a precisão e a revocação, foi apontada uma alta efetividade dessa abordagem para diferentes domínios, que mostrou-se ser interessante sobretudo devido ao contexto ao qual ela é aplicada. Contudo, a utilização de padrões de relações é oportuna para a explicitação de relações, porém esses padrões não permitem extrair relações associativas.

No artigo intitulado "*Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery*", Kliegr (2015) sugere a criação de um banco de dados de hiperônimos vinculados, chamado LHD (*Linked Hypernyms Dataset*), para artigos da Wikipédia nos idiomas inglês, holandês e alemão. Para tal, ele executa quatro passos. No primeiro, um algoritmo chamado THD (*Targeted Hypernym Discovery* – Descoberta de Hiperônimos Marcados) é utilizado para análise linguística do LHD. Esse algoritmo tem as seguintes características: somente a primeira sentença do artigo Wikipédia é processada; somente o primeiro hiperônimo é extraído; alguns tipos de artigos do Wikipédia são excluídos; para hiperônimos de múltiplas palavras, o resultado é o último substantivo; hiperônimos contidos nos nomes de entidade e nos títulos dos artigos Wikipédia são ignorados; hiperônimos genéricos comuns que precedem um hiperônimo mais específico são ignorados. No segundo passo, após a aplicação das características, a saída do THD é lematizada por meio do processo de

retirar as flexões para determinar o seu lema. No terceiro passo determinam-se os *links* de hiperônimo realizando as seguintes tarefas: (a) desambiguação, uma API (*Application Programming Interface*) de buscas na Wikipédia usa um algoritmo como o *PageRank*[39] para determinar a importância do artigo e a saída desse passo são os hiperônimos vinculados; e (b) limpeza dos dados, que executa substituições e eliminações de acordo com regras definidas manualmente. Por fim, no quarto passo, realiza-se um alinhamento entre as bases DBpedia e YAGO[40] para melhor precisão das relações definidas. A avaliação da qualidade é realizada com base em 16.500 avaliações e anotações humanas. Como resultado, obtém-se *f-score* 0,90, isso aponta que a abordagem tem uma precisão eficiente para a descoberta de hiperônimos nos três idiomas analisados. A proposta do autor mostra a extração de relações de hiperônimo em artigos Wikipédia para o uso de bases de conhecimento. Todo o processo é bem detalhado para suprir um tipo de relação muito importante na definição de conceitos na Wikipédia.

Observa-se nesta revisão de literatura que os recursos para a extração de relações são combinados com algoritmos de Inteligência Artificial em diferentes estratégias para a descoberta de relações semânticas em distintos contextos, entre os quais se destaca o ambiente Wikipédia (uma enciclopédia colaborativa) que envolve as próprias páginas da Wikipédia, o DBPedia (a base de dados da Wikipédia) e os *infoboxes* (as caixas de informação localizadas no lado direito das páginas da Wikipédia).

Nesta revisão de literatura não foram encontradas pesquisas cujas relações são extraídas a partir de termos de estruturas classificatórias. Logo, contribuições podem ser realizadas para que os relacionamentos semânticos nessas estruturas possam ser extraídos e explicitados para melhor entendimento dos conceitos envolvidos em uma estrutura de classificação.

Outrossim, nota-se que a maioria das publicações selecionadas podem extrair qualquer tipo de relações e não existe preocupação dos autores em especificar esses tipos de relações. Além disso, prevalecem os artigos cuja aplicação de suas abordagens são independentes do domínio do conhecimento.

---

[39] “PageRank™ é um algoritmo utilizado pela ferramenta de busca Google para posicionar websites entre os resultados de suas buscas”. Disponível em <https://pt.wikipedia.org/wiki/PageRank>. Acesso em 25 mar. 2017.

[40] A YAGO é uma base de conhecimento que extraída automaticamente do Wikipédia. Disponível em <https://en.wikipedia.org/wiki/YAGO_(database)>. Acesso em 25 mar. 2017.

Dos artigos pesquisados, a maioria aborda o problema da extração de relações no idioma inglês, mesmo naqueles em que há participação de pesquisadores brasileiros. Esse fato já foi levantado anteriormente em investigação realizada em anos anteriores a 2012. Abreu, Bonamigo e Vieira (2013) apuraram o estado da arte da extração de relações em português. Na época, eles revelaram que a maioria dos esforços eram de extração de relações no idioma inglês, apesar das contribuições importantes para o português europeu, como as ferramentas computacionais HAREM (HAREM é uma ferramenta de Avaliação de Reconhecedores de Entidades Mencionadas)[41], ReRelME (*Recognition of Relation between Named Entities* – Reconhecimento de Relação entre Entidades Nomeadas)[42], REMBRANDT (Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada de Texto)[43] e o analisador PALAVRAS[44], sendo este o único projeto ativo até o momento, visto que as outras ferramentas foram descontinuadas.

Este capítulo expôs sobre a revisão de literatura realizada para apurar o estado da arte das pesquisas sobre extração de relações. Nesse sentido, primeiramente, detalhou-se sobre a metodologia de revisão de literatura utilizada e sem seguida, apresentou-se os esforços encontrados sobre extração de relações semânticas. O próximo capítulo apresenta o Modelo de Extração de Relações Semânticas elaborado nesta tese.

---

[41] Disponível em <http://www.linguateca.pt/HAREM/>. Acesso em 02 abr. 2017.

[42] Disponível em <http://www.linguateca.pt/HAREM/>. Acesso em 02 abr. 2017.

[43] Disponível em <http://xldb.di.fc.ul.pt/Rembrandt/>. Acesso em 02 abr. 2017.

[44] Disponível em < http://visl.sdu.dk/visl/pt/info/portsymbol.html>. Acesso em 02 abr. 2017.

5 MODELO DE EXTRAÇÃO DE RELAÇÕES SEMÂNTICAS

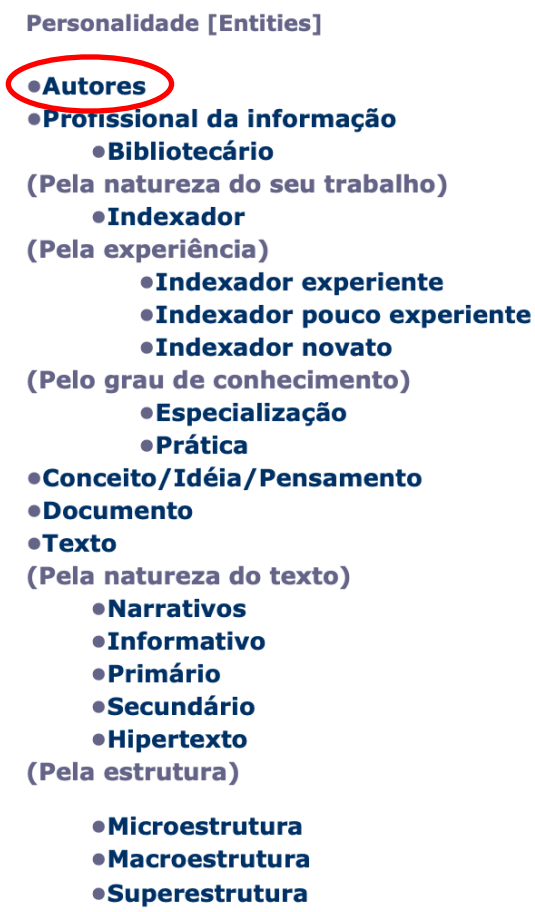
Este capítulo apresenta o Modelo de Extração de Relações Semânticas. Para fins de entendimento, o termo Modelo, quando utilizado nesta tese, refere-se a esse Modelo de Extração de Relações Semânticas, que é uma declaração conceitual de uma abordagem de extração de relações semânticas que será implementada computacionalmente em um protótipo de uma aplicação Web, que recebeu o nome **Semantizar**. Nesse contexto, os termos Modelo de Relações Semânticas e Semantizar serão utilizados intercambiavelmente, uma vez que o modelo e o sistema desenvolvido se convergem. O Semantizar está disponível no *site* www.semantizar.ufop.br.

A prototipação do Semantizar contemplou as atividades de desenvolvimento de *software*, conforme mencionado em capítulo anterior sobre a metodologia, quais sejam: especificação, modelagem de dados, projeto de arquitetura e implementação computacional. Essas atividades estão descritas a seguir.

5.1 Especificação do Modelo de Extração de Relações Semânticas

A especificação do Modelo de Extração de Relações Semânticas contempla uma contextualização para a compreensão sobre qual é a proposta para a qual o Modelo pretende contribuir. Para fins de entendimento, nessa especificação foi utilizado o conceito *autores*, da categoria Personalidade, da estrutura facetada do Protótipo Mapa Hipertextual – MHTX. O conceito autores é o primeiro termo da faceta personalidade, conforme pode ser visto na Figura 41.

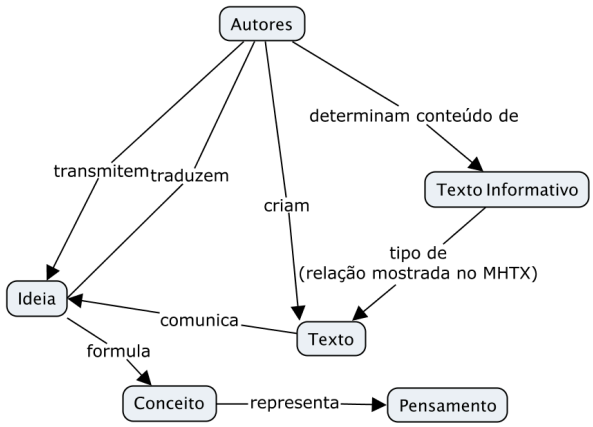
Figura 41 – Recorte da estrutura facetada do MHTX



Fonte: Disponível em <http://www.gercinalima.com/mhtx/pages/prototipo-btdeci/teses/naves-mml/estrutura-facetada.php>. Acesso em 20 mar. 2014.

Ao analisar manualmente essa estrutura, observou-se que o conceito *autores* está relacionado com outros conceitos da estrutura, como pode ser visto na Figura 42. Tal constatação se deu ao examinar manualmente as ocorrências de *autores* na tese de Naves (2000), amostra que foi utilizada para implementação do protótipo MHTX, em decorrência dos outros conceitos da categoria a qual ele está inserido.

Figura 42 – Representação do conceito *autores* explicitamente relacionado a outros conceitos da amostra



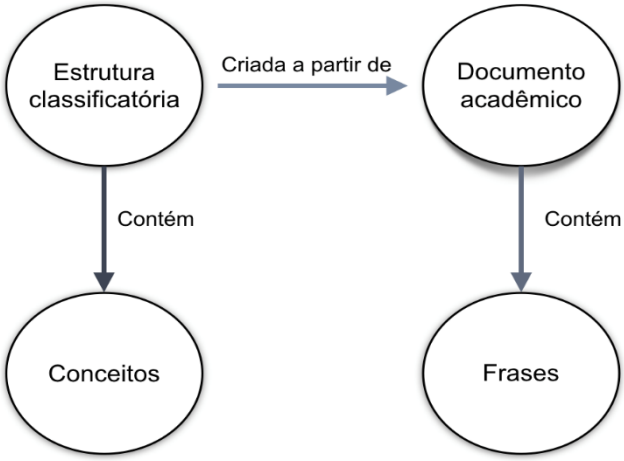
Fonte: Elaborada pela autora.

Pelo que se percebe, as relações entre os conceitos mostradas na Figura 42 estavam implícitas na estrutura classificatória do protótipo do MHTX. Somente após uma análise foi possível tornar claro a existência dos relacionamentos entre tais conceitos e a natureza de sua associação. Como pode ser observado, foi possível descobrir relações entre os conceitos *autores* e *texto, autores* e *texto informativo* e *autores* e *ideia*[45], relações essas que não são explicitadas em uma estrutura facetada por ela ser estritamente excludente. Nesse contexto, criam-se as facetas e os focos, depois de inseridos em uma delas, não poderão estar em outra. Constata-se também, observando a Figura 42, que é possível criar relações entre outros conceitos – como entre *texto* e *texto informativo* – e explicitar relações – como entre *ideia*, *conceito* e *pensamento*.

A estrutura classificatória que trata o Modelo é originária de um documento acadêmico, como pode ser visto na Figura 43; portanto, a estrutura classificatória, por sua natureza, representa o documento acadêmico, isto é, ela contém conceitos que representam um domínio de uma dissertação ou tese. O documento acadêmico, por sua vez, é decomposto em frases.

[45] De acordo com o Acordo Ortográfico de 1990, que entrou em vigor em 2009 e em obrigatoriedade em 2016, a palavra *ideia* não tem acento agudo na vogal *e*. Contudo, na época da tese de Naves, em 2000, época em que o protótipo do MHTX foi implementado, a palavra ainda era acentuada. Por isso, quando em citações, tabelas ou imagens retiradas diretamente do documento original que deu suporte ao MHTX, a palavra idéia será acentuada visando fidedignidade.

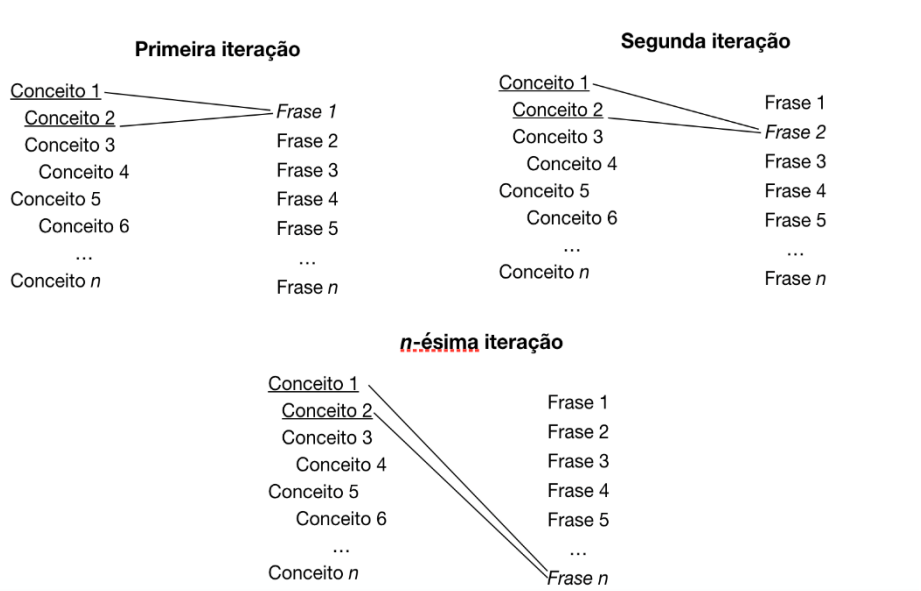
Figura 43 – Composição de uma estrutura classificatória e de um documento acadêmico



Fonte: Elaborada pela autora.

Dos conceitos contidos na estrutura classificatória, para cada par formado por eles, uma busca é realizada em cada frase da dissertação ou tese para verificar se esses dois conceitos existem na frase. A Figura 44 mostra uma representação das iterações das buscas de pares de conceitos nas frases da dissertação ou tese. Como pode ser observado na ilustração, no primeiro momento utilizou-se o par de conceitos 1 e 2. Dado esse par de conceitos, examina-se a existência dele em cada frase até a última delas. Caso o par de conceitos seja encontrado em uma frase, essa frase é destacada para que uma verificação manual confirme a existência, ou não, de uma relação semântica entre os conceitos. Se a confirmação for verdadeira, identifica-se a relação semântica encontrada entre os dois conceitos em determinada frase, indicado pela tripla: sujeito-predicado-objeto.

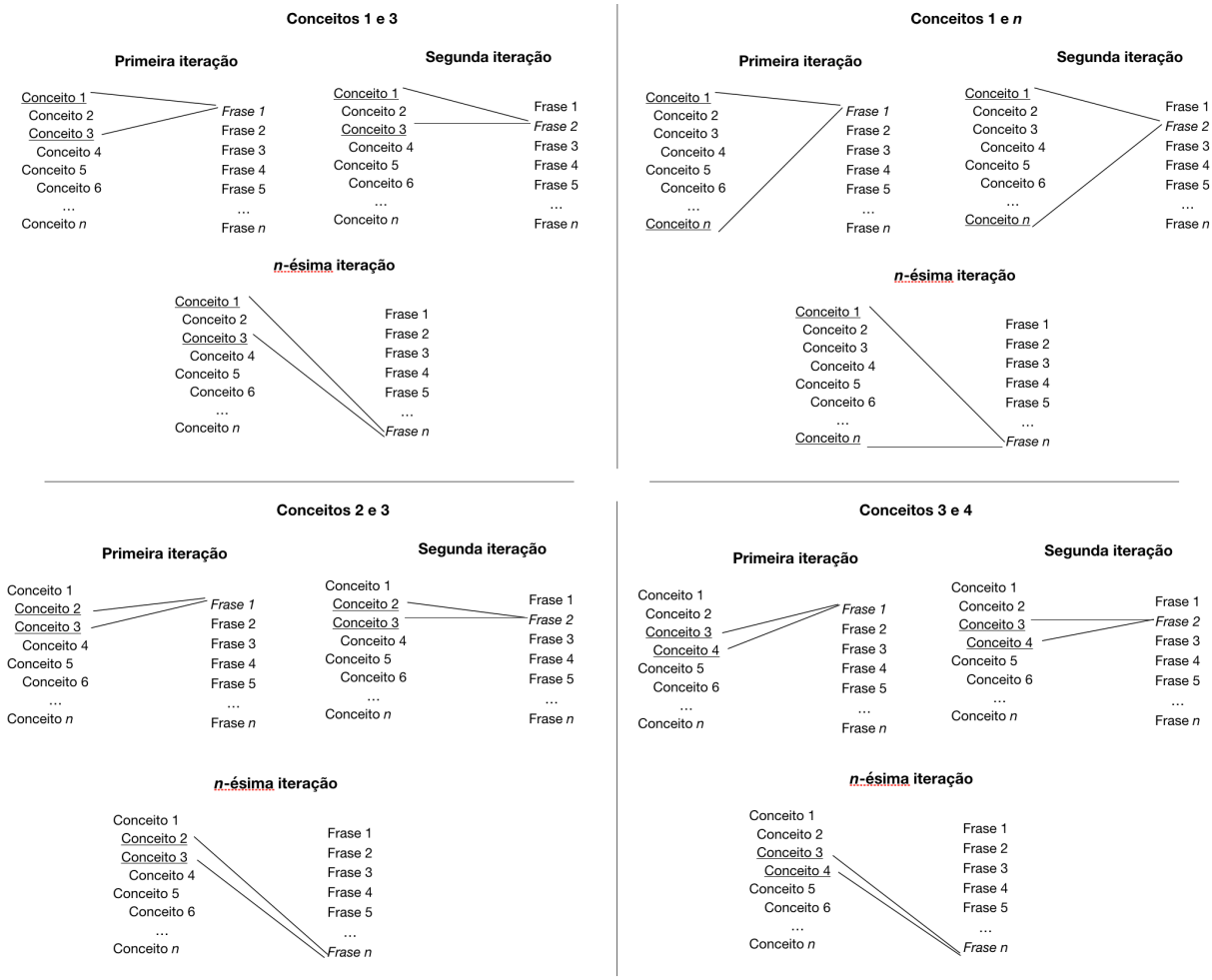
Figura 44 – Iterações das buscas do par de conceitos 1 e 2 nas frases



Fonte: Elaborada pela autora.

Em seguida, verifica-se se existe a combinação dos conceitos 1 e 3, conforme ilustrado na Figura 45. Essa verificação é realizada na primeira frase do documento acadêmico e continua até a última delas. Da mesma forma como entre os conceitos 1 e 2, explicitada anteriormente, ao se constatar os conceitos 1 e 3 em uma frase, essa frase é destacada para que seja verificada a existência de uma relação semântica entre eles. Seguidamente, o Modelo combina o conceito 1 com todos os outros conceitos da estrutura classificatória, verificando em cada combinação se elas existem em uma frase, desde a primeira frase até a última. Ao findar as combinações com o conceito 1, o Modelo subsequentemente faz combinações com o conceito 2 e verifica se as mesmas existem em todas as frases e assim sucessivamente até o último par de conceitos da estrutura classificatória e a última frase do documento acadêmico.

Figura 45 – Iterações de buscas de pares de conceitos nas frases do documento acadêmico



Fonte: Elaborada pela autora.

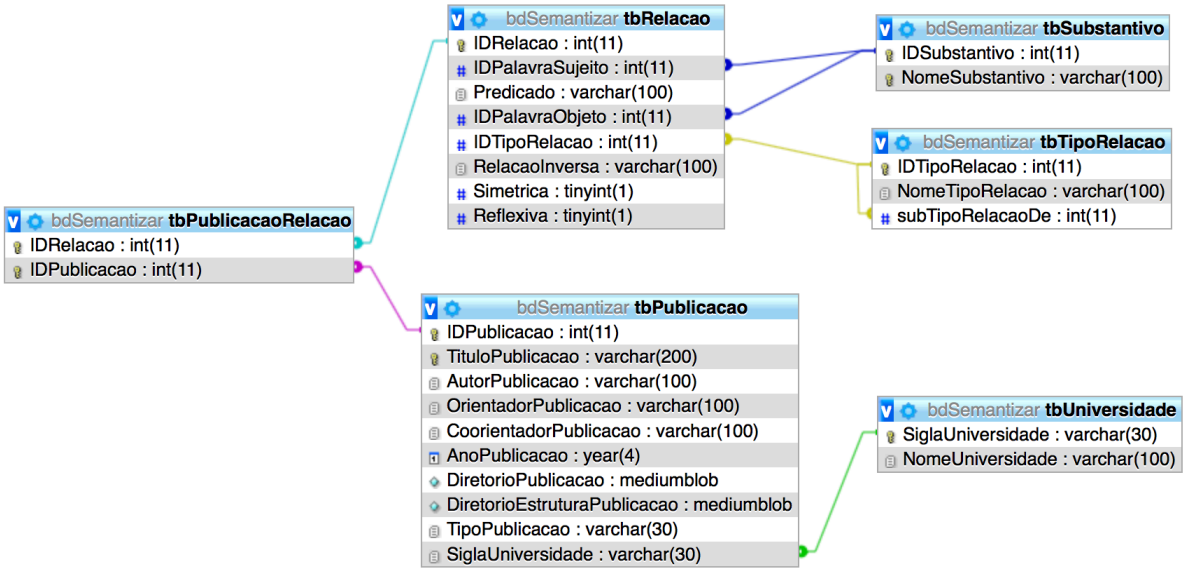
5.2 Modelagem de dados

A modelagem de dados considerou a proposta de criar um sistema chamado Semantizar para implementar o Modelo de Extração de Relações Semânticas, especificado na seção anterior. Nessa fase, foram desenvolvidos o diagrama de entidade-relacionamento (Figura 46) e o dicionário de dados (Apêndice A). Nesses dois instrumentos foram criadas tabelas para armazenar dados de substantivos, publicações[46], relações semânticas e relações por publicação. Além disso, foram

[46] No Semantizar, os documentos acadêmicos foram representados pelo termo “publicação”.

elaboradas as tabelas para armazenar as siglas e os nomes das universidades e os tipos de relações semânticas.

Figura 46 – Diagrama de Entidade Relacionamento do Semantizar



Fonte: Elaborada pela autora.

Iniciando pela tabela *tbTipoRelacao,* ela armazena os tipos de relações semânticas existentes tomando como base a taxonomia definida na seção 3.3.2. Nessa taxonomia, foram levantados 63 tipos de relações semânticas classificadas como relações hierárquicas, associativas e equivalentes. Essas classes, por sua vez, têm subdivisões. Para dar suporte à taxonomia das relações semânticas, a tabela *tbTipoRelacao* contempla o autorrelacionamento por meio do campo *subTipoRelacaoDe*, que permite que um tipo de relação semântica possa ser um subtipo em outra relação. Isso pode ser observado na Figura 47, em que no tipo de relação *Hierárquica,* por exemplo, o subtipo é nulo, conforme destacado na figura. Já o tipo de relação *Objeto Estruturado* tem como subtipo a relação semântica cujo código é 19, que remete ao *Merônimo-Holônimo,* que, por sua vez, direciona para a relação *Hierárquica.*

Figura 47 – Recorte da base de dados da tabela *tbTipoRelacao*

IDTipoRelacao ▲ 1	NomeTipoRelacao	subTipoRelacaoDe
7	Hierárquica	*NULL*
8	Hipônimo-Hiperônimo	7
9	Hipônimo Simples	8
10	Instância	8
11	Inclusão de Classe	8
12	Perceptivelmente Subordinado	11
13	Funcionalmente Subordinado	11
14	Estado Subordinado	11
15	Atividade Subordinada	11
16	Geograficamente Subordinado	11
17	Taxonômica	11
18	Inclusão Espacial	8
19	Merônimo-Holônimo	7
20	Objeto Estruturado	19
21	Componente-Complexo	20
22	Unidade-Organização	20

Fonte: Elaborada pela autora.

Conforme apresentou-se na Figura 46, a tabela *tbsubstantivo* foi criada para armazenar os termos das estruturas classificatórias. A nomeação dessa tabela considerou que todos os termos da estrutura pertencem à classe gramatical dos substantivos; contudo, inicialmente não interessa discriminar o tipo de substantivo, a saber: comum, próprio, simples, composto, concreto, abstrato, primitivo, derivado ou coletivo. O cadastro dos registros dessa tabela é realizado automaticamente pelo sistema. Essa tabela tem a característica de ser considerada simples por armazenar apenas o identificador do substantivo e o substantivo em si. Ressalta-se que a tabela foi planejada para restringir a duplicação de registros de substantivos.

Já a tabela *tbUniversidade* armazena o registro das principais universidades do Brasil, mas não inclui todas. Foram cadastradas 67 siglas e os respectivos nomes das universidades, sendo que as siglas denotam a chave primária da tabela. Esse cadastro baseou-se na lista de universidades federais do Brasil[47], porém outras universidades estaduais e particulares também foram cadastradas. Essa tabela é importante para a tabela *tbPublicacao*, que será apresentada a seguir. A tabela *tbPublicacao* armazena, além da sigla da universidade, dados como o título da publicação (projetado de modo a não permitir duplicação); os nomes do autor, do

[47] Disponível em <https://pt.wikipedia.org/wiki/Lista_de_universidades_federais_do_Brasil>. Acesso em 20 mai. 2017.

orientador e do co-orientador (se houver); o ano da publicação; o tipo do documento, sendo permitidos um dos tipos: tese ou dissertação; e os campos que armazenam os arquivos relativos à publicação e à estrutura classificatória.

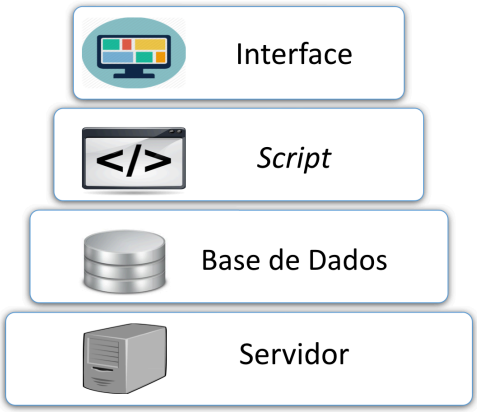
A tabela *tbrelacao* armazena a tripla sujeito-predicado-objeto, em que o sujeito e o objeto são provenientes da tabela *tbsubstantivo* e o predicado (que é a relação semântica) é definido pelo usuário, conforme explicado na seção anterior. Além disso, registra-se o tipo da relação (proveniente da tabela *tbTipoRelacao*, também explicada anteriormente), a relação inversa (quando ela existir) e as propriedades relativas à simetria e reflexividade, conforme explicado na seção 3.3.3.

Por fim, a tabela *tbPublicacaoRelacao* contém campos que armazenam o código que identifica a relação semântica proveniente da tabela *tbRelacao*, explicada logo acima, e o código identificador da publicação, advindo da tabela *tbPublicacao*. Esta última é importante para associar a relação semântica ao seu contexto, que é a publicação em questão.

5.3 Projeto de arquitetura do sistema

Esta seção dispõe sobre a arquitetura computacional do Semantizar, que é um sistema ou aplicação *web* desenvolvido para implementar o Modelo de Extração de Relações Semânticas. A Figura 48 mostra os quatro componentes que compõem essa arquitetura.

Figura 48 – Arquitetura computacional do sistema



Fonte: Elaborada pela autora.

A base da arquitetura criada para esse sistema é a camada do *servidor*. Como o sistema está projetado para funcionar no ambiente *web*, um servidor de aplicações *web* foi configurado para dar suporte à linguagem de programação escolhida. Nesse sentido, foi instalado o Apache HTTP Server 2.4.27[48], que é licenciado pela *Apache Software Foundation*, com licença permissiva que consente o uso não comercial[49].

Além do suporte para a linguagem de programação escolhida, a camada de servidor suporta a camada de *base de dados*. A escolha da base de dados considerou o tipo de licença de uso de *software* e a sua interface com o servidor Apache e a linguagem de programação. Logo, optou-se pelo Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL 5.7.18[50], que é um código livre, sob licença da Oracle e, assim como o Apache, o uso não comercial é permitido.

A próxima camada, a de *script*, é a camada lógica em que o código fonte da aplicação será construído. Para isso, a linguagem de programação PHP (*Hypertext Preprocessor*) versão 7, que é código-aberto, foi utilizada para, entre outras funcionalidades, conectar a aplicação à base de dados, coletar dados de formulários, gerar páginas *web* com conteúdo dinâmico e manipular arquivos com as extensões .txt e .pdf, que são necessários no sistema.

Na linguagem de programação PHP, o código fonte é executado no servidor que, por sua vez, gera o código HTML (*Hypertext Markup Language*) que é enviado para o navegador do cliente, ou seja, para a última camada da arquitetura, a *interface*. Logo, a interface, nessa arquitetura, denota a linguagem de marcação HTML juntamente com o CSS (*Cascading Style Sheets*) para a descrição dos estilos da página, o que configura a formatação dos elementos HTML, como fontes, formulários, cabeçalhos etc. A Figura 49 mostra a comunicação entre os elementos da arquitetura do sistema.

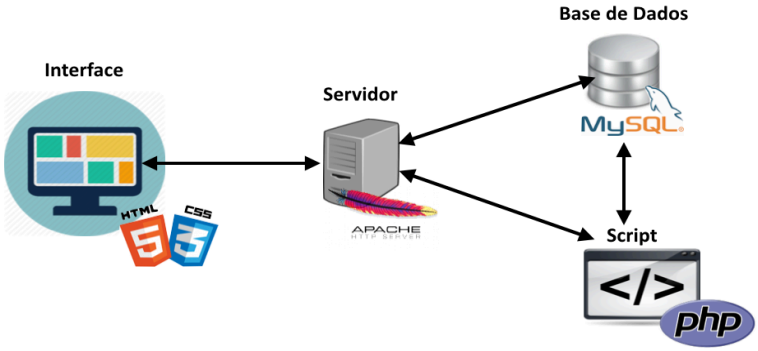
---

[48] Disponível em <https://www.apache.org/>. Acesso em 30 mai. 2017.

[49] Disponível em <http://httpd.apache.org/docs/2.2/en/license.html>. Acesso em 30 mai. 2017.

[50] Disponível em <https://www.mysql.com/> Acesso em 30 mai. 2017.

Figura 49 – Comunicação entre os elementos da arquitetura do sistema



Fonte: Elaborada pela autora.

5.4 Implementação computacional do Modelo de Extração de Relações Semânticas – Semantizar

Inicialmente, o Semantizar apoiaria a extração *automática* de relações semânticas. Para isso, foi criada uma tabela de verbos na base de dados onde cadastrou-se 3.865 verbos em português brasileiro que definiriam essas relações automaticamente[51]. Contudo, durante os testes, percebeu-se que os radicais dos verbos não estavam sendo precisos na recuperação das relações semânticas. Um exemplo recorrente foi a *string* do radical *indexa*, do verbo indexar. Toda vez que o

[51] Esses verbos foram retirados da base de dados *br.ispell* (disponível em: https://www.ime.usp.br/~ueda/br.ispell/), que é uma base de dados de conjugação de verbos da língua portuguesa, versão 1.1, de junho de 1999. Dessa forma, cadastrou-se na base de dados do sistema Semantizar os radicais dos verbos regulares nos três tipos de conjugação, ou seja, os verbos terminados em ar, er e ir. Ao excluir as terminações dos verbos para cadastrar seus radicais, percebeu-se que alguns verbos tinham o mesmo radical. Esses radicais foram unificados, como é o caso de fundar e fundir, gerar e gerir, parar e parir, provar e prover, recobrar e recobrir, cobrar e cobrir, regar e reger, sentir e sentar, vendar e vender, entre outros. Outra observação sobre os radicais diz respeito aos radicais monossílabos, como ge do verbo gear, iç de içar, mi de miar, op de opor, or de orar, pi de piar, entre muitos outros. Nesse caso, optou-se por excluir os verbos cujos radicais tinham apenas duas letras porque eles não teriam tanta relevância na busca pelos verbos nas frases, uma vez que eles seriam fáceis de ocorrer nas palavras. Dessa forma, foram apagados 21 registros de radicais de verbos. Outrossim, durante o cadastro dos verbos, percebeu-se que alguns deles poderiam não ser úteis para o propósito do Modelo, como os verbos que expressam emoções, tais como gargalhar, chorar, rir, entre outros. Da mesma forma foram excluídos os verbos que expressam atividades físicas, como correr e andar, e os verbos relacionados ao tempo, como garoar e chover. Como a base de dados é de 1999, alguns verbos foram atualizados para atender ao acordo ortográfico atual, como é o caso dos verbos que tinham o trema: frequentar, sequestrar e tranquilizar. Além dos radicais dos verbos regulares cadastrados, como mencionado acima, optou-se por incluir na base de dados os verbos auxiliares ser, estar e ter. Nesse caso, eles foram conjugados na 3ª pessoa do singular e do plural no indicativo, nos tempos presente, pretérito perfeito, pretérito imperfeito, pretérito mais que perfeito, futuro do presente e futuro do pretérito. Esses tempos verbais foram escolhidos por se acreditar que eles são usuais em teses e dissertações.

sistema encontrava essa *string* em uma frase em que dois conceitos eram detectados, ele retornava um falso positivo, ou seja, encontrava-se o radical *indexa,* mas a palavra poderia ser *indexação* ou *indexador,* por exemplo, e essas palavras não denotam verbos. É como o caso apresentado na frase: “Nas literaturas inglesa e americana, o termo *indexador* é aplicado tanto àquela pessoa que elabora índices de *textos* ou livros quanto àquela que faz a indexação acadêmica” (NAVES, 2000, p.16, grifos nossos). Nesse caso, o sistema encontrou os conceitos que estão grifados: indexador e textos. Porém, ele erroneamente encontrou o radical do verbo indexar (index) e definiu que indexador também era um verbo e sugeriu a seguinte relação: indexador *indexador* textos**.**

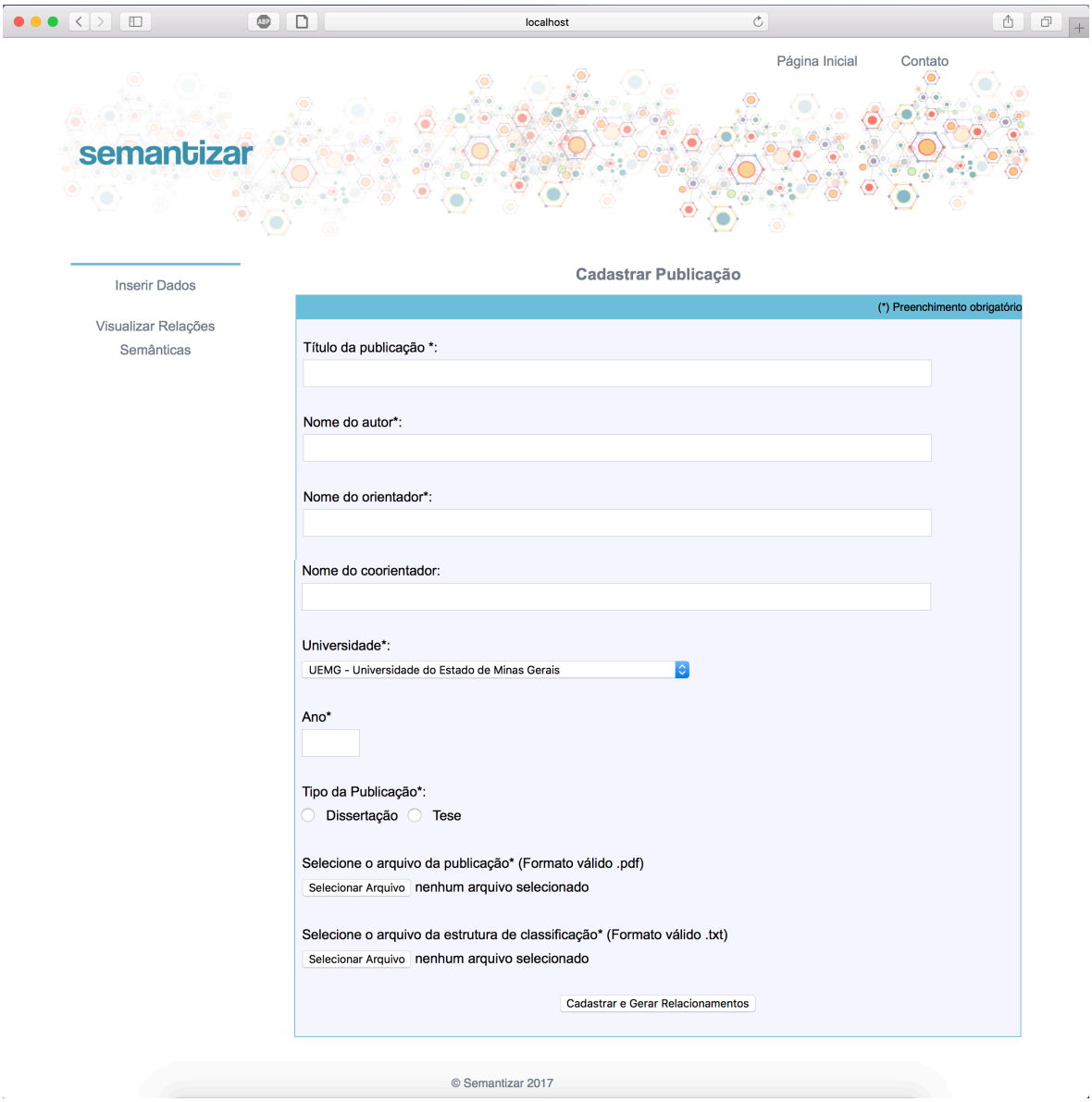
Do mesmo modo, verificou-se que nem sempre as relações semânticas entre os conceitos construídas em linguagem natural nos textos são do tipo sujeito-predicado-objeto. Isso porque as frases são complexas e envolvem outras classes gramaticais, como adjetivos, advérbios, preposições e outros, que ajudam na definição da relação entre os conceitos nas frases. No exemplo apresentado acima, a relação semântica entre *indexador* e *texto* seria: indexador *elabora índices de* textos, sendo que *elabora índice de* seria a relação semântica. Desse modo, ao analisar essa frase e muitas outras, percebeu-se que a construção sujeito-predicado-objeto não é tão direta e algumas vezes não é simples de ser encontrada. Nesse sentido, optou-se pela extração semiautomática de relações semânticas, em que o usuário decide a relação entre dois conceitos em uma frase.

A implementação do Semantizar foi dividida em quatro atividades: (1) entrada de dados, (2) leitura e preparação, (3) extração de relações semânticas e (4) representação do conhecimento.

A primeira atividade, a de *entrada de dados*, é responsável por receber os metadados do documento acadêmico do qual se pretende extrair relações semânticas. Os metadados incluem o título da publicação, o nome do(a) autor(a), o nome do(a) orientador(a) e, se houver, o nome do(a) co-orientador(a), o local de publicação, o ano em que o documento foi defendido e, por fim, o tipo do documento acadêmico, se tese ou dissertação. Além disso, na entrada de dados, envia-se os arquivos referentes à publicação e à estrutura classificatória, sendo que a publicação é um arquivo do tipo .pdf (*Portable Document Format* – Formato Portátil de Documento) e a estrutura classificatória é um arquivo de texto simples de

extensão .txt. A Figura 50 mostra a interface de cadastro implementada para esta etapa do Semantizar[52].

Figura 50 – Interface da atividade de entrada de dados do Semantizar



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

A atividade subsequente, de *leitura e preparação*, verifica se cada termo da estrutura classificatória existe na base de dados. Se o termo não existir, ele é

[52] Nesta versão do Semantizar, não foi realizada a validação dos dados que o usuário informa. Desse modo, na próxima versão, deverá ser implementada, além da validação da entrada de dados, os tipos dos arquivos que o usuário envia para o sistema. Assim sendo, deverá ser realizado, entre outras coisas: (1) um tratamento do arquivo em formato .pdf para validar se está bloqueado ou não; (2) um tratamento de erro caso o tamanho do arquivo exceda o limite permitido para o armazenamento na base de dados e; (3) um tratamento para a estrutura classificatória em formato .txt quando o usuário informar, nesse tipo de estrutura, marcadores, tais como números, traços e letras antes dos conceitos.

cadastrado automaticamente pelo sistema. A linguagem de programação PHP, escolhida para a implementação do Modelo, permite que arquivos de texto sejam convertidos em vetores. Dessa forma, o arquivo referente à estrutura classificatória é convertido automaticamente em um vetor. Cada linha da estrutura (que se refere a um termo) é transformada em uma posição do vetor. Logo, o algoritmo percorre cada posição do vetor verificando se o conteúdo da posição, que é o termo da estrutura classificatória, existe na tabela criada para esse fim no banco de dados; caso não exista, o termo é automaticamente cadastrado pelo sistema. A Figura 51 mostra um trecho de código em PHP que implementa essa atividade de leitura e preparação. Como já foi mencionado, na base de dados os termos da estrutura classificatória foram tratados como substantivos. Isso foi definido porque, analisando gramaticalmente, os termos da estrutura classificatória denotam os substantivos.

Figura 51 – Implementação em PHP do cadastro dos termos da estrutura classificatória

cadastrarRelacaoSem.php | cadastraRelacao.php | gerar_mapa_tese.php | upload_arquivos.php | index.php

```
53
54
55      $f = file($arquivo);//arquivo referente à estrutura classificatória
56      $i=0;//variável criada para manipular o vetor (índice do vetor)
57      $cont=0;//conta a quantidade de termos que foram cadastrados
58      foreach($f as $item){ //percorre o vetor
59          $termos[$i] = trim($item);//retorna uma string com os espaços retirados do ínicio e do final
60          // Verifica se cada substantivo da estrutura existe na base de dados, se não existir, ele adiciona
            na tabela substantivo
61          $query = "SELECT NomeSubstantivo from tbSubstantivo where NomeSubstantivo = '".$termos[$i]."'" or
                die($mysqli->error);
62          if ($stmt = $mysqli->prepare($query))
63              if($stmt->execute()){
64                  $stmt->store_result();
65                  $palavra_check= "";
66                  $stmt->bind_result($palavra_check);
67                  $stmt->fetch();
68                  if ($stmt->num_rows != 1){//significa que é falso, ou seja, a palavra não existe no banco
                    de dados
69                      $sql = "INSERT INTO tbSubstantivo (NomeSubstantivo) VALUES ('".$termos[$i] ."')";
70                      if (mysqli_query($mysqli, $sql)== True)
71                          $cont++;
72                      else
73                          echo mysqli_error ($mysqli);
74                  }
75              }
76          $i = $i+1;
77      }
78
79      echo "<BR> <h2>Foram cadastrados " .$cont. " conceitos para a ".$tipoPub .": <i>".$tituloPub ."</i>. <
            br></h2>";
80
81
```

Fonte: Recorte do código-fonte do Semantizar, elaborado pela autora.

A outra tarefa da atividade de leitura e preparação é a preparação do arquivo referente à publicação (documento acadêmico: tese ou dissertação) para a manipulação que ocorrerá na próxima etapa do Modelo. Devido à linguagem de programação escolhida para a implementação, foi necessário converter o texto em formato .pdf para o formato .txt. Essa conversão foi necessária porque o manuseio de um arquivo de texto simples é menos complicado computacionalmente do que o de

um arquivo .pdf. Nessa conversão, o arquivo .txt resultante é temporário, ou seja, ele não é armazenado no servidor; logo, não há sobrecarga provocada pelo acúmulo de arquivos com o mesmo conteúdo. A conversão dos arquivos utilizou a classe PdfToText.phpclass[53]. A Figura 52 mostra a conversão do arquivo .pdf em .txt.

Figura 52 – Código com a conversão da publicação em formato PDF para texto simples

```
cadastrarRelacaoSem.php × | cadastraRelacao.php × | gerar_mapa_tese.php ● | upload_arquivos.php × | index.php ×

33
34      //*** conversão o pdf em txt ***//
35
36      include ( 'PdfToText.phpclass' ) ;//inclui a classe PdfToText
37      $file = $_SESSION['publicacao'];//a variável $file é configurada para receber a publicação inserida na
        etapa de leitura de dados
38      $pdf =  new PdfToText ("$file" ) ;//a publicação é convertida em um arquivo texto temporário
39      $publicacao = fopen("file.txt", "w+");// abertura do arquivo para leitura e escrita
40      fwrite($publicacao, "$pdf");//o conteúdo do pdf é inserido no arquivo .txt temporário
41
```

Fonte: Recorte do código-fonte do Semantizar, elaborado pela autora.

A atividade de *extração de relações semânticas* é considerada a mais importante e o cerne do Modelo de Extração de Relações Semânticas. Ela é composta de duas tarefas, sendo que na primeira realiza-se uma busca por pares de termos da estrutura classificatória em frases da publicação à qual a estrutura se refere. A Figura 53 mostra o código-fonte em PHP dessa fase. Como pode ser visto, o arquivo temporário criado na etapa anterior *file.txt* é transformado em uma *string* (variável que armazena caracteres alfanuméricos). Posteriormente, essa *string* é decomposta em uma *string* menor toda vez que o sinal de pontuação ponto-final (;) é encontrado no arquivo. Dessa forma, a variável *$frase*, que pode ser vista na linha 79 da Figura 53, pode ser tratada como um vetor em que cada posição é uma frase da publicação separada pelo ponto final. Em seguida, percorre-se o vetor *$frase* e verifica-se se existe na frase em questão o primeiro termo do vetor *$termo*. Esse vetor foi criado anteriormente, como apresentado na Figura 52. Caso exista, ele percorre as demais posições desse mesmo vetor *$termo* para verificar se existe outro termo da estrutura na mesma frase. Se existir, um formulário é criado para o usuário validar a relação semântica.

[53] Disponível em <https://github.com/christian-vigh-phpclasses/PdfToText/blob/master/PdfToText.phpclass>. Acesso em 02 ago. 2017.

Figura 53 – Busca por pares de termos nas frases da publicação

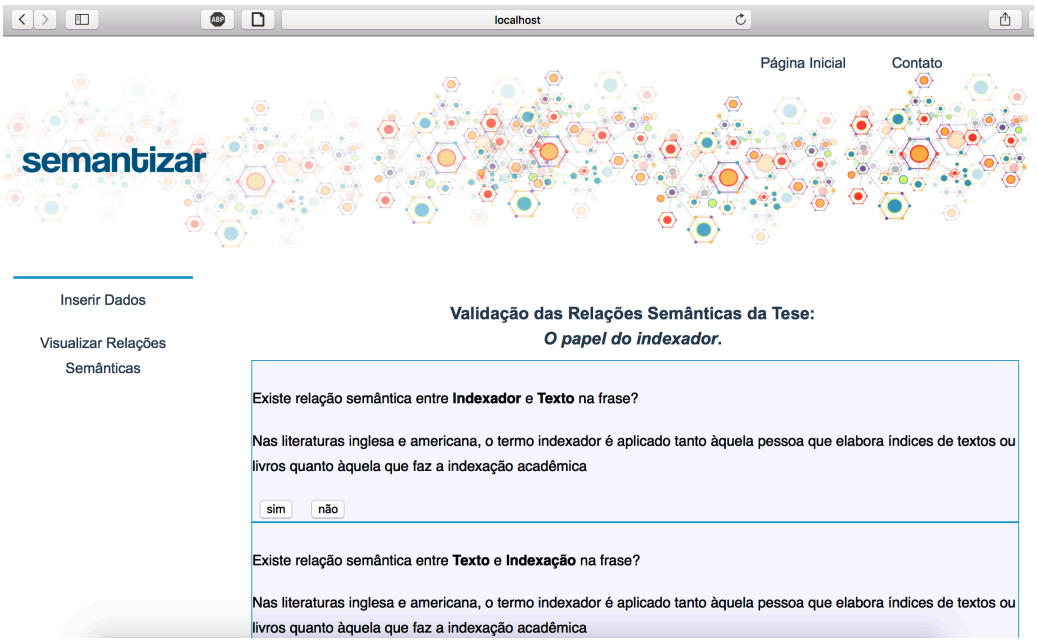
cadastrarRelacaoSem.php | cadastraRelacao.php | gerar_mapa_tese.php | upload_arquivos.php | index.php

```
78      $arquivo = file_get_contents("file.txt");//arquivo temporário da publicação, convertido anteriormente
79      $frase = explode(".", $arquivo); //decompõe a publicação em frases cujo separador é o ponto final. A variável
        $frase se transforma em um vetor e cada frase se transforma em uma posição do vetor.
80      for($z = 0; $z < sizeof($frase); $z++) {//percorre o vetor $frase
81          for($j=0; $j < (sizeof($f)-1); $j++){//percorre o vetor correspondente ao arquivo da estrutura classificatória
82              $termo1 = stripos($frase[$z], "$termo[$j]");//a variável $termo1 recebe valor verdadeiro ou falso retornado
                pela função do stripos.
83              if ($termo1 !== FALSE){//significa que encontrou o termo da estrutura na frase
84                  for($k=$j+1; $k<sizeof($f); $k++){//percorre o mesmo vetor de estrutura
85                      $termo2 = stripos($frase[$z], "$termo[$k]");//verifica se o segundo termo da estrutura existe na
                        frase
86                      if ($termo2 !== FALSE){//significa que encontrou o segundo termo na frase
87                          //cria um formulário para o usuário validar se existe a relação.
88                          echo "<FORM id='existe_relacao' action= '". $_SERVER['PHP_SELF']."' enctype='multipart/
                                form-data' method='post'>
89                           <BR> Existe relação semântica entre <b> " . $array_1[$j] . " </b> e <b>" .$array_1[$k] . "<
                                /b> na frase?  <p> <blockquote>" .$frase[$z] . " </blockquote></p>
90                          <INPUT TYPE='hidden' NAME='idPublicacao' VALUE= '".$idPublicacao."'>
91                          <INPUT TYPE='hidden' NAME='sujeito' VALUE= '".$array_1[$j]."'>
92                          <INPUT TYPE='hidden' NAME='objeto' VALUE= '".$array_1[$k]."'>
93                          <INPUT TYPE='hidden' NAME='frase' VALUE= '".$frase[$z]."'>
94                          <input type= 'submit' id = 'existeRelacao' name='existeRelacao' value = 'sim'>
95                          <input type= 'submit' id = 'naoExisteRelacao' name= 'existeRelacao' value = 'não'>
96                          <BR>
97                          </form>";
98
99                          break;
100                     }
101                 }
102             }
103         }
104     }
105
```

Fonte: Recorte do código-fonte do Semantizar, elaborado pela autora.

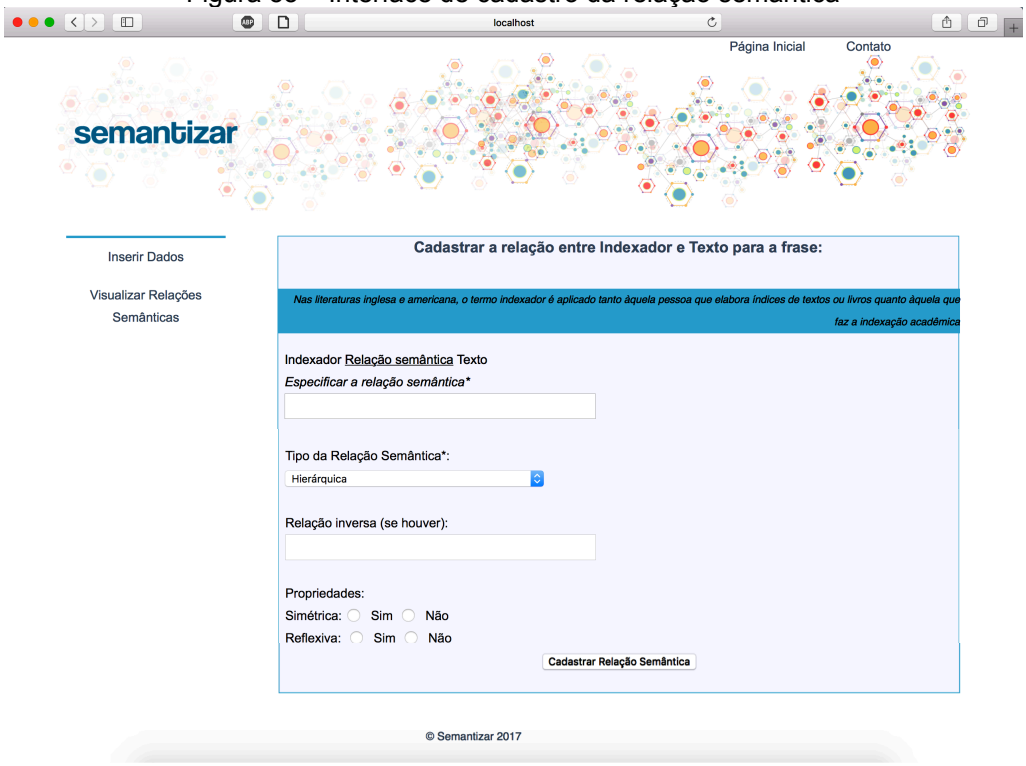
A validação da relação semântica é a segunda tarefa da atividade de extração de relações. A Figura 54 mostra a interface implementada para essa fase. Caso o usuário concorde que existe uma relação semântica entre os dois conceitos da estrutura classificatória encontrados em uma frase da publicação, ele é submetido a outro formulário, apresentado na Figura 55, em que a relação semântica é cadastrada.

Figura 54 – Interface de validação das relações semânticas



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

Figura 55 – Interface de cadastro da relação semântica



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

A interface de cadastro apresentada na Figura 55 é criada para cada indício de existência de relação semântica, considerando que, quando existem dois conceitos da estrutura classificatória na mesma frase, pode haver a ocorrência de uma relação entre esses conceitos. O usuário pode especificar a relação semântica conforme seu julgamento ao analisar a frase (por exemplo, *elabora índices de,* conforme a análise do trecho da publicação apresentado na Figura 55). Em seguida, o usuário especifica qual o tipo da relação semântica. Conforme mencionado na seção 5.2, os tipos de relações semânticas consideraram a taxonomia de relações semânticas elaborada e apresentada na seção 3.3.2. Posteriormente, o usuário informa a relação inversa, se houver. No caso do exemplo apresentado, a seguinte relação inversa só seria possível se o termo da estrutura classificatória fosse *índices de texto* e não *texto***.** Nesse caso, a tripla de relação semântica inversa seria: Índices de texto *são elaborados pelo* indexador. Logo, no campo relação inversa, o usuário informaria: “são elaborados pelo”. Finalizando o cadastro das relações semânticas, o usuário assinala as propriedades: simetria e reflexividade. Conforme mencionado na seção 3.3.3, a propriedade de transitividade não foi considerada por entender-se que ela se aplica a relações ternárias, o que não é o caso desta tese.

Essa funcionalidade do Semantizar para cadastrar as relações semânticas foi criada parcialmente. Dessa forma, as relações semânticas explicitadas nessa tese não serão armazenadas no banco de dados por questões de tempo para implementação e teste.

Do mesmo modo, a última atividade, de *representação do conhecimento,* não foi implementada até o momento. Nessa atividade, pretende-se criar uma visualização gráfica dos conceitos e as respectivas relações semânticas descobertas no contexto do documento acadêmico ao qual a estrutura classificatória representa. Obviamente nessa atividade, interfaces de consulta e de geração de relatórios serão criadas.

Como o Semantizar irá representar vários documentos, no futuro, um documento acadêmico de um domínio poderá ser ligado a outro documento acadêmico por meio das relações semânticas. Desse modo, além da visualização da representação de um documento o Semantizar irá gerar uma visualização do domínio no escopo dos documentos por ele tratados. Nesses estudos, deverão ser incorporadas pesquisas sobre *typed links, knowledge graph, linked data* e criação automática de *links*.

Com o estabelecimento dos tipos e subtipos e das propriedades das relações semânticas e das relações inversas que o Semantizar implementou, pretende-se criar, também na atividade de representação do conhecimento, um mecanismo para gerar arquivos com as relações semânticas que poderão ser exportadas para *softwares* de criação de ontologias, como o Protégé[54]. Da mesma forma, acredita-se que as triplas de relações semânticas armazenadas permitirão que elas sejam migradas para esquemas RDF (*Resource Description Framework*) e para ferramentas de elaboração de mapas conceituais, a exemplo do CmapTools[55].

No capitulo seguinte, apresenta-se um estudo de caso utilizando o Semantizar na extração e explicitação de relações semânticas e desse modo avaliar suas contribuições e dificuldades apresentadas.

---

[54] Disponível em <https://protege.stanford.edu>. Acesso em 03 ago. 2017.

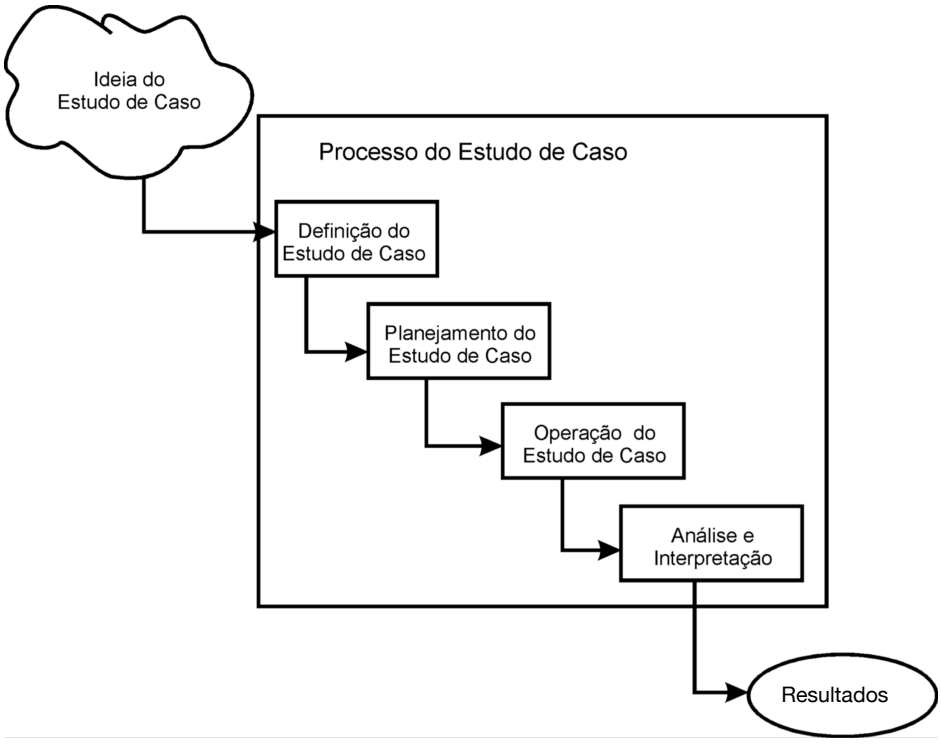
[55] Disponível em <https://cmap.ihmc.us/products/>. Acesso em 03 ago. 2017.

6 ESTUDO DE CASO: AVALIAÇÃO DA IMPLEMENTAÇÃO DO MODELO DE EXTRAÇÃO DE RELAÇÕES SEMÂNTICAS – SEMANTIZAR

Este capítulo apresenta o estudo de caso que examina as contribuições do Modelo de Extração de Relações Semânticas assim como sua implementação computacional, o Semantizar, elaborado nesta tese para facilitar a explicitação de relações semânticas a partir de estruturas classificatórias ou listas de termos de documentos acadêmicos.

A metodologia para a realização do estudo de caso está organizada de acordo com a proposta de Processo de Experimentação de Wohlin *et al.* (2000), porém, adaptado para esta tese. O processo está representado na Figura 56.

Figura 56 – Processo do estudo de caso



Fonte: Adaptada de Wohlin *et al.* (2000).

Na primeira etapa, *definição do estudo de caso***,** determinou-se conforme Wohlin *et al.* (2000), (a) o objeto do estudo de caso, qual seja, as relações semânticas; (b) o objetivo, que é verificar a eficiência do modelo de extração de relações semânticas implementado na aplicação *web* Semantizar; e (c) o contexto,

formado de teses e dissertações do domínio da Organização e Representação do Conhecimento.

Após a fase de definição do estudo de caso, segue a etapa de *planejamento.* De acordo com Wohlin *et al.* (2000), essa etapa indica basicamente "como" o estudo de caso será conduzido. Dessa forma, no planejamento, estabeleceu-se: (a) a amostra conforme está referida no Capítulo 2 e (b) as análises dos dados que serão coletados. Nesse caso, decidiu-se realizar análises quantitativa e qualitativa para identificar: (I) o número de relações semânticas sugerido pela aplicação em decorrência da quantidade de relações semânticas que realmente existem (esse fator é importante para sinalizar se a aplicação tem o potencial de extrair automaticamente relações semânticas); (II) os conceitos que têm mais probabilidade de serem relacionados semanticamente (esse parâmetro pode apontar os conceitos-chave da publicação analisada); (III) as características das relações semânticas encontradas; e (IV) o paralelo entre as relações entre os conceitos da estrutura classificatória e a representação resultante a partir do Semantizar.

Seguindo o processo do estudo de caso, a próxima etapa é a *operação*. Essa fase consiste da execução do estudo de caso definido e planejado anteriormente (WOHLIN *et al.*, 2000). Para isso, três procedimentos foram necessários. O primeiro, a preparação, envolveu a realização do recorte da estrutura classificatória e da publicação. Na estrutura classificatória, os termos ideia, pensamento e conceito foram desmembrados. Já em relação ao documento acadêmico, conforme mencionado no Capítulo 2, optou-se por considerar os capítulos 2, 3 e 4 da tese de Naves (2000). Essa escolha decorreu do fato de esses capítulos referirem-se à parte de definições conceituais da tese em questão. Decidiu-se também retirar as figuras dos capítulos, uma vez que o Semantizar não tem suporte para analisá-las. O segundo procedimento, a execução, compreendeu o processamento dos recortes da amostra no Semantizar no ambiente computacional em que foi implementado. Ressalta-se que, no momento, o Semantizar está hospedado em um servidor de acesso somente local. Por fim, o último procedimento foi o de validação, em que se realizou um refinamento dos dados levantados para evitar ruídos na análise e interpretação dos dados. Da mesma forma, na validação buscou-se compilar os dados para facilitar a compreensão durante a análise e interpretação dos dados. Tanto na validação dos dados quanto nos resultados, os

conceitos serão observados individualmente e com seus pares. A próxima seção trata desse procedimento.

6.1 Validação dos dados

A amostra executada no Semantizar totalizou 954 frases, nas quais foram detectados 199 indícios de relações semânticas entre conceitos, pertencentes à estrutura classificatória, em 131 frases. Isso corresponde a 14% das frases dos capítulos da tese utilizada na amostra. O gráfico da Figura 57 mostra a porcentagem de frases em que foram detectados indícios de relações semânticas entre os conceitos, assim como o contrário.

Figura 57 – Gráfico com a porcentagem de frases onde foram detectados pares de conceitos

Fonte: Elaborada pela autora.

Nas 131 frases do documento acadêmico em que ocorreram relações semânticas, observou-se que, por vezes, uma frase continha mais de um indício de relação semântica, conforme pode ser visto na Figura 58.

Figura 58 – Recorte do Semantizar mostra que uma frase pode ter mais de um indício de relação semântica

Fonte: Recorte do Semantizar. Elaborada pela autora.[56]

6.1.1 Validação dos indícios de relações semânticas

Conforme revela a Tabela 1, os 199 indícios de relações semânticas identificados pelo Semantizar foram encontrados em 48 pares de conceitos. Nesse contexto, entre um par de conceitos pode existir mais de uma relação semântica. Por exemplo: entre *indexador* e *documento* podem existir as relações “indexador *analisa* documento e indexador *agrupa* documento”. E, ainda, uma relação semântica pode ocorrer outras vezes em frases diferentes.

[56] Todas as passagens de texto nos recortes das interfaces referem-se a Naves (2000).

Tabela 1 – Indícios de Relações Semânticas entre os pares de conceitos

Pares de conceitos	**Quantidade de indícios**
Indexador e Documento	28
Conceito e Texto	18
Indexador e Texto	16
Conceito e Documento	14
Indexador e Conceito	12
Ideia e Texto	12
Documento e Texto	12
Conceito e Ideia	6
Autores e Texto	5
Texto e Superestrutura	5
Autores e Conceito	4
Autores e Documento	4
Indexador e Prática	4
Conceito e Pensamento	4
Autores e Indexador	3
Bibliotecário e Especialização	3
Indexador e Ideia	3
Ideia e Pensamento	3
Pensamento e Documento	3
Texto e Narrativos	3
Texto e Informativo	3
Profissional da informação e Indexador	2
Prática e Conceito	2
Prática e Texto	2
Texto e Primário	2
Texto e Macroestrutura	2
Narrativos e Informativo	2
Microestrutura e Macroestrutura	2
Autores e Prática	1

Pares de conceitos	Quantidade de indícios
Autores e Ideia	1
Profissional da Informação e Bibliotecário	1
Bibliotecário e Indexador	1
Bibliotecário e Prática	1
Bibliotecário e Ideia	1
Bibliotecário e Documento	1
Indexador e Pensamento	1
Especialização e Prática	1
Especialização e Texto	1
Conceito e Secundário	1
Ideia e Documento	1
Documento e Informativo	1
Documento e Secundário	1
Texto e Secundário	1
Texto e Hipertexto	1
Texto e Microestrutura	1
Primário e Secundário	1
Primário e Microestrutura	1
Macroestrutura e Superestrutura	1
Total	**199**

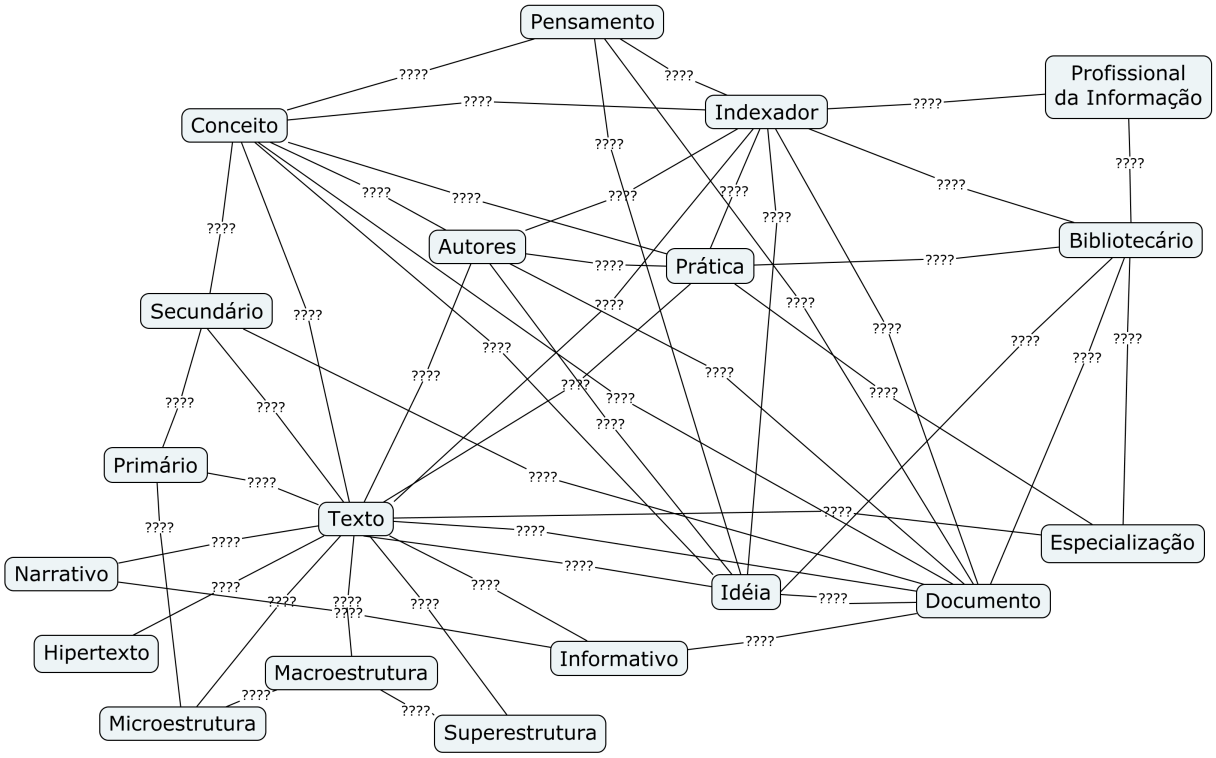
Fonte: Elaborada pela autora.

Prosseguindo com a validação, criou-se um mapa conceitual com os indícios de relações entre os pares de conceitos. Esse mapa é apresentado na Figura 59. Em seguida, verificou-se se cada indício de relação semântica apresentado nesse mapa ocorria de fato, ou seja, se os dois conceitos da amostra encontrados em uma frase relacionavam-se semanticamente no contexto em que eles foram identificados. Assim, os indícios de relações semânticas foram classificados em verdadeiros ou falsos[57]. Desse modo, os indícios de relações verdadeiros totalizaram 105, enquanto os falsos

[57] Cabe destacar a subjetividade da classificação dos indícios em alguns casos.

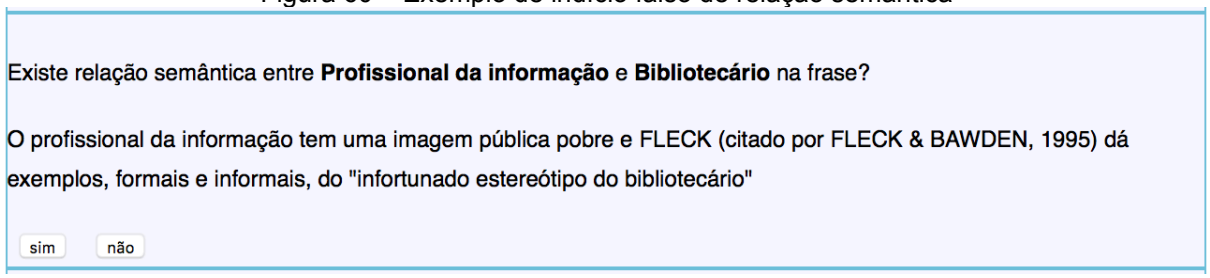
somaram 94. Um exemplo de indício de relação semântica falso está apresentado na Figura 60. Esse exemplo e todos os outros indícios falsos estão no Apêndice B.

Figura 59 – Mapa conceitual com os indícios de pares de conceitos que se relacionam semanticamente



Fonte: Elaborada pela autora.

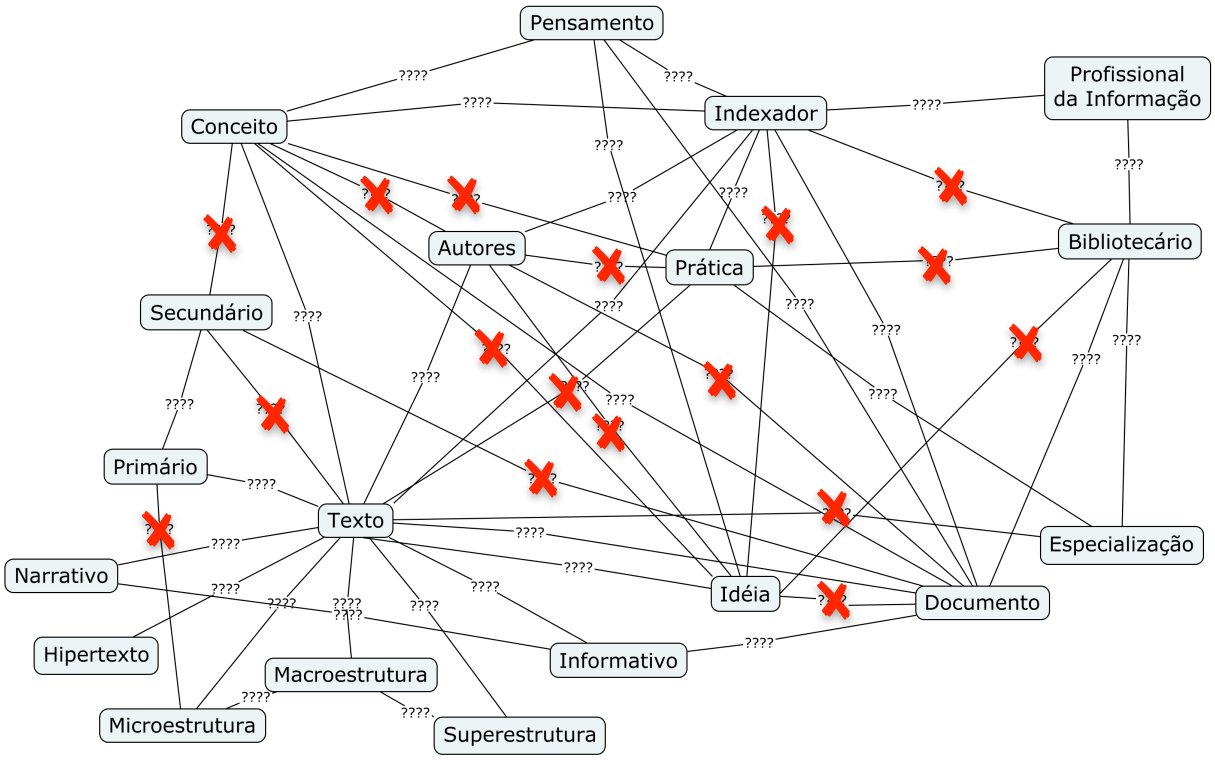
Figura 60 – Exemplo de indício falso de relação semântica



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

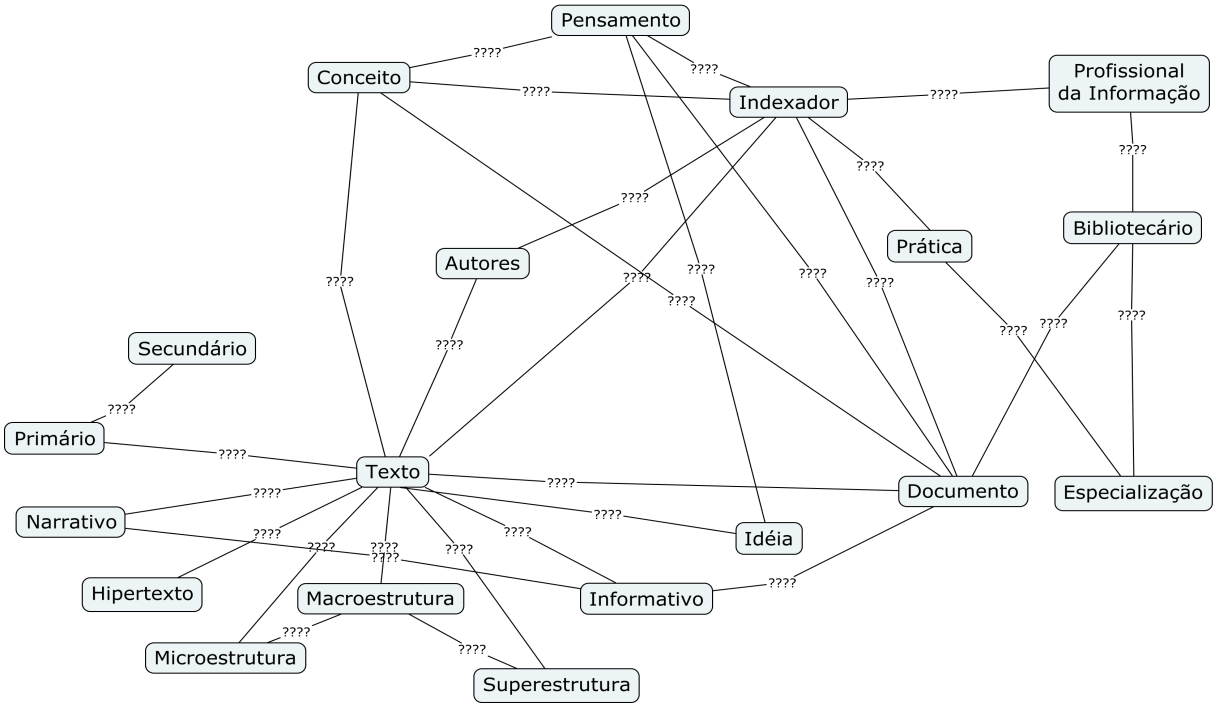
Na Figura 61, um mapa conceitual revela os 17 pares de conceitos em que, durante a validação, verificou-se que não existem relações semânticas entre eles. Adiante, na Figura 62, outro mapa conceitual evidencia somente os pares de conceitos com os indícios de relações semânticas verdadeiros. Esses pares totalizam 31.

Figura 61 – Mapa conceitual mostrando indícios falsos de relações semânticas



Fonte: Elaborada pela autora.

Figura 62– Mapa conceitual com os indícios verdadeiros de relações semânticas



Fonte: Elaborada pela autora.

6.1.2 Validação dos conceitos individualmente a partir dos indícios verdadeiros

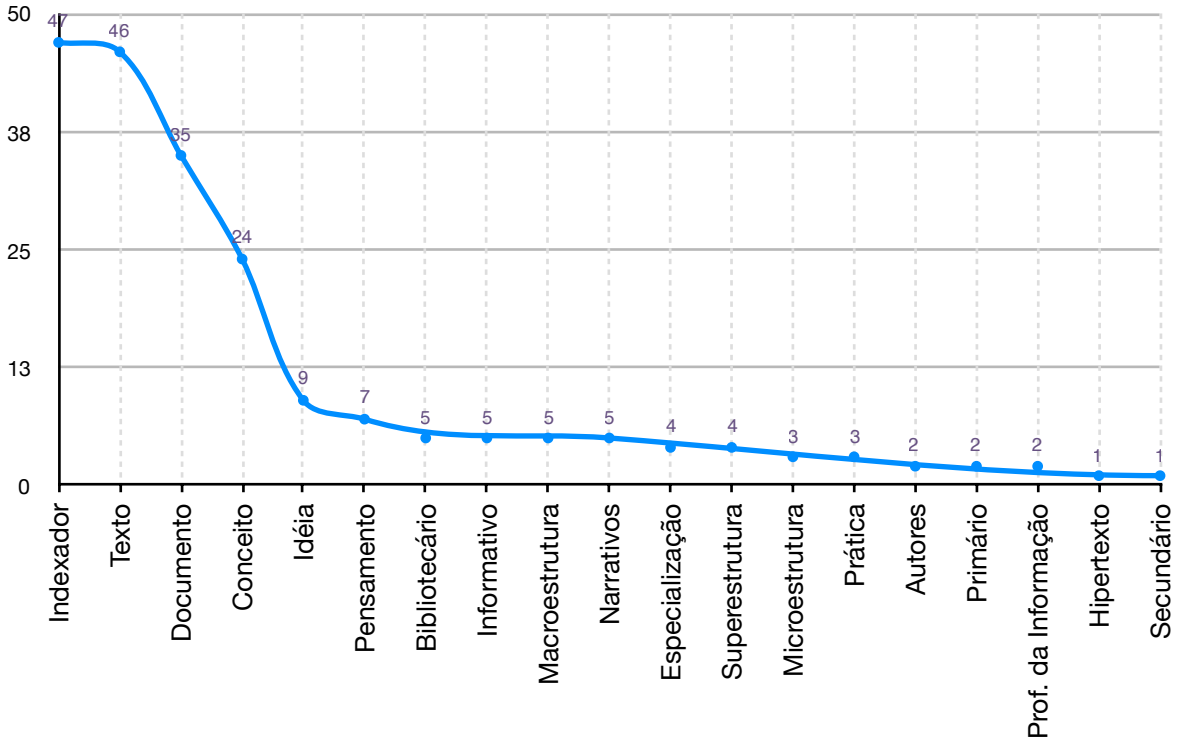
Para a validação dos conceitos, elaborou-se uma matriz (ver Tabela 2) com os conceitos e suas intercessões com seus pares. Com base nessa matriz, contabilizou-se os indícios de relações semânticas por conceito, conforme pode ser visto no gráfico da Figura 63. Para encontrar os valores, considerou-se a soma do total de ocorrências do conceito horizontalmente e verticalmente na matriz. Por exemplo: o conceito *texto* tem horizontalmente 32 indícios de ocorrências de relações semânticas e verticalmente 14, conforme destacado. Logo, o valor resultante dos indícios de relações semânticas com o conceito *texto* é 46.

Tabela 2 – Matriz de conceitos de indícios de relações semânticas verdadeiros

	Bibliotecário	Indexador	Especialização	Conceito	Idéia	Documento	Texto	Informativo	Primário	Macroestrutura	Total
Autores		1					1				**2**
Profissional da Informação	1	1									**2**
Especialização	3										**3**
Prática		2	1								**3**
Conceito		5									**5**
Pensamento		1		3	2	1					**7**
Documento	1	23		7				1			**32**
Texto		14		9	7	2					**32**
Narrativo							3	2			**5**
Informativo							2				**2**
Primário							1				**1**
Microestrutura							1			2	**3**
Secundário									1		**1**
Macroestrutura							2				**2**
Superestrutura							3			1	**4**
Hipertexto							1				**1**
Soma	**5**	**47**	**1**	**19**	**9**	**3**	**14**	**3**	**1**	**3**	**105**

Fonte: Elaborada pela autora.

Figura 63 – Gráfico com a quantidade de indícios verdadeiros por conceito



Fonte: Elaborada pela autora.

Foi observado durante a validação individual que os conceitos *indexador experiente*, *indexador pouco experiente* e *indexador novato* não foram detectados pelo Semantizar. Assim, dos 22 conceitos da amostra, 19 tiveram alguma relação semântica com outro.

A validação dos dados reduziu os resultados para concentrar apenas nos indícios verdadeiros de relações semânticas. Dessa forma, como mostra a Tabela 3, inicialmente existiam 22 conceitos na amostra que compuseram 48 pares de conceitos em 199 indícios de relações semânticas. Após a validação, permaneceram 19 conceitos organizados em 31 pares em 105 indícios verdadeiros de relações semânticas.

Tabela 3 – Valores atualizados dos dados que serão analisados

Dados	Quantidade de conceitos	Quantidade de pares de conceitos	Quantidade de Indícios
Início	22	48	199
Analisados como falsos	3	17	94
Analisados como verdadeiros	19	31	105

Fonte: Elaborada pela autora.

6.2 Estabelecimento das relações semânticas

Ainda na etapa de Operação do Estudo de Caso, ocorreu o *Estabelecimento das Relações Semânticas*. Nessa tarefa, realizou-se a explicitação das relações semânticas entre os conceitos.

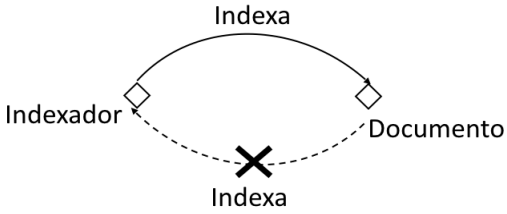
Conforme relatado anteriormente, existem 105 indícios de relações semânticas classificados como verdadeiros. Assim, notou-se que esse número é expressivo para uma descrição detalhada de cada relação semântica explicitada. Nesse sentido, optou-se por especificar 10 relações semânticas entre os conceitos *indexador* e *documento* que, na amostra, demonstrou ser o par de conceitos com mais indícios de relações semânticas. As demais relações semânticas que puderam ser explicitadas estão descritas no Apêndice C.

- *Caso 1* – Nesse primeiro caso, a seguinte frase foi detectada: “No caso do processamento técnico do acervo de *documentos*, o profissional responsável pela catalogação, classificação e indexação, costuma ter a formação bibliotecária, e receber o nome de *indexador*.” (NAVES, 2000, p. 16, grifo nosso)[58]. Percebe-se, especificamente nessa frase detectada, que existem não uma, mas três relações semânticas entre *documento* e *indexador*. São

[58] Respeitou-se nesta tese, o texto original de Naves (2000). Sendo assim, problemas textuais que por acaso ocorram nesse texto, serão ignorados.

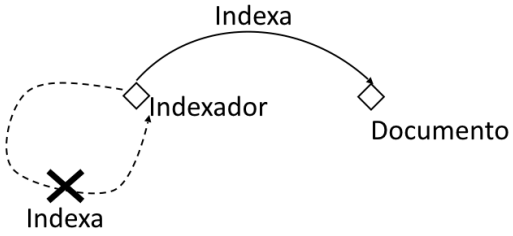
elas: indexador *cataloga* documento[59] indexador *classifica* documento e indexador *indexa* documento. Esses três casos são relações associativas do tipo causal e subtipo agente-objeto, em que o agente é o *indexador* e o objeto é o *documento* em que o agente realiza sua ação. Com respeito às propriedades de simetria e reflexividade, as relações apresentadas nesse caso são assimétricas e irreflexivas, conforme mostram as Figuras 64 e 65. As relações inversas são: documento *é catalogado por* indexador, documento *é classificado por* indexador e, documento *é indexado por* indexador. Esta última está representada na Figura 66.

Figura 64 – Representação da assimetria da relação entre indexador e documento



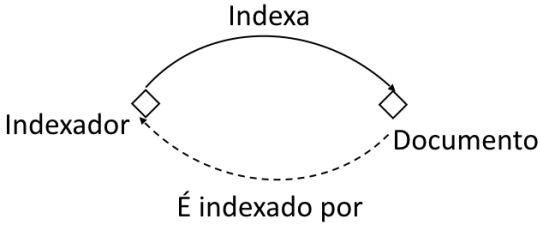
Fonte: Elaborada pela autora.

Figura 65 – Representação da irreflexividade da relação entre indexador e documento



Fonte: Elaborada pela autora.

Figura 66 – Representação da relação inversa entre indexador e documento



Fonte: Elaborada pela autora.

[59] Optou-se por tratar alguns conceitos das relações semânticas no singular.

- *Caso 2* – Na frase “Já para o segundo autor, Análise de assunto é um processo muito subjetivo e um aspecto importante a considerar é a visão, do *indexador*, de qual assunto trata um *documento*” (NAVES, 2000, p. 19, grifos nossos), percebe-se que a relação entre *indexador* e *documento* não é direta e ainda não é o principal assunto da frase. Contudo, determinou-se que a relação semântica nesse caso é: indexador *determina assunto do* documento. Assim como no Caso 1, a relação é associativa do tipo causal e subtipo agente-objeto, cujas características são assimétrica e irrefletiva. A relação inversa, nesse caso, seria: documento é determinado assunto pelo indexador. Contudo, como pode ser visto, essa relação não é possível, pois não se trata simplesmente do documento e sim do assunto do documento. Logo, o mais correto seria: o assunto do documento *é determinado pelo* indexador. Como o conceito da amostra é *documento* e não *assunto do documento*, resolveu-se não estabelecer a relação inversa.
- *Caso 3* – Nesse caso, na frase: “Um dos estudos é o de BLAIR (1986), já mencionado anteriormente, que se refere à inconsistência inter *indexadores* ao determinar assuntos de *documentos*” (NAVES, 2000, p. 32, grifos nossos), a relação semântica detectada entre indexador e documento é a mesma do caso anterior: indexador *determina assunto do* documento, e, por conseguinte, tem as mesmas classificações e características.
- *Caso 4* – “O trabalho de DAVID et al. (1995), também já mencionado anteriormente, registra o estudo de quatro *indexadores* experientes, que foram solicitados a analisar, cada um, dois *documentos*” (NAVES, 2000, p. 33, grifos nossos). Nessa frase, a relação semântica é indexador *analisa* documento. Ela é associativa do tipo causal e subtipo agente-objeto. Assim como nas propriedades das relações semânticas anteriores, essa relação é assimétrica e irreflexiva. A relação inversa nesse caso é documento *é analisado por* indexador. Nesse caso, apesar dos quantificadores “quatro indexadores” e “dois documentos”, existe a mensagem do autor que “cada um” ou seja, cada indexador analisa documentos, que podem ser dois, como está descrito, mas pode ser um documento se colocado em outro contexto. Assim, julgou-se que essa relação semântica procede.

- *Caso 5* – Na frase “A observação do trabalho do *indexador*, mais precisamente da leitura que esse faz para análise, síntese e representação do conteúdo de um *documento* […]” (NAVES, 2000, p. 33, grifos nossos), assim como no Caso 1, observa-se que existem três relações semânticas entre indexador e documento: indexador *analisa o conteúdo do* documento, indexador *sintetiza o conteúdo do* documento e indexador *representa o conteúdo do* documento. Como nos casos anteriores, essas relações são associativas do tipo causal e subtipo agente-objeto. Da mesma forma, elas são assimétricas e irreflexivas. Como no Caso 2, a relação inversa não é possível de ser estabelecida, pois não se trata simplesmente do documento e sim do conteúdo do documento.

Como todas as relações entre *indexador* e *documento* analisadas até o momento têm a mesma classificação e as mesmas propriedades de simetria e reflexividade, optou-se por suprimir essa informação. Caso durante a análise ocorra alterações nessas informações, elas serão devidamente apresentadas.

- *Caso 6* – “Resultados indicaram que os *indexadores* utilizam estratégias metacognitivas de leitura, realizam associação com a linguagem documentária do sistema durante a leitura e conhecem a estrutura textual dos *documentos* de Odontologia” (NAVES, 2000, p. 34, grifos nossos). Relação semântica: indexador *conhece a estrutura textual do* documento. Assim como nos Casos 2 e 5, não é possível estabelecer a relação inversa.
- *Caso 7* – “E com relação ao leitor/*indexador*? Sabe-se que um *documento*, inserido num Sistema de Recuperação da Informação, antes de ser lido pelo leitor, usuário final do sistema, é lido por um leitor técnico, o *indexador*, aquele que faz a leitura para fins documentários” (NAVES, 2000, p. 53, grifos nossos). Nesse caso, a relação semântica é: indexador *faz leitura documentária do* documento. Naves (2000) destacou nessa frase, que a leitura realizada pelo indexador é uma leitura técnica, por isso, na relação semântica utilizou-se o termo “leitura documentária”. A relação inversa não foi possível.

- *Caso 8* – “Para eles, não é clara a noção de ‘trata de’ e sugerem o procedimento abaixo: *indexadores* rotulam e agrupam *documentos* de acordo com o conteúdo desses; pesquisadores formulam suas questões de acordo com o conteúdo dos documentos; questões relativas à pesquisa são cotejadas com o que os documentos do sistema oferecem” (NAVES, 2000, p. 65, grifos nossos). Nesse caso, duas relações semânticas foram identificadas: indexador *rotula* documento e indexador *agrupa* documento. As relações inversas são: documento *é rotulado pelo* indexador e, documento *é agrupado pelo* indexador.
- *Caso 9* – “O procedimento geral é que o *indexador* tome uma cadeia de pontos de vista claramente definidos, em círculo (por exemplo, orientação teórica, método de pesquisa) e gere uma lista de palavras-chave para cada *documento*, cada frase tomando como seu próprio foco um aspecto diferente do mesmo, mas referindo-se ao todo” (NAVES, 2000, p. 65, grifos nossos). Nesse caso, a relação semântica é: indexador *gera palavras-chave para* documento. Como em casos anteriores (2, 5, 6 e 7), a relação entre indexador e documento não é direta; logo, a relação inversa não é possível.
- *Caso 10* – “O artigo de BEGHTOL (1986) […] levanta algumas das implicações do trabalho do lingüísta Van Dijk […] para a teoria da classificação bibliográfica, o que pode ser transposto para a Análise de assunto, por envolver também a atividade intelectual, pelo *indexador*, de interpretação do conteúdo temático do *documento*” (NAVES, 2000, p. 70, grifos nossos). Assim como nos casos 1 e 5, a relação semântica entre *indexador* e *documento* é denotada por um substantivo: *interpretação,* oriundo de um verbo *interpretar.* Desse modo, a relação semântica nesse caso é indexador *interpreta o conteúdo temático do* documento. E, da mesma forma que os casos 2, 5, 6, 7 e 9, a relação inversa não é possível.

Durante a análise para o estabelecimento das relações semânticas, percebeu-se que algumas delas exigiam um esforço maior, pois sua interpretação não era trivial. Essas relações foram consideradas complexas. A próxima subseção trata desses casos.

6.2.1 Determinação de relações complexas

Durante a determinação das relações semânticas, observou-se que algumas relações exigiam uma interpretação mais criteriosa no contexto em que os pares de conceitos foram encontrados. Desse modo, dos 105 indícios de relações semânticas, 25 foram considerados complexos, o que corresponde a 23,8% dos indícios verdadeiros. Nesta seção, alguns desses casos serão tratados. Eles foram selecionados considerando a particularidade de suas complexidades.

Nesse primeiro caso, o Semantizar detectou os conceitos *pensamento* e *ideia* na frase: “Segundo o dicionário, *pensamento* é o ato ou efeito de pensar, refletir, meditar; processo mental que se concentra nas *idéias*; poder de formular *idéias*; atividade psíquica que abarca os fenômenos cognitivos, distinguindo-se do sentimento e da vontade” (AURÉLIO apud NAVES, 2000, p. 85, grifos nossos). Nessa sentença, existem duas relações semânticas entre os conceitos *pensamento* e *ideia*. Contudo, julgou-se que somente a segunda relação pode ser simplificada em uma relação semântica. Dessa forma, a relação semântica é: pensamento *formula* ideias. A relação inversa é: ideias *são formuladas pelo* pensamento. Entende-se que essa relação é associativa, do tipo ação subordinada. Com relação às propriedades, acredita-se que essa relação é assimétrica e reflexiva.

Em um outro momento, os conceitos *pensamento* e *ideia* foram encontrados pelo Semantizar:

> É necessário incluir considerações de como o *pensamento* é produzido, isto é, de que maneira estímulos externos ou internos resultam em atividade mental ou registro mental, e como os pensamentos são retidos na memória, ou dela desaparecem, e podem ser diferenciados de outros *pensamentos* (percepção, conceitos, *idéias*). (NAVES, 2000, p. 86, grifos nossos)

Diferentemente do caso anterior, no contexto em que os conceitos estão apresentados, a relação entre eles é de sinônimo. Como já mencionado no Capítulo 3, conforme Arnold e Rahm (2014), os sinônimos algumas vezes podem aparecer entre parênteses em uma frase. Logo, essa relação é de equivalência, do tipo sinônimo e subtipo sinônimo parcial, pois, no contexto abordado pela autora, ideia é de alguma forma um tipo de pensamento, e pensamento é um conceito mais amplo que ideia. A relação semântica explicitada é: ideia *é um tipo de* pensamento. Já a relação inversa

não é possível de ser estabelecida. A relação semântica encontrada é assimétrica – pois não se pode afirmar que pensamento é um tipo de ideia – e irreflexiva – porque não se pode afirmar que a relação entre ideia e pensamento é diferente da relação entre a ideia e ela mesma.

Esse próximo caso é uma situação considerada mais comum nas frases observadas. Por exemplo, o Semantizar detectou os conceitos *bibliotecário* e *documento*: “Para as autoras, normalmente são duas as situações nas quais os *bibliotecários* fazem Análise de assunto: (a) quando recebem um *documento* e devem dar entrada deste num sistema de informações […]” (NAVES, 2000, p. 35, grifos nossos). Nesse exemplo, a relação semântica é: bibliotecário *faz análise de assunto do* documento. Contudo, isso está implícito na frase, ou seja, na mensagem que ela transmite isso pode ser inferido. Nesse caso, a relação inversa não é possível, pois a análise de assunto do documento é feita pelo bibliotecário e não o documento faz a análise de assunto. Essa relação é do tipo associativa, causal, agente-objeto. Outrossim, ela é assimétrica e irreflexiva.

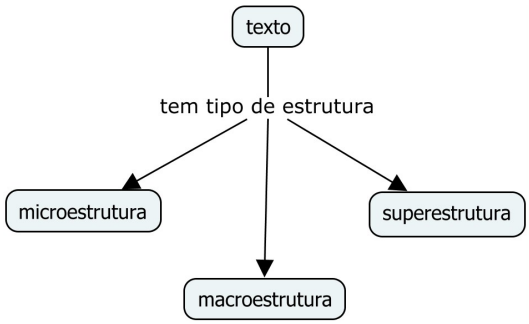
As três situações que seguem dizem respeito ao conceito *texto* e algumas classificações feitas pela autora sobre ele. No primeiro momento, o Semantizar encontrou os conceitos *texto* e *microestrutura*, conforme pode ser visualizado na frase:

> O sentido geral do *texto* é baseado na seguinte trilogia estrutural: *microestrutura* (estrutura superficial, que corresponde à realidade física do texto e a seus símbolos de significação, as palavras), macroestrutura (concebida como um tópico representativo hierárquico e coerente da unidade textual, envolvendo mínima estrutura da representação textual sintática-semântica) e superestrutura (estrutura retórica-esquemática, um tipo de esquema de produção convencional para o qual o texto é adaptado, podendo ser considerado como transição entre estruturas de superfície e de profundidade). (NAVES, 2000, p. 43, grifos nossos)

Contudo, como pode ser visto nessa frase outros conceitos presentes na amostra que podem se relacionar com *texto*. Analisando a passagem, entende-se que microestrutura, macroestrutura e superestrutura são tipos de estruturas de texto. Logo, as relações semânticas são: texto *tem tipo de estrutura* microestrutura, texto *tem tipo de estrutura* macroestrutura e texto *tem tipo de estrutura* superestrutura. A Figura 67 mostra essas relações a partir do conceito *texto*. Essas relações são associativas do tipo conceito-propriedade, assimétricas e reflexivas. As relações

inversas são: microestrutura *é um tipo de estrutura de* texto, macroestrutura *é um tipo de estrutura de* texto e superestrutura *é um tipo de estrutura de* texto.

Figura 67 – Relação entre texto e os tipos de estruturas textuais



Fonte: Elaborada pela autora.

A relação entre *texto* e *superestrutura* foi detectada em outros três momentos, quais sejam, em (a) “Para CINTRA (1987), os constituintes básicos de um determinado tipo de *texto* é que definem a sua *superestrutura*” (NAVES, 2000, p. 43, grifos nossos); em (b) “Segundo Van Dijk (citado por KOBASHI, 1996), a *superestrutura* é considerada um elemento fundamental para a apreensão do significado do *texto*” (NAVES, 2000, p. 43, grifos nossos); e em (c) “A *superestrutura* é, por definição, uma estrutura convencional, uma organização paradigmática, e o processo de compreensão supõe a transferência das unidades semânticas identificadas no *texto* para esse esqueleto conceitual” (NAVES, 2000, p. 43, grifos nossos). Nessas três situações, as relações semânticas existem; contudo, se explicitadas, não expressam a mensagem da autora. Observa-se que as relações entre os conceitos não são diretas, elas possuem outros elementos que afetam as relações entre *texto* e *superestrutura*.

Já entre *texto* e *macroestrutura*, a relação semântica foi determinada na frase: “Diante disso, pode-se afirmar que, para a Análise de assunto de textos, interessa mais o estudo da coerência textual, já que se lida com a *macroestrutura* do *texto* e a estruturação do sentido” (NAVES, 2000, p. 46, grifos nossos). Nesse caso, a relação semântica foi determinada pela preposição “de”. Conforme assinalado por Arnold e Rahm (2014), essa classe de palavras pode indicar a relação de merônimo-holônimo. Contudo, entendeu-se que, nesse caso, a relação “texto *tem atributo* macroestutura” é

do tipo associativa e subtipo atributo convidado. Essa relação foi classificada como assimétrica e irreflexiva. Já a relação inversa é: macroestrutura *é atributo de* texto.

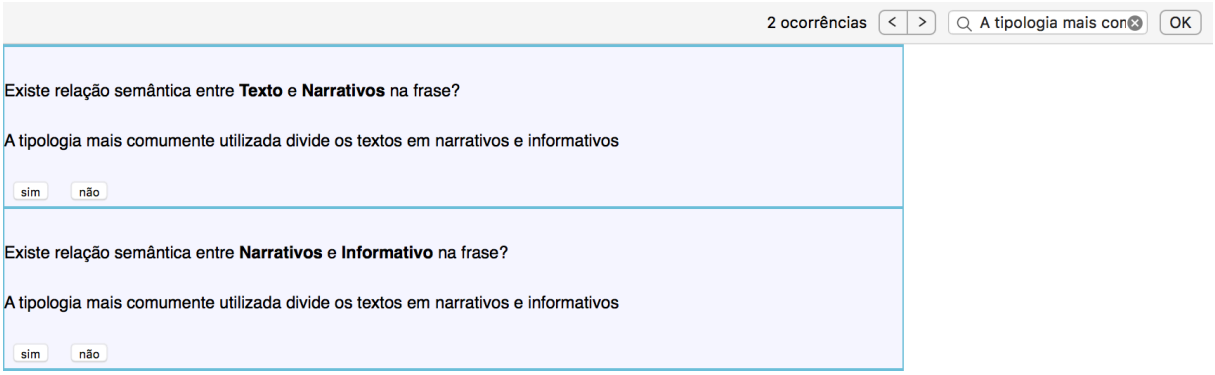
Ao analisar a frase em que as relações semânticas da Figura 68 foram identificadas e ao observar tal figura, compreende-se que existe uma relação semântica entre os conceitos considerados “irmãos”, ou seja, que estão no mesmo nível em uma hierarquia (por mais que essa relação não seja hierárquica): *microestrutura, macroestrutura* e *superestrutura*. No caso de conceitos “irmãos”, conforme mencionado no Capítulo 3, pode-se afirmar que existe alguma conexão semântica entre *microestrutura* e *macroestrutura*, *microestrutura* e *superestrutura* e *macroestrutura* e *superestrutura*, pois a relação entre esses conceitos e o conceito *texto* é a mesma. Desses três pares de relações semânticas, o Semantizar não detectou relação entre *microestrutura* e *superestrutura* nessa frase. Essa falha do Semantizar será explicada na seção 6.5.

No caso de conceitos “irmãos”, infere-se que as relações são associativas. Analisando o contexto, entende-se que as relações são desmistificadas nos subtipos: contraste – de oposição – assimetricamente contrárias. Nessas situações, os conceitos não são antônimos, mas são contrários em alguma proporção. Portanto, tais relações semânticas podem ser explicitadas de três modos. (1) Microestrutura *é assimetricamente contrária a* macroestrutura. Isso pode ser corroborado na frase em que o Semantizar detectou esses conceitos: “A estrutura semântica é caracterizada em dois níveis: *microestrutura* (nível primário, no qual se designa, por um lado, a estrutura das proposições individuais e, por outro lado, as relações entre as seqüências das frases no texto) e a *macroestrutura* (que representa as relações entre os grupos de frases, ou a organização geral do texto)” (NAVES, 2000, p. 83, grifos nossos). (2) Microestrutura *é assimetricamente contrária a* superestrutura. E (3) macroestrutura *é assimetricamente contrária a* superestrutura. As relações inversas são possíveis por meio da propriedade de simetria, ou seja, todas as relações são simétricas e essa simetria pode ser vista como relação inversa. Com relação à reflexividade, as relações são irreflexivas.

A segunda ocorrência correlata ao caso acima foi constatada na frase: “A tipologia mais comumente utilizada divide os *textos* em *narrativos* e *informativos*” (NAVES, 2000, p. 44, grifos nossos). Como pode ser verificado na Figura 68, o Semantizar localizou essa frase em dois momentos, no indício de relação entre

texto e *narrativo* e *narrativo* e *informativo*. Contudo, nessa frase ainda pode existir a relação entre *texto* e *informativo*.

Figura 68 – Recorte da interface do Semantizar que mostra os indícios de relações entre *texto* e *narrativos* e *narrativos* e *informativo*



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

Nesse último caso, entre *texto* e *informativo*, os indícios ocorreram nas frases: (1) “Para os *textos informativos*, a classificação mais conhecida, segundo GIASSON (1993) é a de Meyer, e compreende: descrição (dá informações de um sujeito e especifica alguns de seus atributos e características), […]” (NAVES, 2000, p. 44, grifos nossos) e (2) “Dentre os *textos informativos*, pode-se reconhecer o texto científico e, nesse tipo de texto, o conteúdo é quase inteiramente determinado pelo autor; […]” (NAVES, 2000, p. 44, grifos nossos).

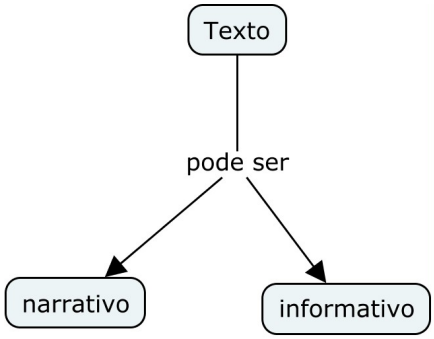
Da mesma forma como na situação descrita acima sobre *texto* e *informativo*, outras situações de indícios de relações semânticas entre *texto* e *narrativo* também ocorreram nas frases: (1) “Os *textos narrativos* se caracterizam pela marcação temporal cronológica […]” (NAVES, 2000, p. 44, grifos nossos) e (2) “Essa informação importante pode variar segundo os tipos de texto: nos *textos narrativos*, a idéia principal tem a ver com os acontecimentos e a sua interpretação […]” (NAVES, 2000, p. 61, grifos nossos).

Diferentemente dos casos de relações semânticas analisados até o momento, que foram estabelecidos por verbos, substantivos e preposições, nessas duas situações que seguem, as relações consideraram que a composição dos substantivos *narrativos* e *informativo* foi suficiente para indicar as relações semânticas entre eles e *texto*. Logo, foram definidas as seguintes relações: texto *pode ser* narrativo e texto *pode ser* informativo. Essas relações são do tipo hierárquicas, cujo subtipo pode ser detalhado como hipônimo-hiperônimo – inclusão de classe – taxonômica. Elas são

simétricas e reflexivas. As relações inversas podem ser: narrativo *é um tipo de* texto e informativo *é um tipo de* texto.

Como observa-se na Figura 69, *narrativo* e *informativo* são conceitos irmãos na hierarquia apresentada. Nesse caso, a relação é associativa, de contraste e assimetricamente contrária. Logo, essa relação pode ser explicitada como: narrativo *é assimetricamente contrário a* informativo. Isso é corroborado na frase em que esses dois conceitos foram encontrados: “Essa informação importante pode variar segundo os tipos de texto: nos textos *narrativos*, a idéia principal tem a ver com os acontecimentos e a sua interpretação, enquanto nos textos *informativos* o que é importante pode ser um conceito, uma generalização, uma regra” (NAVES, 2000, p. 61, grifos nossos). Essa relação segue todas as características das relações entre microestrutura, macroestrutura e superestrutura, detalhadas anteriormente.

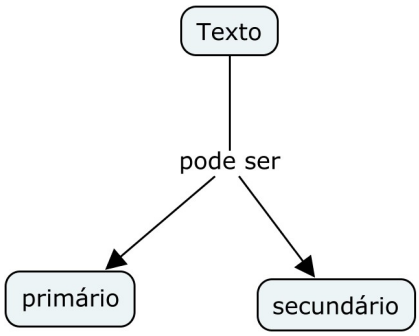
Figura 69 – Relação entre texto e seus tipos



Fonte: Elaborada pela autora.

A outra situação como a do exemplo acima pode ser encontrada na frase: “O texto original é chamado texto primário, um sumário ou resumo, texto secundário, e a expressão do texto primário numa linguagem documentária, texto terciário” (NAVES, 2000, p. 48, grifos nossos). Nesse caso, o Semantizar detectou indícios de relação entre os conceitos *primário* e *secundário*, e entre *texto* e *primário*. Contudo, ele não relacionou os conceitos *texto* e *secundário*, cuja relação pode ser observada claramente na frase. Na Figura 70, essas relações estão representadas. As relações entre *texto* e *primário* e *texto* e *secundário* também são taxonômicas e a relação entre *primário* e *secundário* também é assimetricamente contrária. As propriedades das relações semânticas desses pares de conceitos são as mesmas do caso anterior.

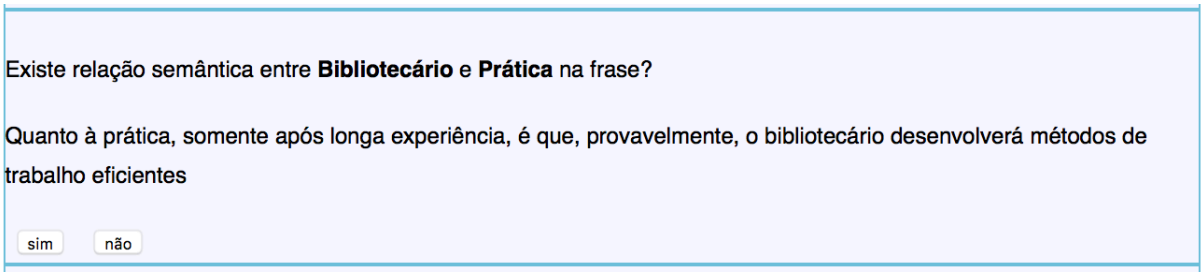
Figura 70 – Relação entre os conceitos *texto*, *primário* e *secundário*



Fonte: Elaborada pela autora.

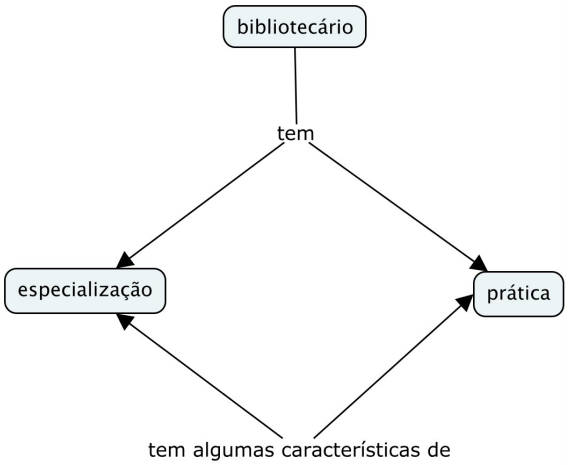
Por fim, ao analisar a relação entre *bibliotecário*, *especialização* e *prática*, percebeu-se que esses três conceitos podem se relacionar, como nos três casos acima relacionados ao *texto*. Isso foi percebido na frase: "A *especialização* e a *prática* do *bibliotecário* são tratadas por INGWERSEN (1982)" (NAVES, 2000, p. 22, grifos nossos). Essa frase foi detectada para os pares de conceitos *bibliotecário* e *especialização* e *especialização* e *prática*. A relação entre *bibliotecário* e *prática* foi identificada pelo Semantizar. Contudo, o indício no contexto verificado foi falso, conforme pode ser visto na Figura 71. Logo, a relação entre *bibliotecário* e *prática* foi determinada pela análise da frase do início deste parágrafo. A Figura 72 mostra as três relações: bibliotecário *tem* especialização, bibliotecário *tem* prática e especialização *tem algumas características de* prática.

Figura 71 – Indício falso de relação semântica entre *bibliotecário* e *prática*



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

Figura 72 – Relações entre *bibliotecário*, *especialização* e *prática*



Fonte: Elaborada pela autora.

As relações entre *bibliotecário* e *prática* e entre *bibliotecário* e *especialização* foram percebidas por meio da preposição “de”. A situação ocorreu na frase: “Segundo ele, a *especialização* de *bibliotecários* poderia ser melhorada” (NAVES, 2000, p. 22, grifos nossos). Essas relações foram classificadas como associativas e subtipo atributo convidado; elas são assimétricas e irreflexivas. Considerou-se não ser possível a relação inversa entre elas. Já a relação “especialização *tem algumas características de* prática” foi classificada como associativa do subtipo similaridade de atributo. Ela é simétrica e irreflexiva. A simetria entre os conceitos permite a relação inversa.

6.2.1.1 Relações ternárias e autorrelacionamento

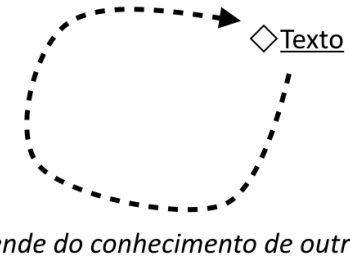
Somente a título de registro, durante a análise de relações ocorreu a manifestação de relações ternárias, como na frase: “Não há dúvida de que o *indexador* interponha suas próprias idéias e preconceitos na atuação de intermediário entre *autores* e usuários” (NAVES, 2000, p. 19, grifos nossos). Observa-se na frase que pode haver uma relação ternária entre *indexador*, *autores* e *usuários*: indexador *intermedeia* autores *e* usuários. De acordo com Khoo e Na (2006), uma relação ternária pode ser decomposta em relações binárias. Dessa forma, as seguintes relações podem se originar: indexador *intermedeia* autores e indexador *intermedeia*

usuários. Como o conceito *usuários* não está no escopo da amostra deste estudo de caso, a relação semântica entre ele e *indexador* não será aproveitada. Ressalta-se que um novo subtipo de relação semântica associativa foi criado para suprir a relação entre *autores* e *indexador*. Esse subtipo foi nomeado *agente subordinado.* Essa relação está descrita no Apêndice C.

Outro exemplo de relação ternária, mas um em que não é possível estabelecer uma relação semântica, foi observado na frase: "*Conceito* é a representação dum objeto pelo *pensamento*, por meio de suas características gerais" (NAVES, 2000, p. 55, grifos nossos). Nesse caso, a relação entre *conceito* e *pensamento* é indireta, ou seja, entre esses dois conceitos existe o conceito *objeto*: conceito *representa* objeto *pelo* pensamento. Como no exemplo anterior, essa relação pode ser decomposta em: conceito *representa* objeto e objeto *é representado pelo* pensamento. Contudo, *objeto* não faz parte da amostra e nas duas relações semânticas encontradas não existe uma entre *conceito* e *pensamento*.

Um caso de autorrelacionamento foi encontrado ao analisar a frase em que foram detectados os conceitos *autores* e *texto*: "E o último componente definido pelos autores acima é a intertextualidade, que concerne aos fatores que fazem a utilização de um *texto* dependente do conhecimento de outro(s) *texto*(s)" (NAVES, 2000, p. 47, grifos nossos). O autorrelacionamento configura-se entre *texto* e ele mesmo,em que texto *depende do conhecimento de outro* texto. Esse foi o único caso de autorrelacionamento percebido e ele está fora do escopo desta tese. A Figura 73 o exemplifica.

Figura 73 – Representação do autorrelacionamento entre *texto* e *texto*



Fonte: Elaborada pela autora.

6.2.1.2 Relações semânticas não explicitadas

Durante a análise dos dados, observou-se que nem sempre as relações semânticas entre os conceitos poderiam ser explicitadas. No contexto em que os conceitos se encontravam existia uma relação entre eles, contudo, por vezes, a explicitação poderia comprometer a mensagem do autor. Um dos fatores que impediu a determinação das relações semânticas foi a presença de algumas classes de palavras encontradas nas frases. É o caso dos advérbios “se” e “não” e de adjetivos, como nos exemplos 1, 2 e 3, respectivamente.

- *Exemplo 1* – Nesta frase foram identificados os conceitos *indexadores* e *texto*: “Certamente, se *indexadores* abordam um *texto* apenas com a intenção de decidir a questão de assunto em sistema de vocabulário, eles podem perder algumas nuances que poderiam acrescentar aos subseqüentes termos do índice” (NAVES, 2000, p. 27, grifos nossos). Nesse caso, resumir a relação entre *indexador* e *texto* como indexador *aborda* texto suprime a ideia do autor. Nesse exemplo, o adjunto adverbial de dúvida “se” antes do conceito *indexadores* acrescenta que a autora pretendeu questionar sobre a abordagem que o indexador realiza no texto. Dessa forma, entende-se que existe relação entre indexador e texto nesse contexto, mas que ela não pode ser traduzida na tripla sujeito-predicado-objeto, como se pretende nesta tese.
- *Exemplo 2* – “Recomenda-se que o *indexador* não focalize exclusivamente o conteúdo de *documentos*, mas tente antecipar o impacto e o valor de um documento para seu uso potencial (ALBRECHTSEN, 1993)” (NAVES, 2000, p. 61, grifos nossos). Nesse caso, observa-se que existe a relação entre os conceitos *indexador* e *documento*, contudo, existe uma negativa quanto à ação que não pode ser expressa na relação semântica.
- *Exemplo 3* – “Concepção orientada para o conteúdo – envolve uma interpretação adicional do conteúdo, que vai além dos limites da estrutura léxica e gramatical, com o estabelecimento de assuntos que não estão explicitamente colocados no texto, mas que são facilmente identificados pelo *indexador*, envolvendo, portanto, uma abstração mais indireta do

documento” (NAVES, 2000, p. 61, grifos nossos). Nesse exemplo, o advérbio “mais” e o adjetivo “indireta” na relação entre *indexador* e *documento* qualificam a relação *abstrai.* Logo, expressar a relação como “indexador *abstrai mais indiretamente o* documento” ou com a simplificação “indexador abstrai documento”, retirando os qualificadores, faz com que a explicitação da relação semântica seja considerada imperfeita à luz da mensagem que a autora da tese pretendeu transmitir.

O Apêndice D mostra todos os casos em que existem relações semânticas entre os conceitos em seus contextos, mas que não puderam ser explicitadas.

6.3 Análise e interpretação dos dados

Conforme relatado, na etapa de execução do estudo de caso, validou-se os dados e determinou-se as relações semânticas possíveis. Prosseguindo com o processo do estudo de caso conforme Wohlin *et al.* (2000), a etapa seguinte é a *Análise e interpretação dos dados*, caracterizada por análises quantitativa e qualitativa.

A organização desta seção será organizada conforme as linhas da Tabela 4, ou seja, primeiro serão analisados e interpretados os dados coletados no início, em seguida os dados considerados falsos e, por fim, os verdadeiros, que serão subdivididos em não explicitados e explicitados.

Tabela 4 – Dados coletados na fase de operação do estudo de caso

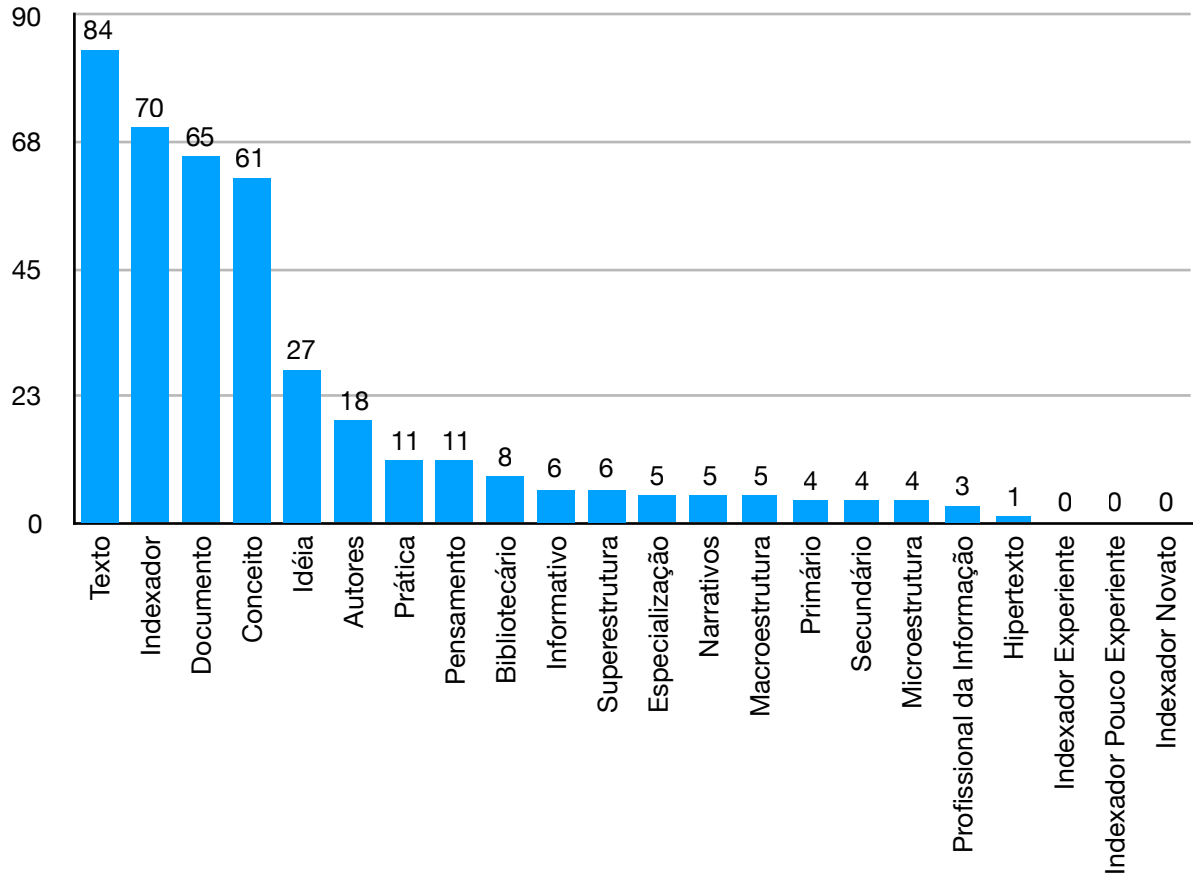
Dados	Quantidade de conceitos	Quantidade de pares de conceitos	Quantidade de indícios
Coletados no início	22	48	199
Analisados como falsos	3	17	94
Analisados como verdadeiros	**19**	**31**	**105**
Não explicitados	**6**	**14**	**35**
Explicitados	**18**	**17**	**70**

Fonte: Elaborada pela autora.

6.3.1 Análise e interpretação dos dados coletados no início do estudo de caso

A primeira análise considera as ocorrências dos 22 conceitos da amostra coletados do Semantizar. Como pode ser visto no gráfico da Figura 74, os conceitos que inicialmente tinham mais possibilidade de se relacionarem semanticamente com outros eram: *texto* (84 ocorrências), *indexador* (70 ocorrências), *documento* (65 ocorrências) e *conceito* (61 ocorrências). Esses quatro conceitos corresponderam a 70,35% das ocorrências nas frases analisadas. Isso indica que esses podem ser os principais conceitos abordados na tese da amostra. Por outro lado, nota-se que não houve ocorrências de indícios de relações semânticas com os conceitos *indexador experiente*, *indexador pouco experiente* e *indexador novato*; contudo, isso não indica que esses conceitos não existam no documento acadêmico da amostra, já que isso pode ter ocorrido por uma falha no algoritmo do Semantizar.

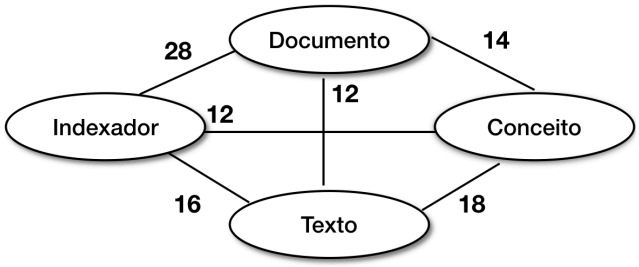
Figura 74 – Quantidade de ocorrência individual dos conceitos da amostra inicialmente no estudo de caso



Fonte: Elaborada pela autora.

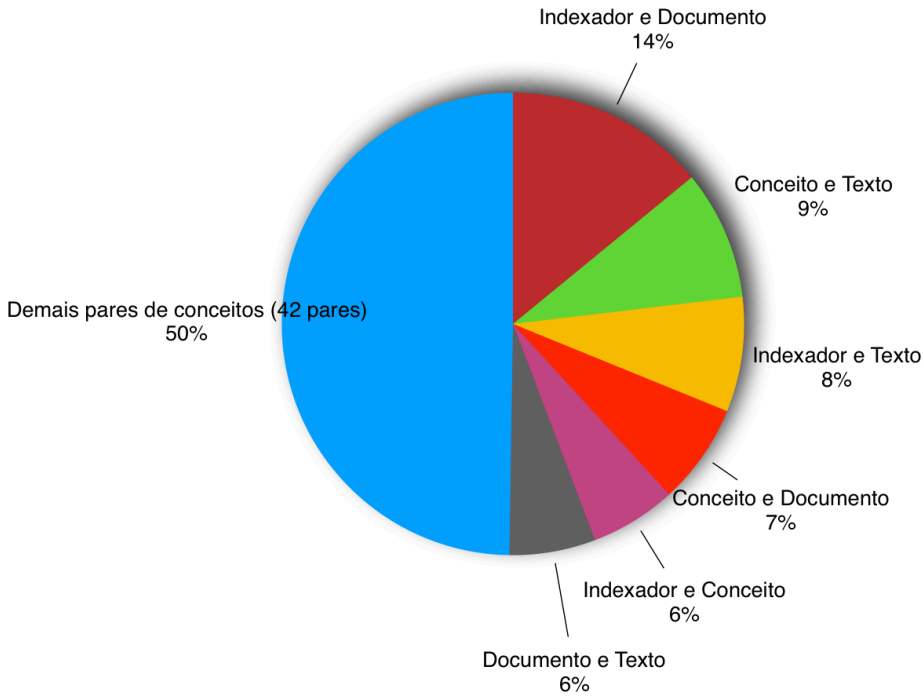
A Figura 75 mostra a quantidade de indícios de relações semânticas entre os quatro conceitos considerados mais relevantes. Logo a seguir, pode ser observado no gráfico da Figura 76 que os 6 pares formados com esses conceitos correspondem a 50% dos indícios de relações semânticas; os outros 50% são formados pelos demais 42 pares de conceitos. Isso significa que, inicialmente, a metade dos indícios encontrados de relações semânticas está concentrada em 12,5% dos pares de conceitos (6 pares), enquanto os demais 42 pares de conceitos, que correspondem a 81,5%, podem indicar relações entre assuntos complementares aos pares de conceitos mais relevantes nos capítulos do documento acadêmico da amostra.

Figura 75 – Quantidade de indícios de relações entre os quatro conceitos mais recorrentes na operação do estudo de caso.



Fonte: Elaborada pela autora.

Figura 76 – Gráfico da porcentagem de indícios dos pares de conceitos e relevância desses pares em relação aos demais

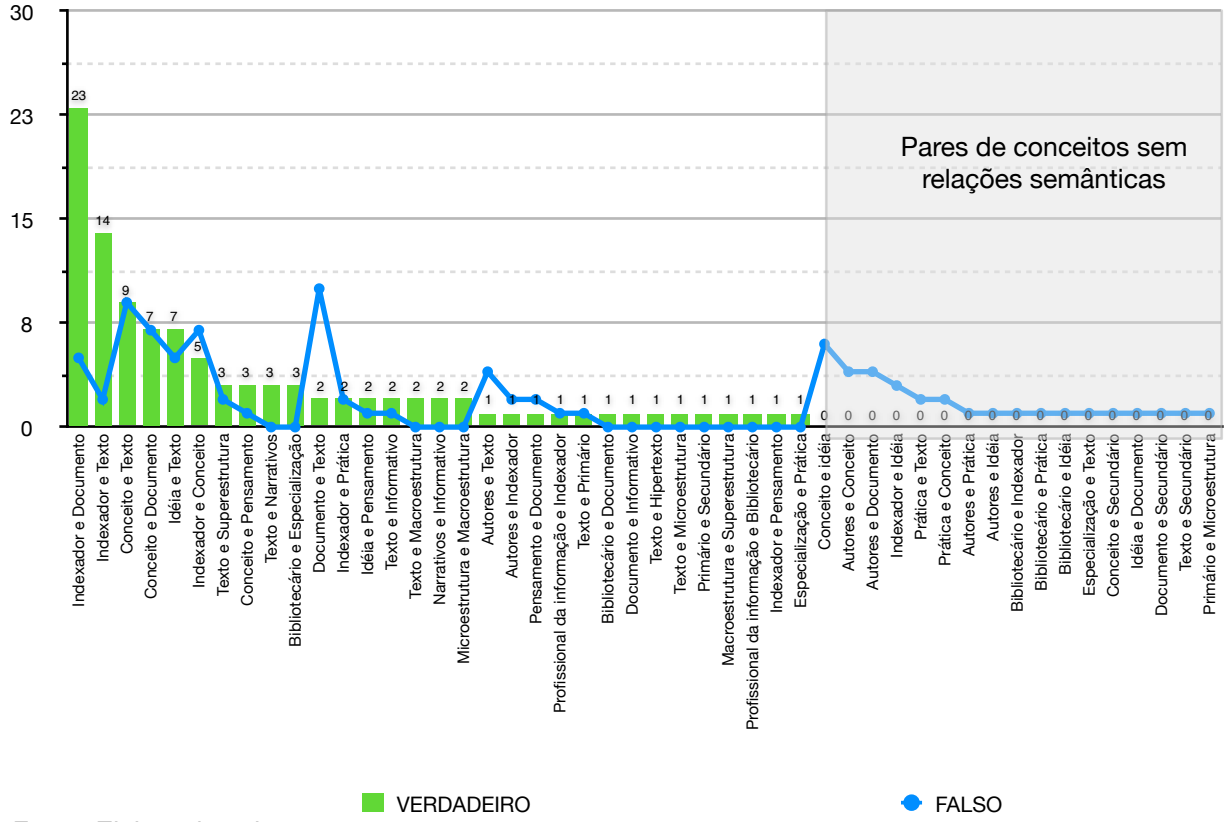


Fonte: Elaborada pela autora.

Essa primeira análise realizada baseou-se nos dados coletados em sua totalidade, sem a verificação se os indícios de relações semânticas encontrados pelo Semantizar correspondiam realmente às relações semânticas. Dessa forma, cada um dos 199 indícios de relações foram manualmente investigados quanto à sua veracidade. Essa validação resultou em uma redução de 47% dos indícios de relações semânticas considerados falsos no contexto em que se apresentavam.

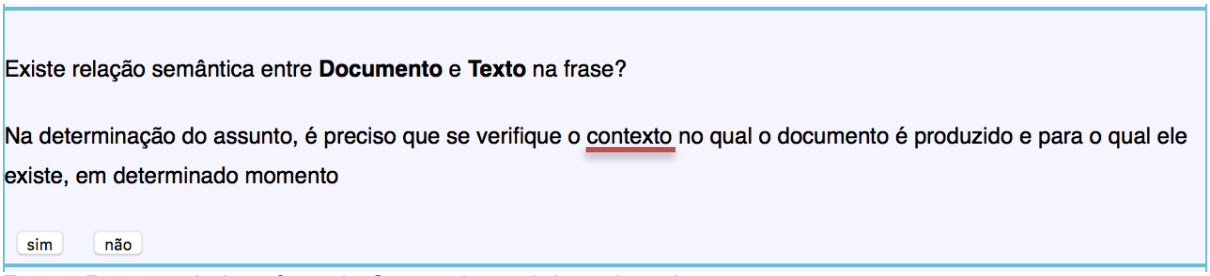
O gráfico apresentado na Figura 77 mostra os pares de conceitos e a quantidade de indícios verdadeiros (representados pelas barras) e falsos (representados pela linha). Como pode ser visto nos últimos pares do gráfico, todos os indícios foram falsos. Ainda nesse gráfico, observa-se uma disparidade no par de conceitos *documento* e *texto*, que foi, inicialmente, indicado como um dos pares com mais indícios de relações semânticas. Em análise sobre as ocorrências falsas desse par de conceitos, dos 10 casos determinados como falsos, três deles são decorrentes de um erro do algoritmo do Semantizar, que identificou a palavra *contexto* como *texto*. Nesse sentido, ao encontrar parte de uma palavra da amostra da estrutura classificatória, ele compreendia que havia encontrado a palavra. A Figura 78 mostra um exemplo desse caso.

Figura 77 – Gráfico com os pares de conceitos e as quantidades de indícios verdadeiros e falsos para cada par



Fonte: Elaborada pela autora.

Figura 78 – Exemplo onde a palavra *contexto* foi identificada erroneamente

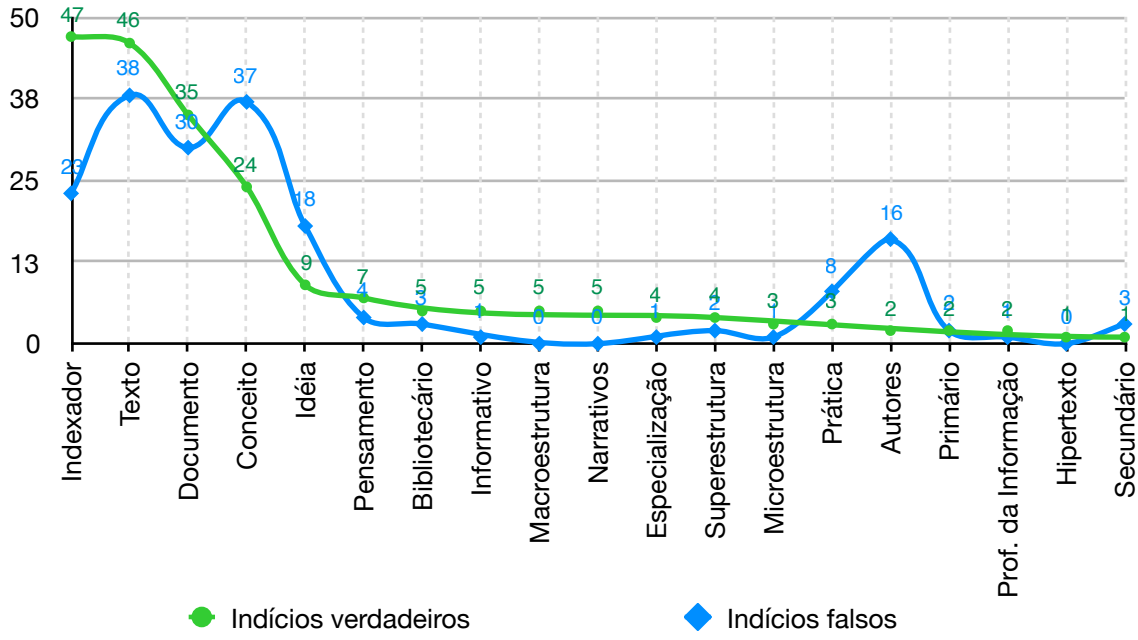


Fonte: Recorte da interface do Semantizar, elaborado pela autora.

Em análise individual dos conceitos com indícios falsos, o gráfico apresentado na Figura 79 retrata que os conceitos com mais indícios falsos do que verdadeiros foram: *conceito*, *ideia*, *prática*, *autores* e *secundário*. Desses, como pode ser observado no gráfico, conceitos *conceito* e *autores* foram os que tiveram maior disparidade entre indícios falsos e verdadeiros. Em análise realizada sobre esses dois conceitos, observou-se que eles não foram empregados com o sentido coeso com o contexto. A Figura 80 mostra um exemplo em que *conceito* não foi empregado com o sentido de unidade de conhecimento, conforme o domínio em que ele está inserido, e sim como uma introdução para significar algo, e em que *autores*, como na maioria das

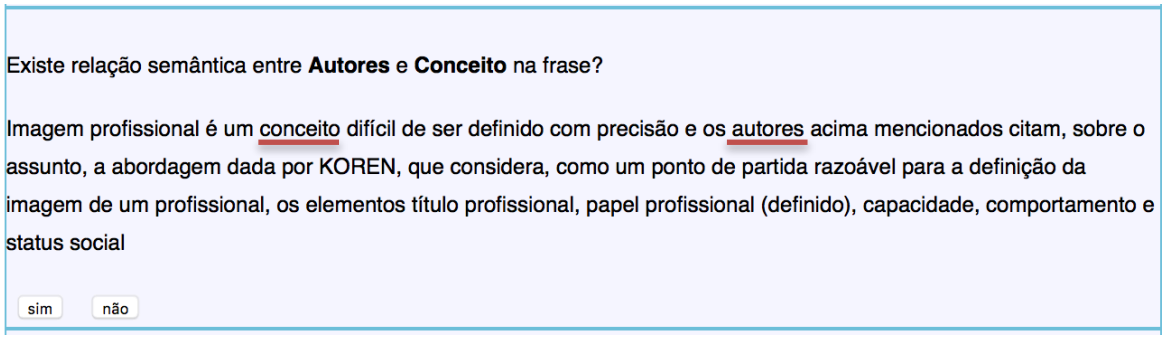
vezes em que ocorreu, remeteu a uma citação, tornando-se inválido para o estabelecimento de uma relação semântica.

Figura 79 – Gráfico com a quantidade de indícios verdadeiros e falsos



Fonte: Elaborada pela autora.

Figura 80 – Recorte do Semantizar com as ocorrências falsas de *autores* e *conceito*



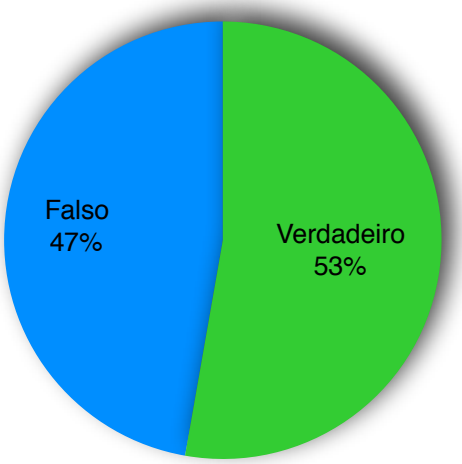
Fonte: Recorte da interface do Semantizar, elaborado pela autora.

6.3.2 Análise e interpretação de indícios verdadeiros de relações semânticas

Conforme pode ser visto no gráfico apresentado na Figura 81, os indícios de relações semânticas verdadeiros ultrapassaram os falsos. Essa diferença foi pouca,

mas isso indica que a ocorrência de dois conceitos em uma frase pode apontar para uma relação semântica entre eles.

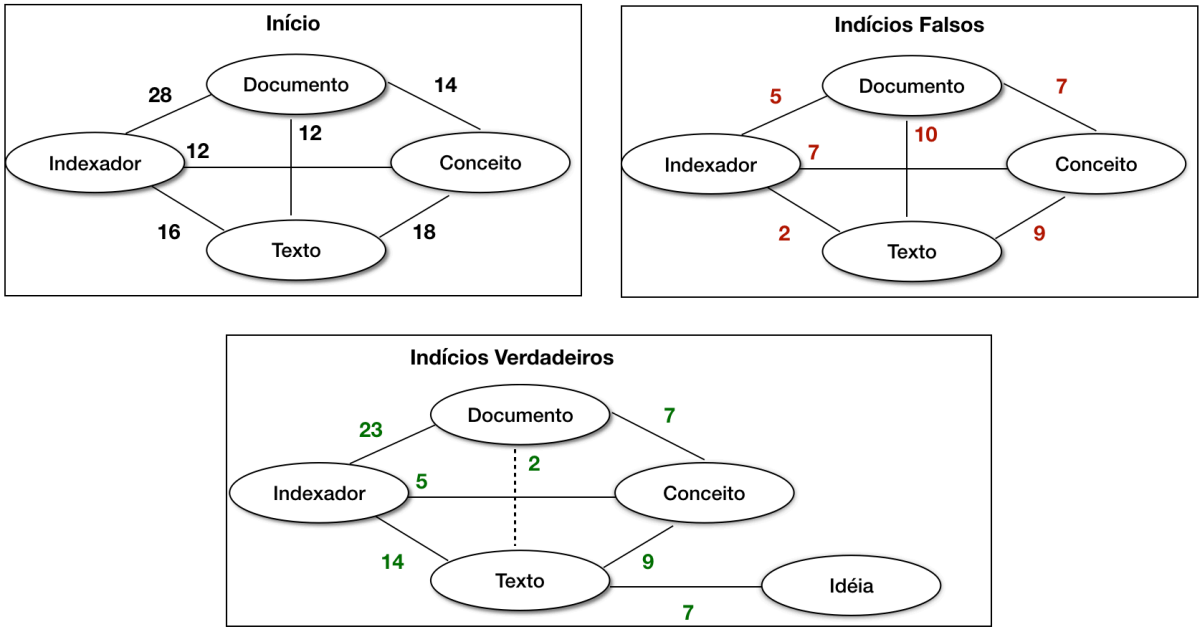
Figura 81 – Gráfico com a confirmação preliminar de existência de relação semântica



Fonte: Elaborada pela autora.

Na análise dos pares de conceitos detectados pelo Semantizar cuja validação atestou a existência de relação semântica no contexto em que eles foram apresentados, o par *documento* e *indexador* manteve-se com mais relações semânticas. Dos outros cinco pares de conceitos com mais relações semânticas, o par *documento* e *texto*, detectado inicialmente como um dos pares com mais indícios de relações semânticas, foi sobreposto pelo par de conceitos *ideia* e *texto*. A Figura 82 mostra esse conjunto de pares considerados mais relevantes devido ao grande número de ocorrências em relação aos demais. Como pode ser visto nessa figura, dentro desse conjunto de pares de conceitos, o conceito *ideia* se relaciona apenas com o conceito *texto*, diferentemente dos demais, que relacionam-se todos entre si.

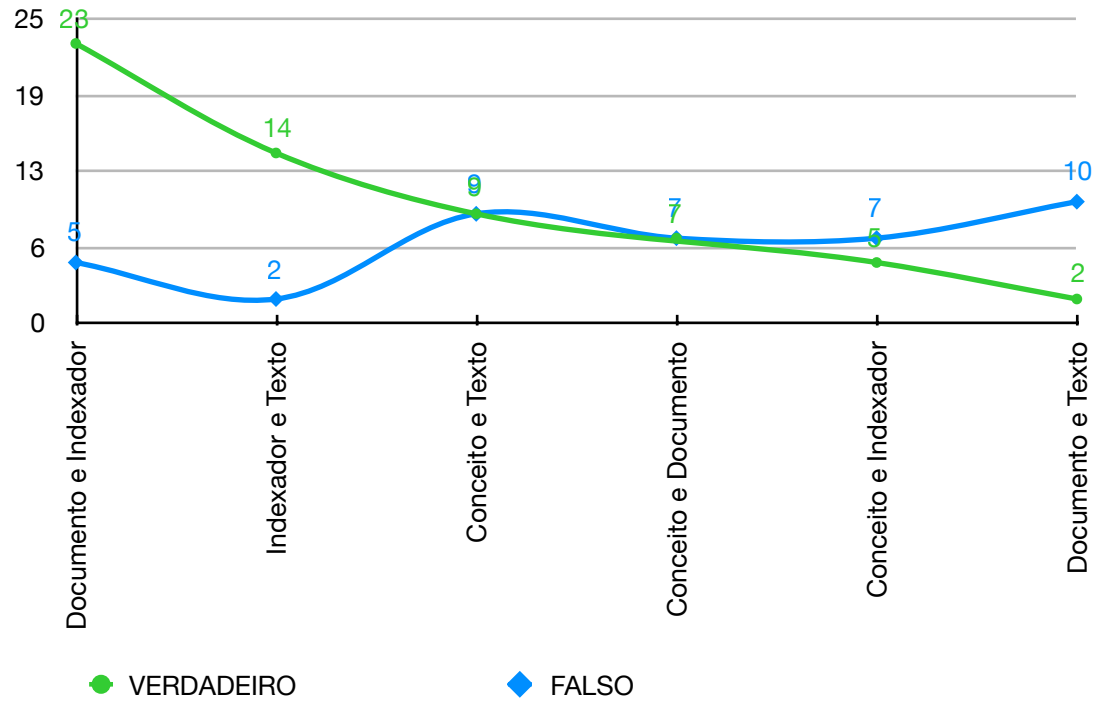
Figura 82 – Pares de conceitos com mais indícios de relações semânticas.



Fonte: Elaborada pela autora.

Como pode ser observado na Figura 83, dos pares de conceitos com mais relações semânticas apresentados na Figura 82, houve aqueles em que o números de indícios verdadeiros superou os falsos, como em *indexador* e *documento* e *indexador* e *texto*; mas também houve aqueles em que os indícios falsos superaram os verdadeiros, como em *indexador* e *conceito* e *documento* e *texto*. Há ainda aqueles em que o número de indícios verdadeiros e falsos foram os mesmos, como em *texto* e *conceito* e *documento* e *conceito*.

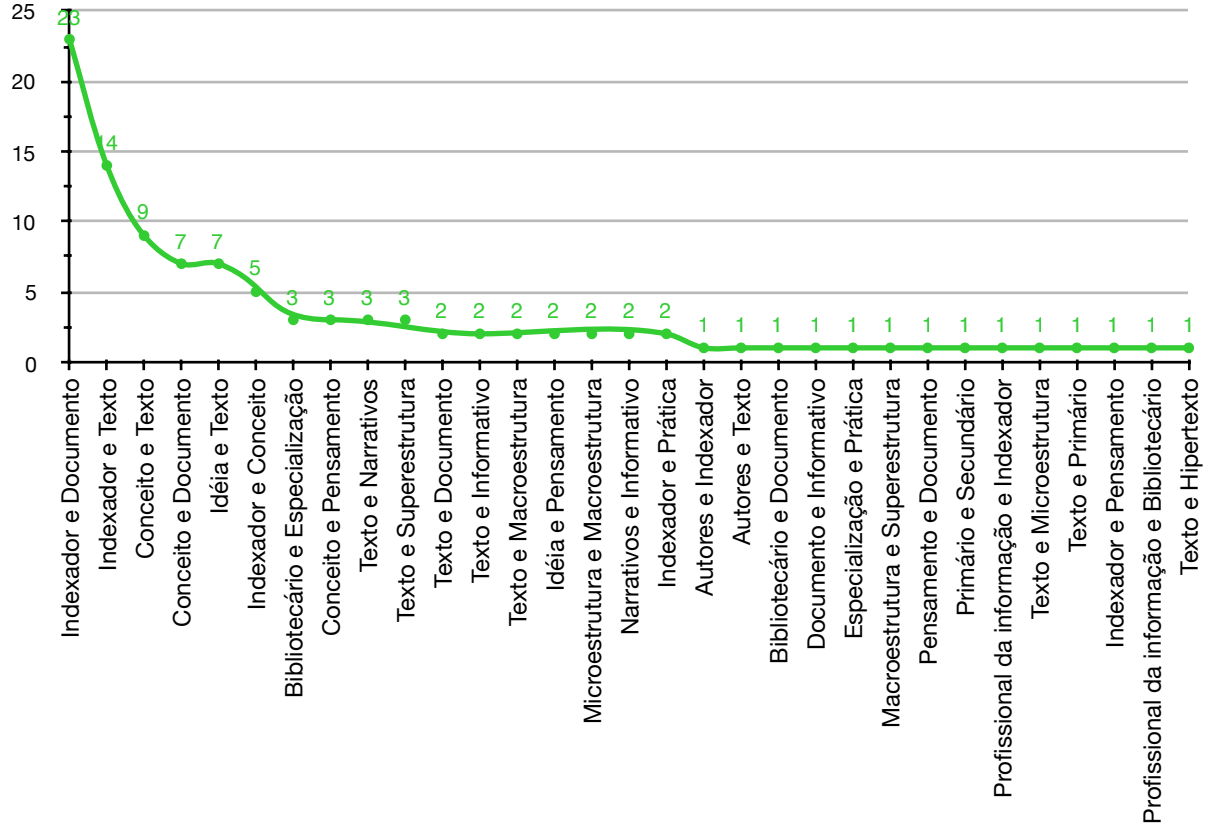
Figura 83 – Gráfico com a marcação dos indícios verdadeiros e falsos dos seis pares de conceitos identificados inicialmente como os mais relevantes.



Fonte: Elaborada pela autora.

A Figura 84 mostra o gráfico com a quantidade de todos os pares de conceitos verdadeiros, como pode ser visto. Com excessão dos seis primeiros pares de conceitos, os demais valores são quase que lineares.

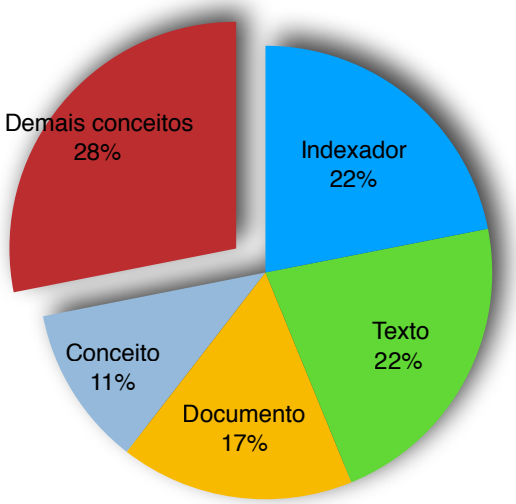
Figura 84 – Gráfico da quantidade de pares de conceitos com relações semânticas



Fonte: Elaborada pela autora.

Conforme relatado na seção 6.3.1, os conceitos com mais indícios de relações semânticas foram *indexador*, *texto*, *documento* e *conceito*. Como esperado, esses conceitos também figuram entre aqueles com mais indícios de relações semânticas verdadeiros: *indexador* (46), *texto* (46), *documento* (35) e *conceito* (24). O gráfico da Figura 85 mostra que, juntos, esses conceitos concentraram 72% dos indícios de relações semânticas verdadeiros.

Figura 85 – Gráfico com a porcentagem dos conceitos com mais indícios verdadeiros de relações semânticas em relação aos demais



Fonte: Elaborada pela autora.

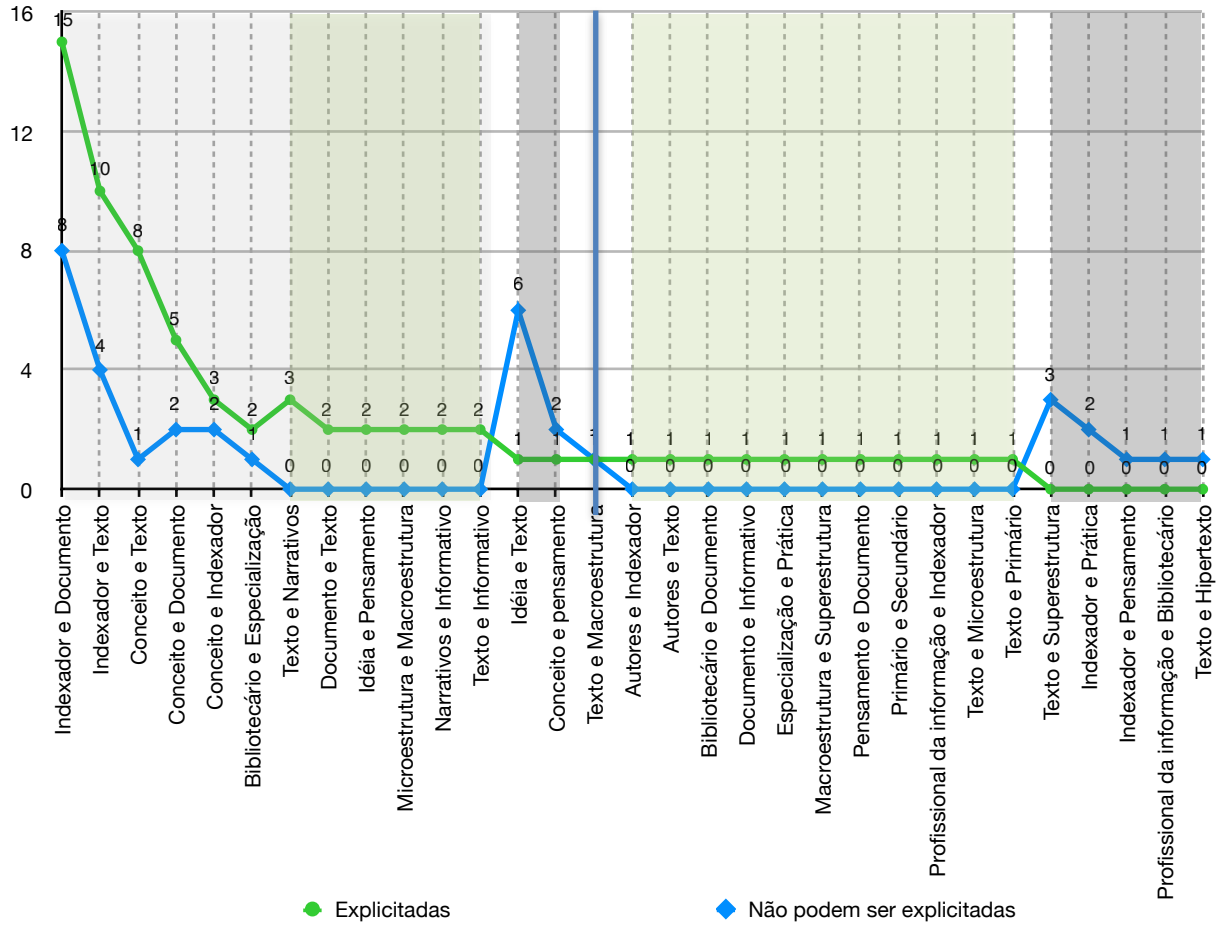
Durante a validação dos dados, observou-se que alguns pares de conceitos detectados pelo Semantizar possuíam relação semântica, contudo elas não poderiam ser explicitadas. Como pode ser visto no gráfico da Figura 86, essa situação ocorreu em 33% dos indícios verdadeiros. Com efeito, os dados desse gráfico indicam que, somente após avaliação humana, os indícios de relações semânticas verdadeiros podem ser explicitados ou não e que a relação semântica pode existir entre dois conceitos, mas, por vezes, ela não pode ser determinada, o que não a isenta de existir.

Figura 86 – Porcentagem de relações semânticas explicitadas e não explicitadas

Fonte: Elaborada pela autora.

O gráfico da Figura 87 mostra a indicação da quantidade de relações semânticas entre pares de conceitos explicitadas e não explicitadas. As áreas do gráfico que estão em destaque mais escuro englobam sete pares de conceitos que tiveram mais relações não explicitadas do que explicitadas. Dessas, a última área, que abrange desde *texto* e *superestrutura* até *texto* e *hipertexto*, denotam relações semânticas que não puderam ser explicitadas de maneira alguma.

Figura 87 – Pares de conceitos com relações semânticas que podem e não podem ser explicitadas



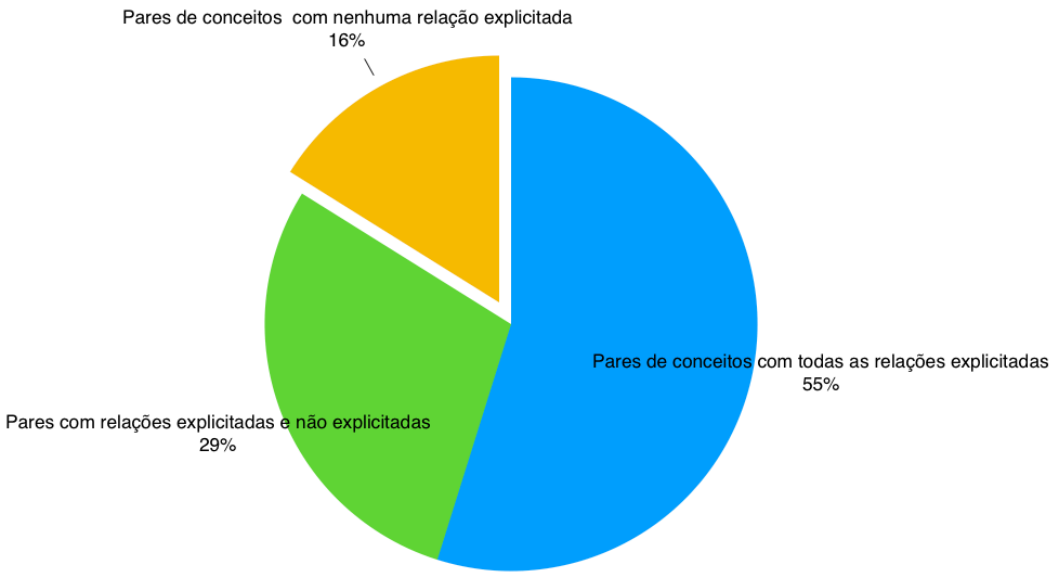
Fonte: Elaborada pela autora.

De modo oposto, as áreas do gráfico da Figura 87 que compreendem os conceitos *indexador* e *documento* a *texto* e *informativo*, e os conceitos *autor* e *indexador* a *texto* e *primário*, correspondem aos pares de conceitos cujas relações explicitadas superaram as não explicitadas. Desses pares, as áreas que abrangem de *texto* e *narrativo* a *texto* e *informativo* e de *autor* e *indexador* até *texto* e *primário* tiveram todas as relações semânticas explicitadas. Já o par de conceitos *texto* e *macroestrutura* tiveram a mesma quantidade de relações explicitadas e não explicitadas.

Conforme já esperado, a maioria dos pares de conceitos com relações semânticas pode ser explicitada (ver o gráfico da Figura 88). Dos 31 pares de conceitos analisados como verdadeiros, 55% correspondem a casos em que todas as relações semânticas dos pares de conceitos podem ser explicitados, 29% reúne pares

de conceitos cujas relações semânticas ora podem ser explicitadas, ora não, e 16% correspondem aos pares de conceitos sem relações semânticas explicitadas.

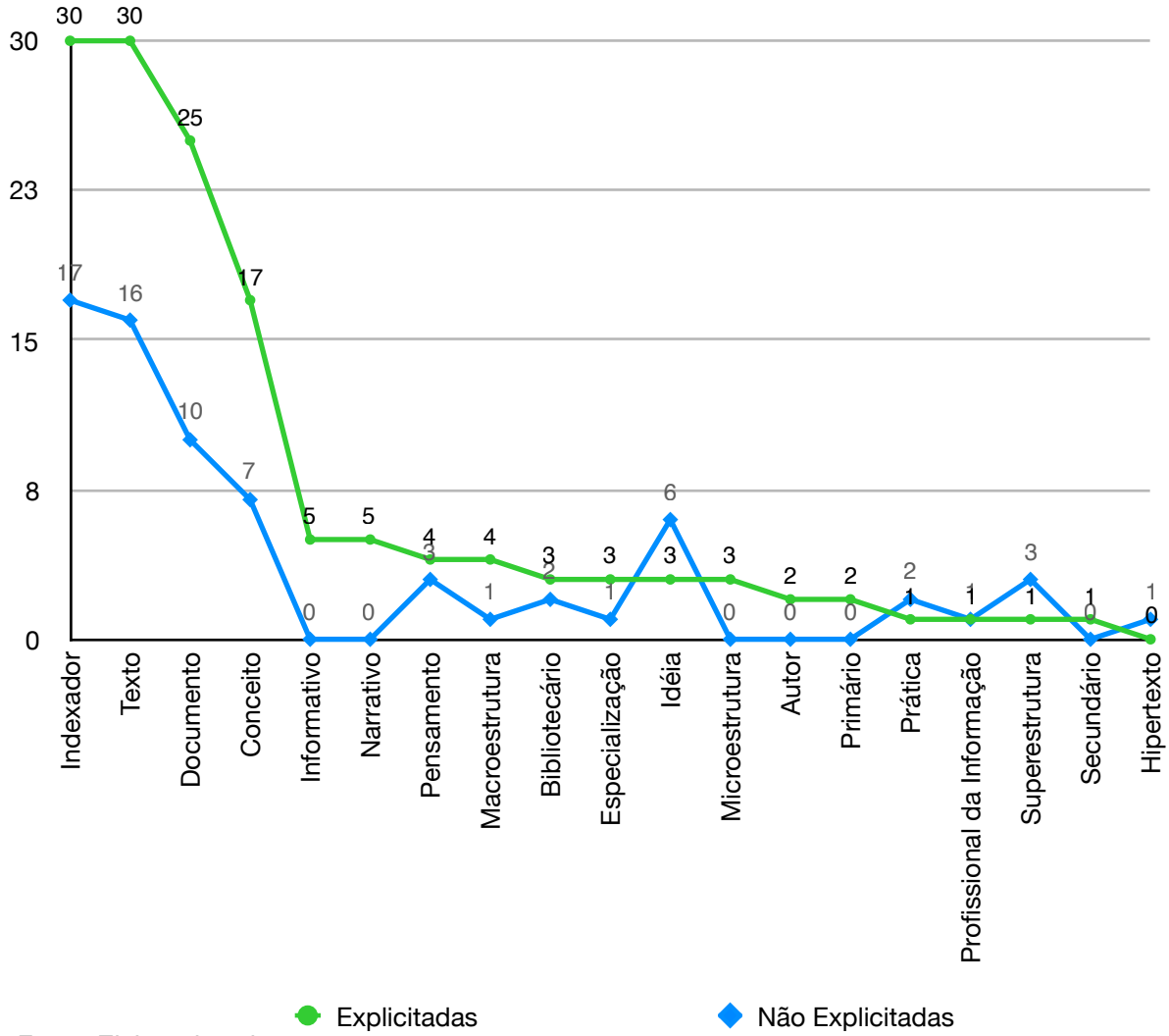
Figura 88 – Porcentagem de pares de conceitos cujas relações podem e/ou não podem ser explicitadas



Fonte: Elaborada pela autora.

Em análise individual dos conceitos cujos pares podem ter relações semânticas explicitadas, observou-se que a maioria dos conceitos teve mais relações semânticas explicitadas do que o contrário. Como pode ser observado no gráfico da Figura 89, com exceção dos conceitos *ideia*, *prática* e *superestrutura*, todos os outros conceitos tiveram quantidade igual ou maior de relações explicitadas. Os conceitos *informativo*, *narrativos*, *microestrutura*, *autor*, *primário* e *secundário* que tiveram todas as relações semânticas explicitadas no contexto em que se apresentavam. O único conceito que não teve relações semânticas explicitadas foi *hipertexto*, conforme pode ser constatado no gráfico.

Figura 89 – Conceitos com relações explicitadas e não explicitadas



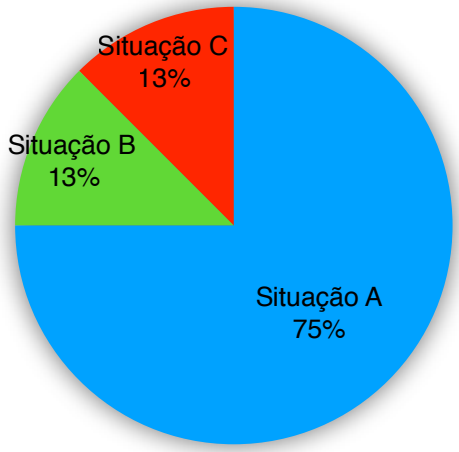
Fonte: Elaborada pela autora.

6.3.2.1 Análise e interpretação das relações semânticas explicitadas

Como relatado anteriormente, os indícios de relações semânticas verdadeiros que poderiam ser explicitados eram 70; contudo, ao determinar as relações semânticas, esse número aumentou para 78 relações. Como pode ser visto no gráfico apresentado na Figura 90, ocorreram três situações diferentes nesse aspecto: (A) relações semânticas cujo número de indícios verdadeiros e que poderiam ser explicitados foi igual ao número de relações explicitadas que surgiram após a análise; (B) pares de conceitos que tiveram mais relações semânticas explicitadas do que o que se previu; e (C) surgimento de pares de conceitos com relações semânticas

descobertos ao longo da determinação das relações semânticas. Essas situações estão explicadas a seguir.

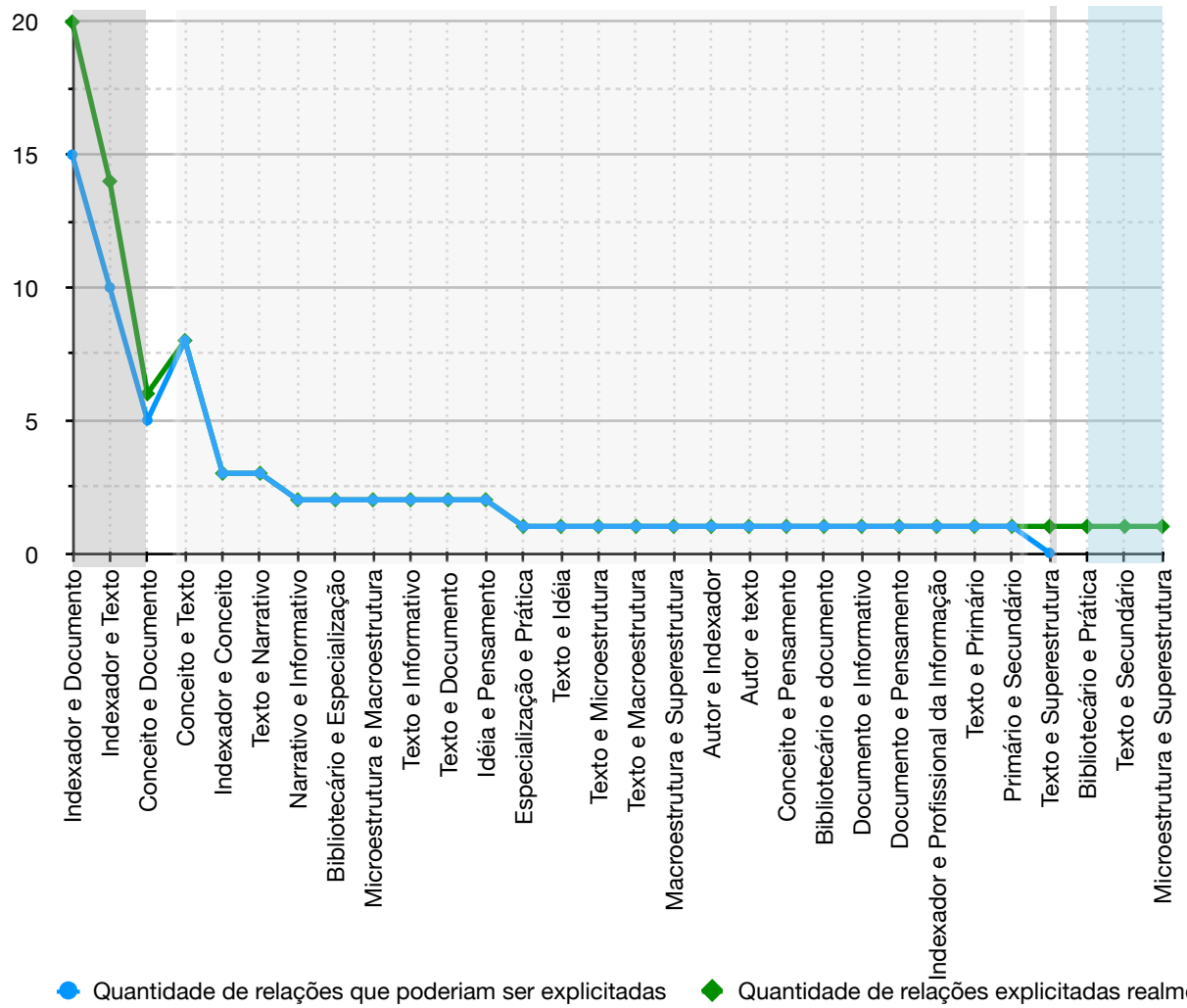
Figura 90 – Gráfico com a porcentagem indicativa das situações de explicitação de relações semânticas surgidas



Fonte: Elaborada pela autora

- *Situação A*: A primeira situação correspondeu à maioria dos casos (ver gráfico da Figura 90), em que o número de relações explicitadas foi semelhante ao número de relações que poderiam ser explicitadas. Como pode ser verificado no gráfico apresentado na Figura 91, essa situação foi identificada nos 24 pares de conceitos que abrangem desde *conceito* e *texto* até *primário* e *secundário*.

Figura 91 – Gráfico com com a diferença das quantidade de relações explicitadas para cada par de conceitos



Fonte: Elaborada pela autora.

- *Situação B*: Nesse caso, os pares de conceitos *indexador* e *documento*, *indexador* e *texto* e *conceito* e *documento* obtiveram mais relações semânticas explicitadas do que o que se antecipou (ver a área mais escura no início do gráfico apresentado na Figura 91). Um exemplo dessa constatação foi o primeiro caso apresentado na seção 6.2, na frase: “No caso do processamento técnico do acervo de *documentos*, o profissional responsável pela catalogação, classificação e indexação, costuma ter a formação bibliotecária, e receber o nome de *indexador*.” (NAVES, 2000, p. 16, grifos nossos). Nessa frase, explicitou-se não uma, mas três relações semânticas: indexador *cataloga* documento, indexador *classifica* documento e, indexador *indexa* documento. Ainda nessa situação, além dos três pares de conceitos supracitados, o par de conceitos *texto* e *superestrutura* inicialmente não teria

relações semânticas explicitadas; contudo, durante a análise das relações complexas na seção 6.2.1, uma relação semântica foi explicitada para esse par: texto *tem tipo de estrutura* superestrutura.

- *Situação C*: Essa situação denota o surgimento de pares de conceitos com relações semânticas descobertos ao longo da explicitação das relações semânticas. Ela compreende especialmente os três últimos pares de conceitos do gráfico apresentado na Figura 91. Quanto aos pares *bibliotecário* e *prática* e *texto* e *secundário*, no contexto em que o Semantizar os detectou, não era possível existir relação semântica entre eles. Contudo, durante a análise das relações complexas, essas relações surgiram em outros fragmentos da amostra, conforme pode ser visto no Quadro 11. Já o par de conceitos *microestrutura* e *superestrutura* não foi identificado pelo Semantizar, mas, também durante a análise de relações complexas, surgiu a relação: microestrutura *é assimetricamente contrário a* superestrutura. Por fim, para a frase “Em síntese, este capítulo mostra o processo em que o indexador faz a leitura de um texto, empreende a extração de conceitos e determina a sua atinência” (NAVES, 2000, p. 69), o Semantizar detectou a presença dos pares de conceitos: *indexador* e *conceito* e *conceito* e *texto*, conforme pode ser visto na Figura 92. Contudo, ao realizar a análise dessa frase, percebeu-se que nela também poderia existir a relação entre *indexador* e *texto*. E mais, constatou-se que duas relações semânticas poderiam existir entre *indexador* e *texto*, quais sejam, indexador *lê* texto e indexador *determina atinência do* texto. Essa última situação aponta que, durante a análise humana para a explicitação de frases em que os pares de conceitos são encontrados pelo Semantizar, relações semânticas podem surgir, dependendo do conhecimento do domínio do agente que está determinando as relações semânticas.

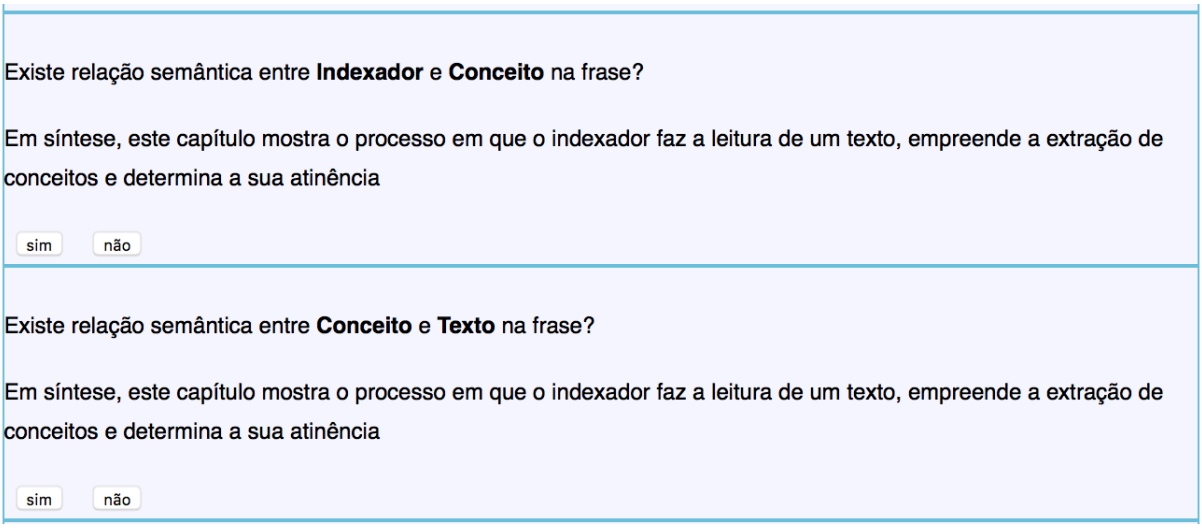
Quadro 11 – Pares de conceitos com indícios falsos no contexto em que foram detectados pelo Semantizar, porém verdadeiros durante a análise em outro contexto

Pares de conceito	**Indícios encontrados pelo Semantizar**	**Relação semântica encontrada após análise**
Bibliotecário e *prática*	“Quanto à *prática*, somente após longa experiência, é que, provavelmente, o *bibliotecário* desenvolverá métodos de trabalho eficientes” (NAVES, 2000, p. 22, grifos nossos).	“A especialização e a *prática* do *bibliotecário* são tratadas por INGWERSEN (1982)” (NAVES, 2000, p. 22, grifos nossos).

Pares de conceito	Indícios encontrados pelo Semantizar	Relação semântica encontrada após análise
Texto e *secundário*	“Segundo essa autora, […] problemas contingentes: são ligados, por exemplo, ao significado de uma palavra desconhecida; problemas táticos: são relacionados à organização do *texto*, por exemplo, a confusão entre os planos principal e *secundário* do discurso […] (NAVES, 2000, p. 52, grifos nossos).	“O texto original é chamado texto primário, um sumário ou resumo, *texto secundário*, e a expressão do texto primário numa linguagem documentária, texto terciário” (NAVES, 2000, p. 48, grifos nossos).

Fonte: Elaborado pela autora.

Figura 92 – Recorte do Semantizar onde ocorreu uma falha e o par de conceitos *indexador* e *texto* não foi identificado



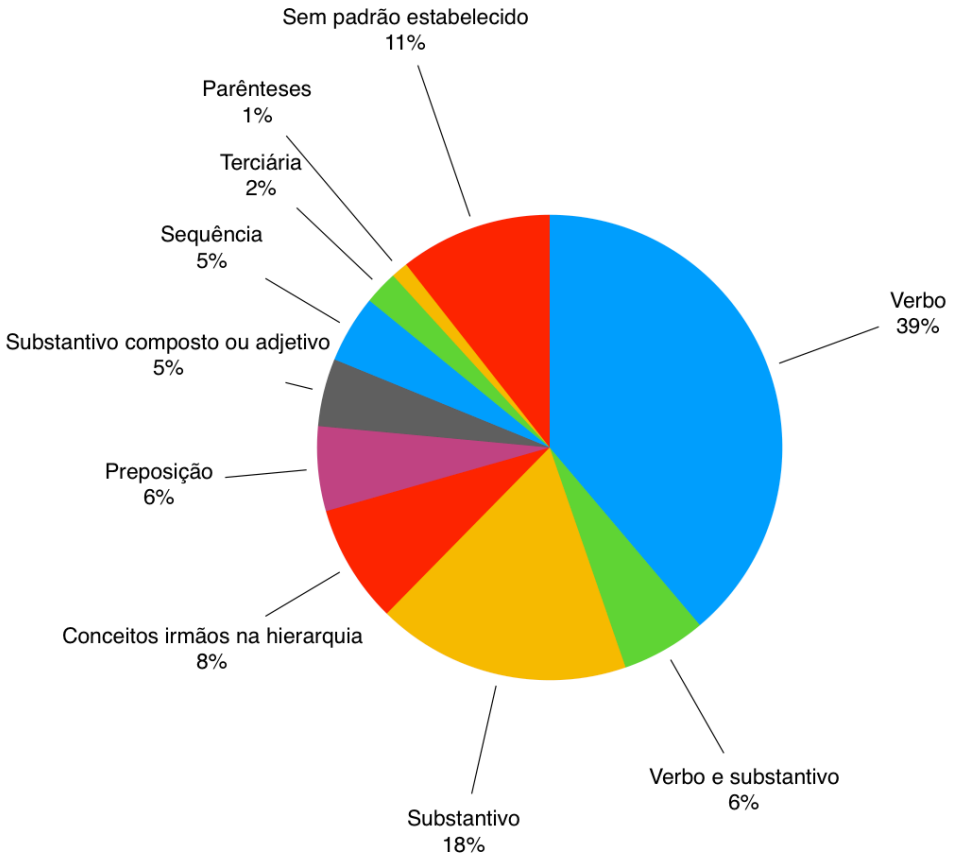
Fonte: Recorte da interface do Semantizar, elaborado pela autora.

6.3.2.2 Características da explicitação das relações semânticas

Durante a análise, constatou-se nessa amostra que as relações semânticas podem ser descobertas por meio de verbos e outras classes gramaticais. Como pode ser visto no gráfico da Figura 93, quase maioria das relações semânticas foram explicitadas por verbos. Portanto, durante a análise dos dados, verificou-se que existe uma certa complexidade relacionada aos verbos, pois em alguns casos eles estavam no particípio, em outros eles estavam acompanhados de verbos auxiliares e, em outros casos, eles estavam posicionados antes dos conceitos, diferentemente do que se imaginou, isto é, que os verbos estariam sempre entre dois conceitos, o que

tornaria a determinação das relações semânticas mais compreensível, como neste exemplo: “Mas pouco é encontrado sobre como indexadores *decidem* qual é o *assunto de um* documento […]” (NAVES, 2000, p. 28, grifos nossos). Diferentemente, na frase “No ato de pensar, quando *faz abstrações, interpreta e define o assunto de um* documento, o indexador sofre influência […]” (NAVES, 2000, p. 70, grifos nossos), nota-se que os verbos (nesse caso foram determinadas três relações diferentes) estão posicionados antes dos dois conceitos. Observa-se ainda que a primeira relação semântica dessa frase é formada por um verbo e um substantivo: *faz abstrações.* Outros casos parecidos com esse, de verbos e substantivos, corresponderam a 6% das formas de descobrir relações semânticas, como pode ser visto no gráfico da Figura 93.

Figura 93 – Gráfico com as formas de descobrir relações semânticas



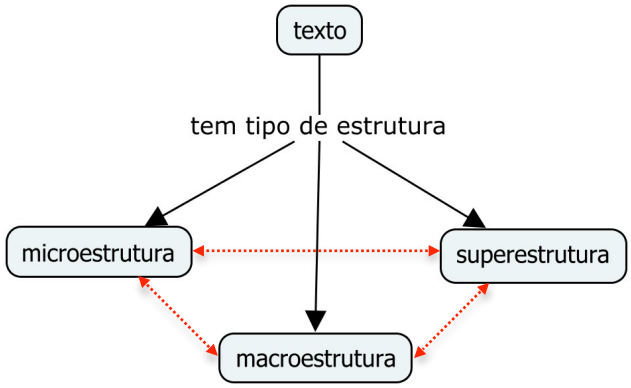
Fonte: Elaborada pela autora.

Outra classe gramatical que sustentou a criação de relações semânticas foram os substantivos (18%), como na frase: “Não há dúvida de que o *indexador* interponha suas próprias idéias e preconceitos na atuação de intermediário entre

autores e usuários" (NAVES, 2000, p. 19, grifos nossos). Nesse caso, a relação semântica descoberta foi: indexador *intermedia* autores.

As relações semânticas determinadas pelos conceitos irmãos (8%) fazem parte daquelas cuja explicitação foi considerada complexa. Nesse caso, as relações semânticas associativas entre os conceitos *microestrutura* e *macroestrutura*, *macroestrutura* e *superestrutura* e *microestrutura* e *superestrutura* foram estabelecidas considerando-se que esses conceitos têm o mesmo relacionamento hierárquico com o conceito *texto* e têm características parecidas. A Figura 94 mostra esse relacionamento. Nela, as setas vermelhas pontilhadas denotam as relações entre os conceitos irmãos.

Figura 94 – Relacionamento entre conceitos irmãos



Fonte: Elaborada pela autora.

As relações semânticas descobertas por meio de preposição corresponderam a 6%. Foi o caso da frase: "Fazem uma análise com o objetivo de determinar o conteúdo *informativo* do *documento*, tendo em vista o objetivo do sistema e as necessidades dos usuários […]" (NAVES, 2000, p. 80, grifos nossos). Nesse caso, a relação foi descoberta pela preposição "de" e foi explicitada da seguinte forma: documento *tem conteúdo* informativo. Nessa situação, especificamente, entende-se que *informativo* é um adjetivo de *documento*. Essas situações de adjetivos ou substantivos compostos aconteceram em outros 5% dos casos.

As descobertas de relações semânticas denotadas por sequência dos conceitos na estrutura sintática das frases também possibilitou a descoberta de

relações semânticas (5%). Esse tipo de explicitação foi considerado complexo. A frase a seguir mostra o exemplo:

> O sentido geral do *texto* é baseado na seguinte trilogia estrutural: *microestrutura* (estrutura superficial, que corresponde à realidade física do texto e seus símbolos de significação, as palavras), *macroestrutura* (concebida como um tópico representativo hierárquico e coerente da unidade textual, envolvendo mínima estrutura da representação textual sintática-semântica), e a superestrutura (estrutura retórica-esquemática, um tipo de esquema de produção convencional para o qual o texto é adaptado, podendo ser considerado como transição entre estruturas de superfície e de profundidade). (NAVES, 2000, p. 43, grifos nossos)

Nesse caso, as relações semânticas são: texto *tem tipo de estrutura* microestrutura, texto *tem tipo de estrutura* macroestrutura e texto *tem tipo de estrutura* superestrutura.

Conforme mencionado, as relações ternárias (somente 2%) não são foco desta tese. Contudo, por meio delas, descobriu-se algumas relações, como mostra a seção 6.2.1.1. A relação semântica determinada por parênteses, que denotou uma relação de sinonímia, correspondeu a 1% das relações encontradas. Ela ocorreu na frase: "[…] e podem ser diferenciados de outros *pensamentos* (percepção, *conceitos*, *idéias*)" (NAVES, 2000, p. 86, grifos nossos). Nesse caso, *conceito* e *ideia* estão entre parênteses com a intenção de reforçar a compreensão do conceito *pensamento*.

Por fim, 11% de relações semânticas encontradas não tiveram um padrão estabelecido para determinação. Elas dependeram da interpretação de quem explicitou a relação, como no caso: "Apesar de este estudo tratar especificamente da pessoa do *indexador*, termo adotado para designar o profissional que faz a indexação, algumas considerações devem ser feitas, inicialmente, sobre o *profissional da informação*" (NAVES, 2000, p. 14, grifos nossos). Para essa frase, a relação inferida foi: indexador *é um* profissional da informação.

Além das diferentes maneiras de encontrar uma relação semântica, outra característica observada durante a explicitação das relações semânticas foi que, algumas vezes, a ordem dos conceitos tinha que ser invertida para que a relação semântica fosse identificada. Esses casos ocorreram principalmente naqueles casos cuja relação inversa não era possível de ser determinada, como na frase: "Dentre os *textos* informativos, pode-se reconhecer o *texto* científico e, nesse tipo de *texto*, o conteúdo é quase inteiramente determinado pelo *autor*; geralmente, relatórios de pesquisa são altamente informativos e os *autores* os constroem numa estrutura

convencional com introdução, metodologia, resultados e discussão, o chamado modelo clássico" (NAVES, 2000, p. 44, grifos nossos). Nesse exemplo, a relação semântica é: autores *determinam conteúdo do* texto. Como pode ser observado, a ordem de *autores* e *texto* na frase e na relação semântica são diferentes.

Por fim, constatou-se, durante a explicitação das relações semânticas, que algumas delas se repetiram ao longo da determinação das relações. Isso ocorreu em quase 9% das relações explicitadas. Dessas, três relações semânticas repetiram duas vezes e outras três repetiram três vezes. Esses dados reafirmam a existência dessas relações. E mais, como pode ser visto no Quadro 12, a relação de sinonímia entre *texto* e *documento* foi reafirmada pelas relações: conceito *é extraído do* texto e conceito *é extraído do* documento. O Quadro 13 mostra um exemplo de uma relação semântica repetida em outras frases.

Quadro 12 – Relações semânticas que se repetiram em outros contextos

Relação semântica	Quantidade de vezes que repetiu
Indexador *lê* texto	3
Conceito *é extraído do* documento	3
Conceito *é extraído do* texto	3
Conceito *é parte do* texto	2
Indexador *extrai* conceito	2
indexador *determina assunto do* documento	2

Fonte: Elaborado pela autora.

Quadro 13 – Exemplo de relação semântica que se confirmou em outras frases

Relação semântica	Frases
Indexador *lê o* texto.	"A esse respeito, FARROW (1995) afirma que a indexação back-of-book permite ao leitor localizar informação sobre um tópico dentro do livro; a tarefa do *indexador* é ler o *texto*, distinguir entre informação relevante e periférica e empregar os tipos de processamento de informação presentes na leitura" (NAVES, 2000, p. 26, grifos nossos).
	"Para ter uma competência textual é preciso que, além de conhecer o *texto* que tem em mãos para análise sob todos os aspectos até aqui abordados, o *indexador* faça dele uma leitura adequada, e sabe-se que um *texto* pode gerar muitas leituras, interessando mais, neste estudo, a leitura para fins documentários" (NAVES, 2000, p. 48, grifos nossos).

Relação semântica	Frases
	“para o indexador, há dois tipos de propósitos: (a) para o *indexador* de livros, a tarefa é ler o *texto*, distinguindo entre informações relevantes e periféricas, e empregando uma mistura dos processamentos top-down (conceitual) e bottom-up (perceptivo) obtidos na leitura fluente normal, e (b) para o indexador acadêmico, a indexação é menos exaustiva e usa predominantemente a abordagem top-down” (NAVES, 2000, p. 85, grifos nossos).

Fonte: Elaborado pela autora.

De maneira análoga às relações semânticas que se repetiram em outras frases e contextos, algumas frases tiveram mais de uma relação semântica entre os conceitos da amostra. Isso ocorreu em nove frases das relações explicitadas. O Quadro 14 mostra três exemplos desses casos. No primeiro caso, as relações foram entre os mesmos pares de conceitos; no segundo caso, as relações foram entre pares de conceitos iguais e um diferente; e o último caso denota a frase em que mais relações semânticas foram explicitadas.

Quadro 14 – Frases com múltiplas relações semânticas

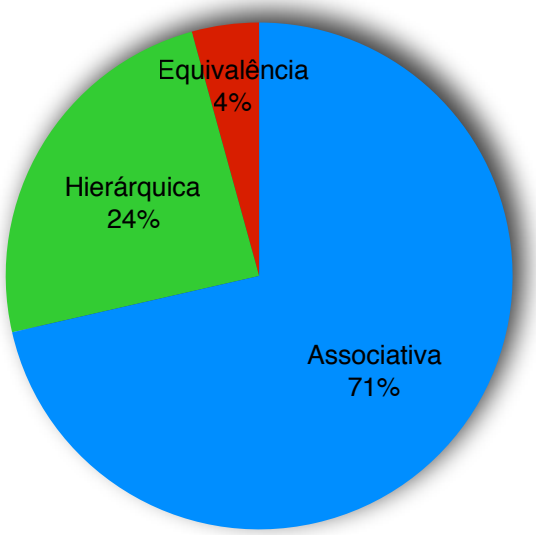
Frases	Relações semânticas
“No ato de pensar, quando faz abstrações, interpreta e define o assunto de um documento, o indexador sofre influência de diversos fatores pertencentes a vários campos, principalmente oriundos da Lingüística, da Ciência Cognitiva e da Lógica” (NAVES, 2000, p. 70).	1. indexador *abstrai assunto do* documento 2. indexador *interpreta assunto do* documento 3. indexador *define assunto do* documento
“Em síntese, este capítulo mostra o processo em que o indexador faz a leitura de um texto, empreende a extração de conceitos e determina a sua atinência”. (NAVES, 2000, p. 69).	1. indexador *extrai* conceito 2. indexador *lê* texto 3. indexador *determina a atinência do* texto
“O sentido geral do texto é baseado na seguinte trilogia estrutural: microestrutura (estrutura superficial, que corresponde à realidade física do texto e seus símbolos de significação, as palavras), macroestrutura (concebida como um tópico representativo hierárquico e coerente da unidade textual, envolvendo mínima estrutura da representação textual sintática-semântica), e a superestrutura (estrutura retórica-esquemática, um tipo de esquema de produção convencional para o qual o texto é adaptado, podendo ser considerado como transição entre estruturas de superfície e de profundidade)” (NAVES, 2000, p. 43).	1. texto *tem tipo de estrutura* microestrutura 2. texto *tem tipo de estrutura* macroestrutura 3. texto *tem tipo de estrutura* superestrutura 4. microestrutura *é assimetricamente contrária a* macroestrutura 5. microestrutura *é assimetricamente contrária a* superestrutura 6. macroestrutura *é assimetricamente contrária a* superestrutura

Fonte: Elaborado pela autora.

6.3.2.3 Características das relações semânticas

A maioria das relações semânticas extraídas foram do tipo associativas, seguida pelas hierárquicas e de equivalência, respectivamente (ver Figura 95).

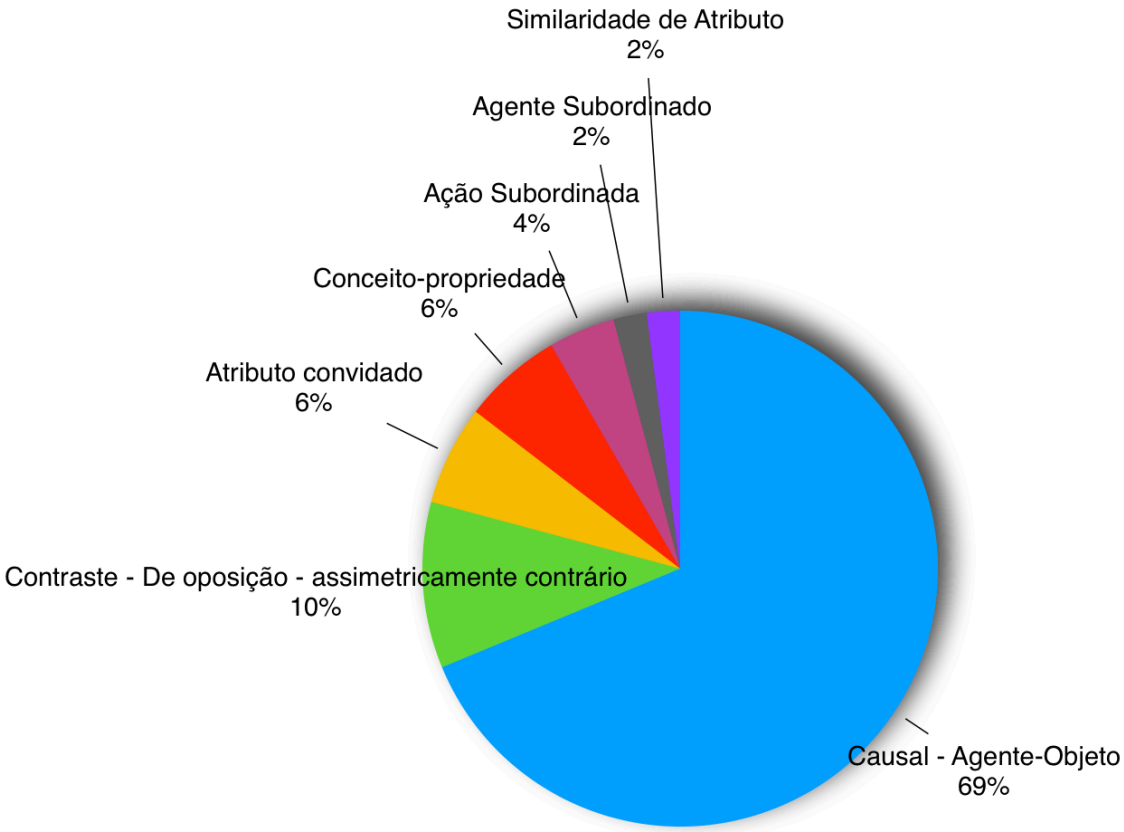
Figura 95 – Gráfico com a porcentagem dos tipos de relações semânticas encontrados



Fonte: Elaborada pela autora.

As relações associativas foram detalhadas em sete subtipos; o subtipo que prevaleceu foi a relação causal agente-objeto, como pode ser visto no gráfico apresentado na Figura 96. Observou-se durante a análise que esse subtipo foi predominante devido ao contexto da amostra, tendo em vista que a tese utilizada aborda o papel do indexador; logo, o indexador foi o agente nas vezes em que esse subtipo de relação aconteceu, enquanto o objeto era o documento ou o texto.

Figura 96 – Gráfico com a porcentagem dos subtipos de relações semânticas associativas encontrados



Fonte: Elaborada pela autora.

O Quadro 15 mostra os subtipos das relações semânticas associativas e os pares de conceitos, como pode ser visto, a relação entre *texto* e *macroestrutura* ocorreu em dois subtipos diferentes, ou seja, em um contexto analisado, *macroestrutura* foi definido como *atributo convidado* de *texto*. Já em outro contexto, *texto* foi determinado como um conceito e *macroestrutura* como uma propriedade de texto. Outra observação a respeito desse quadro é o subtipo agente subordinado, que foi criado para suprir a relação entre *autores* e *indexador*.

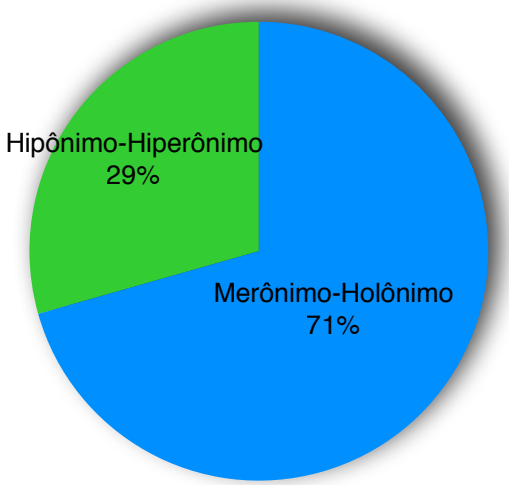
Quadro 15 – Subtipos de relações associativas encontrados

Subtipos de relações semânticas associativas	Pares de conceitos
- Causal - **Agente-Objeto**	• Autores e texto • Indexador e texto • Indexador e documento
- Contraste - De Oposição **- Assimetricamente contrário**	• Microestrutura e macroestrutura • Microestrutura e superestrutura • Macroestrutura e superestrutura • Informativo e narrativo • Primário e secundário
- Atributo convidado	• Bibliotecário e especialização • Bibliotecário e prática • Texto e macroestrutura
- **Conceito-propriedade**	• Texto e macroestrutura • Texto e microestrutura • Texto e superestrutura
- **Ação subordinada**	• Indexador e conceito • Ideia e pensamento
- **Agente subordinado**	• Autores e indexador
- **Similaridade de atributo**	• Especialização e prática

Fonte: Elaborado pela autora.

A respeito das relações hierárquicas, como pode ser visto no gráfico da Figura 97, as relações de merônimo-holônimo foram maioria (71%). Nessas relações, destaca-se que todas foram dos subtipos objeto estruturado componente-complexo. Já nas de hipônimo-hiperônimo (29%), predominou o subtipo inclusão de classe, que ora foi taxonômico, ora funcionalmente subordinado, sendo que o subtipo taxonômico foi maioria. O Quadro 16 especifica os pares de conceitos dos subtipos hierárquicos.

Figura 97 – Gráfico com a porcentagem dos tipos de básicos de relações semânticas hierárquicas encontrados



Fonte: Elaborada pela autora.

Quadro 16 – Subtipos de relações hierárquicas encontrados

Subtipos de relações semânticas hierárquicas	**Pares de conceitos**
- Merônimo-holônimo - Objeto estruturado **- Componente-complexo**	• Conceito e pensamento • Conceito e documento • Conceito e texto • Texto e ideia • Pensamento e documento • Documento e informativo
- Holônimo-hiperônimo - Inclusão de classe **- Taxonômica**	• Texto e informativo • Texto e narrativo • Texto e primário • Texto e secundário
- Hipônimo-hiperônimo - Inclusão de classe **- Funcionalmente subordinado**	• Indexador e profissional da informação

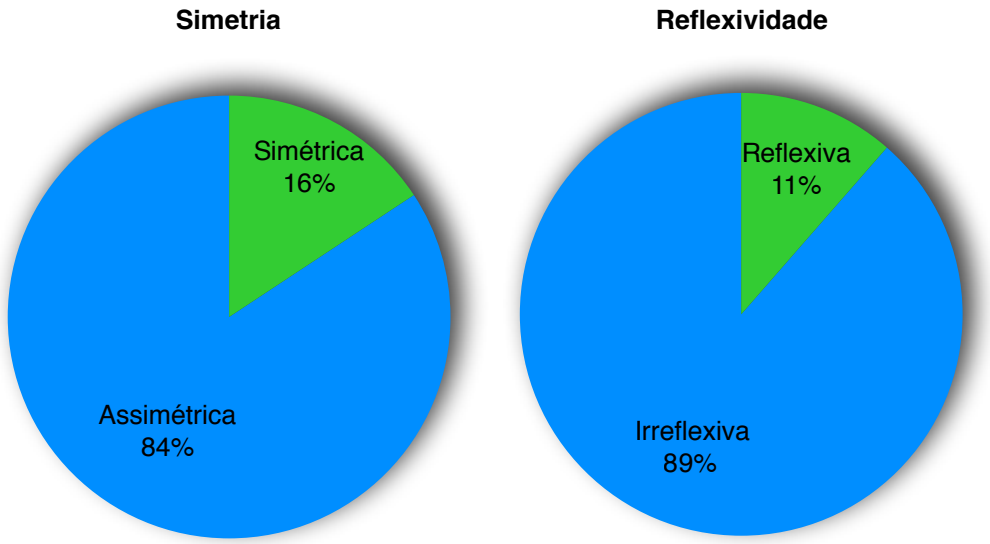
Fonte: Elaborado pela autora.

Por fim, com relação à classificação dos tipos de relações semânticas, foram encontradas apenas três relações de equivalência, que correspondem a 4% das relações. Isso indica que a autora da tese da amostra utilizou poucos sinônimos nos capítulos. Tais relações de equivalência encontradas foram: uma de sinônimo e subtipo sinônimo parcial, cujo par de conceitos é *ideia* e *pensamento*, e três de quase sinônimo, que foram detectadas entre *texto* e *documento*.

6.3.2.3.1 Propriedades das relações semânticas

Na análise das propriedades das relações semânticas explicitadas, observou-se que, quanto à simetria e reflexividade, a maioria das relações semânticas encontradas apresentaram uma maior assimetria e irreflexividade, como mostra a Figura 98.

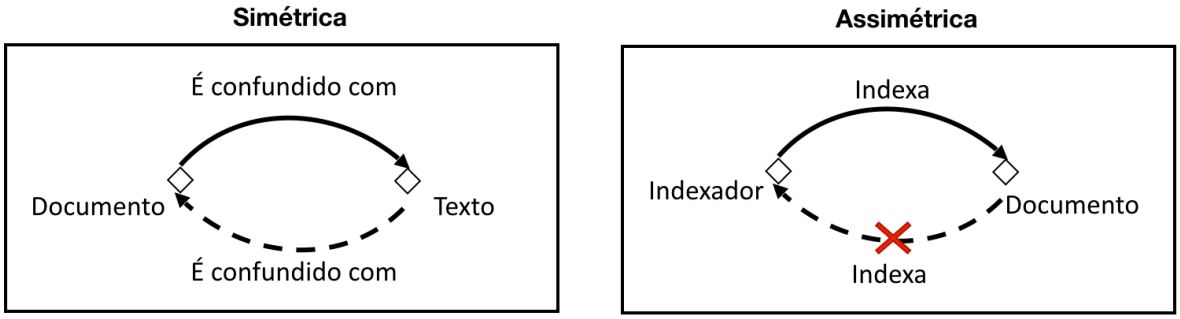
Figura 98 – Propriedades de simetria e reflexividade das relações semânticas explicitadas



Fonte: Elaborada pela autora.

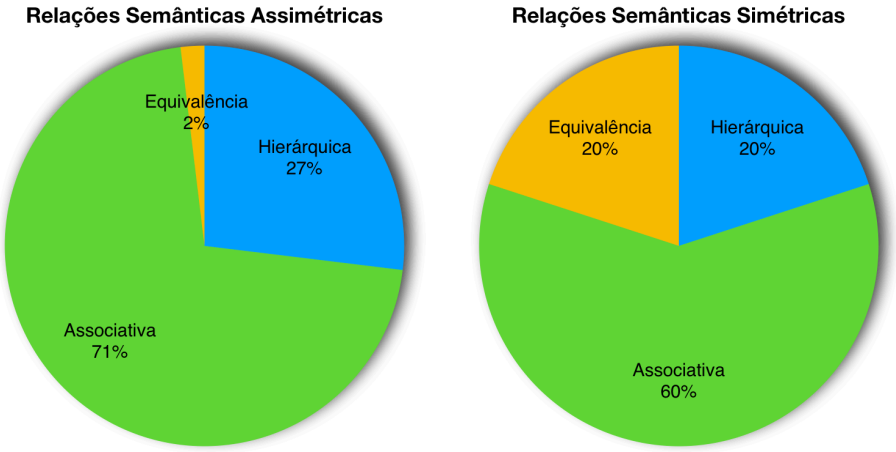
A Figura 99 ilustra as relações quanto à simetria. Percebeu-se que, na determinação da simetria, o tipo básico da relação semântica (hierárquica, equivalência e associativa) não foi o fator resolutivo, pois tanto para as relações simétricas quanto para as assimétricas ocorreram os tipos supracitados, como pode ser visto nos gráficos da Figura 100.

Figura 99 – Exemplos de relações semânticas simétrica e assimétrica



Fonte: Elaborada pela autora.

Figura 100 – Simetria das relações semânticas de acordo com os tipos de relações semânticas



Fonte: Elaborada pela autora.

A análise desses gráficos não permite afirmar qual tipo de relação semântica é mais predominante na propriedade de simetria, pois, no âmbito desse estudo de caso, as relações semânticas explicitadas foram em sua maioria associativas; logo, esse resultado influencia todos os outros cuja análise utiliza os tipos de relações semânticas. Contudo, observa-se que a porcentagem de relações de equivalência aumentou nas relações simétricas, como era esperado, apesar de apenas um par de conceitos ter sido simétrico, como pode ser visto no Quadro 17. O Quadro 18 mostra os tipos e subtipos de relações semânticas e os pares de conceitos considerados assimétricos.

Quadro 17 – Tipos e subtipos de relações semânticas e pares de conceitos classificados como simétricos

Tipo	Subtipo	Pares de conceitos
Hierárquica	- Holônimo-Hiperônimo - Inclusão de Classe **- Taxonômica**	• Texto e informativo • Texto e narrativo • Texto e primário • Texto e secundário
Equivalência	**- Quase sinônimo**	• Texto e documento
Associativa	- Contraste - De Oposição **- Assimetricamente contrário**	• Microestrutura e macroestrutura • Microestrutura e superestrutura • Macroestrutura e superestrutura • Informativo e narrativo • Primário e secundário
	- **Similaridade de atributo**	• Especialização e prática

Fonte: Elaborado pela autora.

Quadro 18 – Tipos e subtipos de relações semânticas e pares de conceitos classificados como assimétricos

Tipo	Subtipo	Pares de conceitos
Hierárquica	- Merônimo-Holônimo - Objeto Estruturado **- Componente-Complexo**	• Conceito e pensamento • Conceito e documento • Conceito e texto • Texto e ideia • Pensamento e documento • Documento e informativo
	- Hipônimo-Hiperônimo - Inclusão de Classe **- Funcionalmente Subordinado**	• Indexador e profissional da informação
Equivalência	**- Sinônimo parcial**	• Ideia e pensamento
Associativa	- Causal - **Agente-Objeto**	• Autores e texto • Indexador e texto • Indexador e documento
	- Atributo convidado	• Bibliotecário e especialização • Bibliotecário e prática • Texto e macroestrutura
	- **Conceito-Propriedade**	• Texto e macroestrutura • Texto e microestrutura • Texto e superestrutura
	- **Ação subordinada**	• Indexador e conceito • Ideia e pensamento
	- **Agente subordinado**	• Autores e indexador

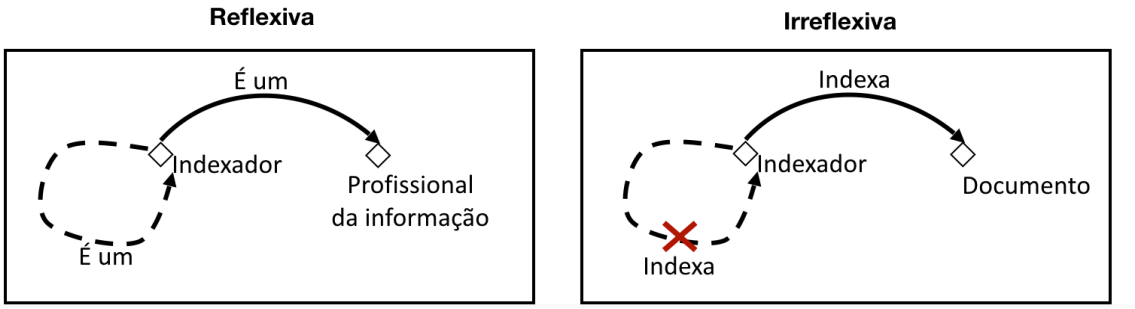
Fonte: Elaborado pela autora.

Como pode ser observado nos Quadros 17 e 18, nenhum subtipo de relação semântica ou par de conceitos foi classificado como ora simétrico, ora assimétrico.

Isso indica consistência na determinação dessa propriedade. Além disso, a simetria não dependeu do contexto em que os pares de conceitos estavam.

Como notou-se no gráfico da Figura 98, quanto à reflexividade, as relações semânticas irreflexivas foram a maioria. A Figura 101 exemplifica relações reflexivas e irreflexivas. Assim como na assimetria, as relações associativas predominaram na análise da reflexividade, isso porque essas relações foram maioria nos resultados desse estudo de caso.

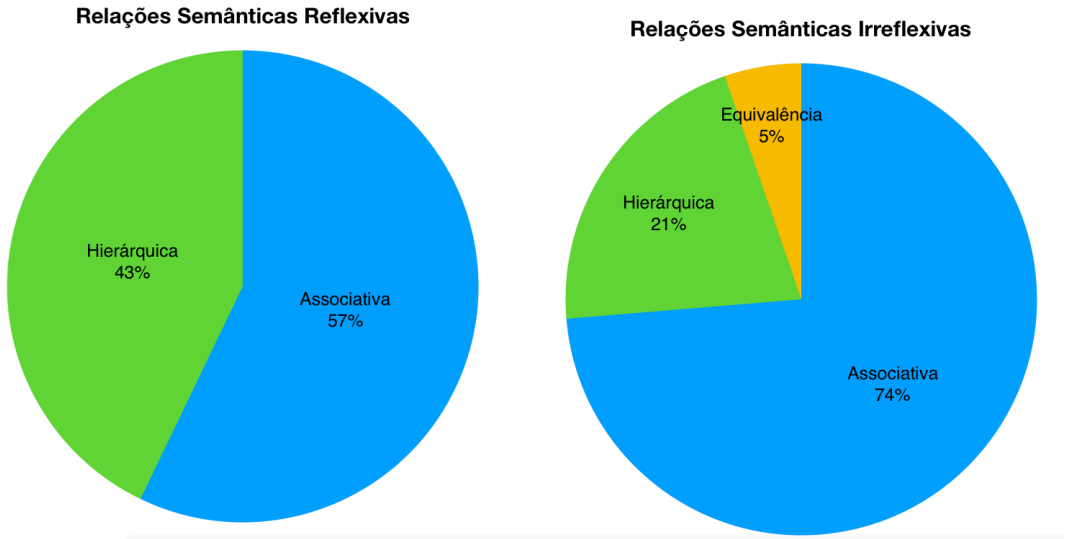
Figura 101 – Exemplos de relações semânticas reflexiva e irreflexiva



Fonte: Elaborado pela autora.

A Figura 102 mostra dois gráficos com a porcentagem de relações semânticas reflexivas e irreflexivas de acordo com o tipo. Observa-se, portanto, que não ocorreram relações de equivalência que fossem reflexivas. Os Quadros 19 e 20 mostram os tipos e subtipos de relações semânticas juntamente com os pares de conceitos classificados como reflexivos e irreflexivos, respectivamente.

Figura 102 – Reflexividade das relações semânticas de acordo com os tipos de relações semânticas



Fonte: Elaborada pela autora.

Quadro 19 – Tipos e subtipos de relações semânticas e pares de conceitos classificados como reflexivos

Tipo	Subtipo	Pares de conceitos
Hierárquica	- Holônimo-hiperônimo - Inclusão de classe **- Taxonômica**	• Texto e informativo • Texto e narrativo
	- Holônimo-hiperônimo - Inclusão de classe **- Funcionalmente subordinado**	• Indexador e profissional da Informação
Associativa	**- Ação subordinada**	• Ideia e pensamento
	- **Conceito-propriedade**	• Texto e macroestrutura • Texto e microestrutura • Texto e superestrutura

Fonte: Elaborado pela autora.

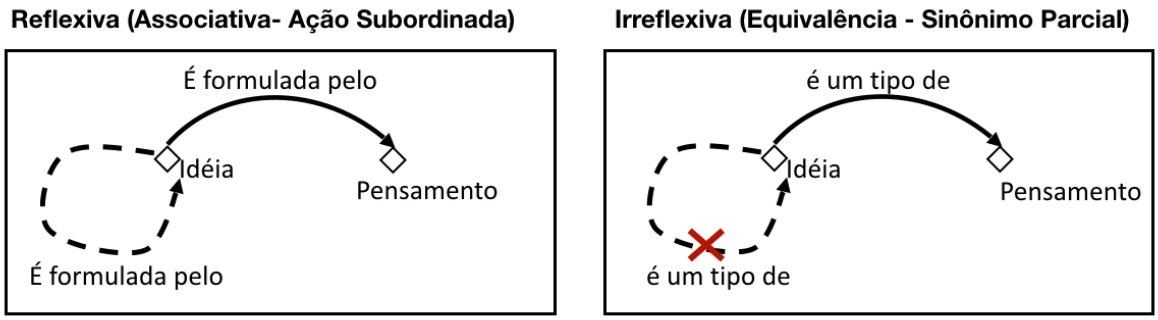
Quadro 20 – Tipos e subtipos de relações semânticas e pares de conceitos classificados como irreflexivos

Tipo	Subtipo	Pares de conceitos
Hierárquica	- Merônimo-holônimo - Objeto estruturado **- Componente-complexo**	• Conceito e pensamento • Conceito e documento • Conceito e texto • Texto e ideia • Pensamento e documento • Documento e informativo
Equivalência	**- Sinônimo parcial**	• Ideia e pensamento
	- Quase sinônimo	• Texto e documento
Associativa	- Causal - **Agente-objeto**	• Autores e texto • Indexador e texto • Indexador e documento
	- Atributo convidado	• Bibliotecário e especialização • Bibliotecário e prática • Texto e macroestrutura
	- Contraste - De Oposição **- A s s i m e t r i c a m e n t e contrário**	• Microestrutura e macroestrutura • Microestrutura e superestrutura • Macroestrutura e superestrutura • Informativo e narrativo • Primário e secundário
	- **Ação subordinada**	• Indexador e conceito
	- **Agente subordinado**	• Autores e indexador
	- **Similaridade de atributo**	• Especialização e prática

Fonte: Elaborado pela autora.

Diferentemente da propriedade de simetria, a reflexividade dos tipos de relações semânticas e dos pares de conceitos dependeu do contexto. Isso foi percebido no tipo de relação semântica associativo – ação subordinada em que, para o par de conceitos *indexador* e *conceito*, a relação é irreflexiva, mas, para o par de conceitos *ideia* e *pensamento*, a relação é reflexiva. Para esse último caso, observou-se ainda que esse par de conceitos manifestou-se diferentemente quanto à reflexividade. Nesse sentido, enquanto ação subordinada, ele foi reflexivo, porém, enquanto sinônimo parcial, ele foi irreflexivo. A partir disso, na amostra do estudo de caso, a reflexividade das relações semânticas, nesses casos, dependeu do contexto. A Figura 103 ilustra a reflexividade e irreflexividade do par de conceitos *ideia* e *pensamento.*

Figura 103 – Reflexividade do par de conceitos *ideia* e *pensamento*.



Fonte: Elaborada pela autora.

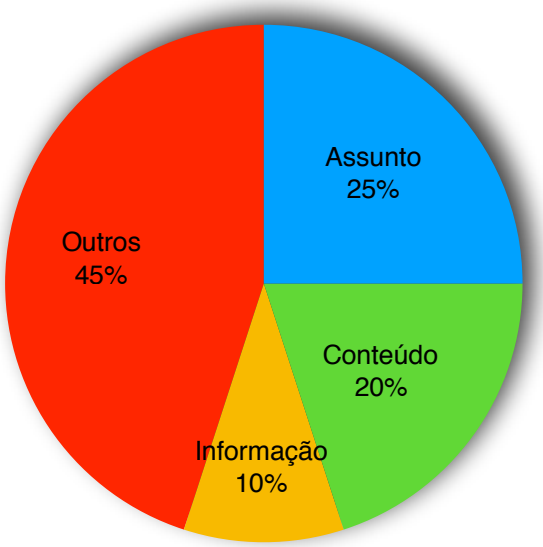
Nem todas as relações semânticas encontradas tiveram relação inversa. Isso pode ser visto no gráfico apresentado na Figura 104, em que quase metade das relações semânticas encontradas não apresentaram essa propriedade.

Figura 104 – Gráfico com a porcentagem de relações semânticas que podem ser invertidas ou não

Fonte: Elaborada pela autora.

Em análise sobre as relações semânticas que não puderam ser explicitadas inversamente (53%), constatou-se que na maioria delas as relações entre os sujeitos e objetos não era direta, como, por exemplo: conceito *sintetiza conteúdo do* texto. Nesse exemplo, *conceito* não sintetiza diretamente o *texto* e sim o conteúdo dele. Logo, a relação inversa mais adequada seria: conteúdo do texto *é sintetizado pelo* conceito. Portanto, o conceito que está sendo analisado é *texto* e não *conteúdo do texto*, por isso a relação inversa não foi possível. Dessas relações semânticas que não puderam ser explicitadas inversamente devido à relação indireta, observou-se que existiam entre os dois conceitos relacionados os conceitos *assunto*, *conteúdo* ou *informação*. Logo, acredita-se que esses conceitos podem ser representativos para o contexto e que eles deveriam compor a amostra de conceitos da estrutura classificatória. O gráfico da Figura 105 mostra a porcentagem em que esses conceitos foram encontrados nas relações semânticas que não puderam ser determinadas inversamente. Já o Quadro 21 mostra essas relações.

Figura 105 – Porcentagem de conceitos que apareceram entre os conceitos das relações semânticas



Fonte: Elaborada pela autora.

Quadro 21 – Relações semânticas que não puderam ser invertidas

Conceitos	**Relações semânticas que não puderam ser invertidas**
Assunto	• indexador *decide assunto do* documento • indexador *abstrai assunto do* documento • indexador *interpreta assunto do* documento • indexador *define assunto do* documento • conceito *sintetiza assunto do* documento
Conteúdo	• autores *determinam conteúdo do* texto • conceito *representa conteúdo do* documento • conceito *sintetiza conteúdo do* texto • documento *tem conteúdo* informativo
Informação	• Indexador *organiza informação do* texto • Indexador *trata informações do* texto
Outros	• Indexador *elabora índices de* texto. • Indexador *constrói sentido do* texto. • Indexador *determina assunto geral do* texto. • Indexador *conhece estrutura do* texto. • Indexador *conhece tipo do* texto. • Indexador *faz leitura específica de* texto. • Indexador *determina atinência do* texto. • indexador *faz leitura documentária do* documento • Conceito *representa conteúdo temático do* texto.

Fonte: Elaborado pela autora.

Ainda sobre as relações inversas, observou-se que em 13% delas a estrutura da relação semântica entre os conceitos era direta, ou seja, entre o sujeito e o objeto não havia outros elementos gramaticais. Contudo, a semântica não permitiu a determinação das relações inversas, como foi o caso das relações: conceito *é extraído do* documento; conceito *é traduzido do* documento; e conceito *é extraído do*

texto. Observa-se, nesses casos, que os verbos estão no particípio, o que geralmente é característica de relações inversas, mesmo elas não sendo inversas no contexto em que foram explicitadas. O Quadro 22 mostra o contexto que essas relações foram detectadas.

Quadro 22 – Relações semânticas que não puderam ser invertidas devido à semântica da relação inversa

Relação semântica	**Contextos**
Conceito *é extraído do* documento	“O processo de ler um *documento* para extrair *conceitos* que traduzam a sua essência é conhecido como ’Análise de assunto’, para alguns, como análise temática, para outros, ou, ainda, como análise documentária, análise conceitual ou, mesmo, análise de conteúdo” (NAVES, 2000, p. 35, grifos nossos).
	“No entanto, a concepção orientada pela demanda já pode ser vista como uma fase posterior à Análise de assunto propriamente dita, considerando ser essa a etapa em que a preocupação é traduzir os *conceitos* extraídos do *documento* para os termos de uma linguagem de indexação” (NAVES, 2000, p. 35, grifos nossos).
	“Verifica-se, no exame da literatura especializada em Biblioteconomia e Ciência da Informação, que o termo ‘Análise de assunto’ é o mais comumente utilizado, mas que grande parte dos autores que tratam do tema o consideram ou como a etapa de tradução dos *conceitos* extraídos dos *documentos* para um vocabulário controlado, ou até mesmo do processo de indexação como um todo” (NAVES, 2000, p. 38, grifos nossos).
Conceito *é traduzido do* documento.	“No entanto, a concepção orientada pela demanda já pode ser vista como uma fase posterior à Análise de assunto propriamente dita, considerando ser essa a etapa em que a preocupação é traduzir os *conceitos* extraídos do *documento* para os termos de uma linguagem de indexação” (NAVES, 2000, p. 35, grifos nossos).
Conceito *é extraído do* texto.	“Para definir em termos adequados o assunto de um *texto* é necessário que primeiro se extraiam os *conceitos* que nele estão contidos” (NAVES, 2000, p. 54, grifos nossos).
	“No processo de extrair *conceitos* de *textos* para definir seu assunto, o silogismo ocorre sempre, pois tanto a dedução quanto a inferência nele estão presentes” (NAVES, 2000, p. 87, grifos nossos)
	“Em síntese, este capítulo mostra o processo em que o indexador faz a leitura de um *texto*, empreende a extração de *conceitos* e determina a sua atinência” (NAVES, 2000, p. 69, grifos nossos)

Fonte: Elaborado pela autora.

À parte dessa questão levantada, observou-se que as relações semânticas determinadas como simétricas automaticamente tinham relações semânticas inversas. Logo, a simetria das relações semânticas, nos contextos analisados, foi um fator determinante para o estabelecimento das relações inversas.

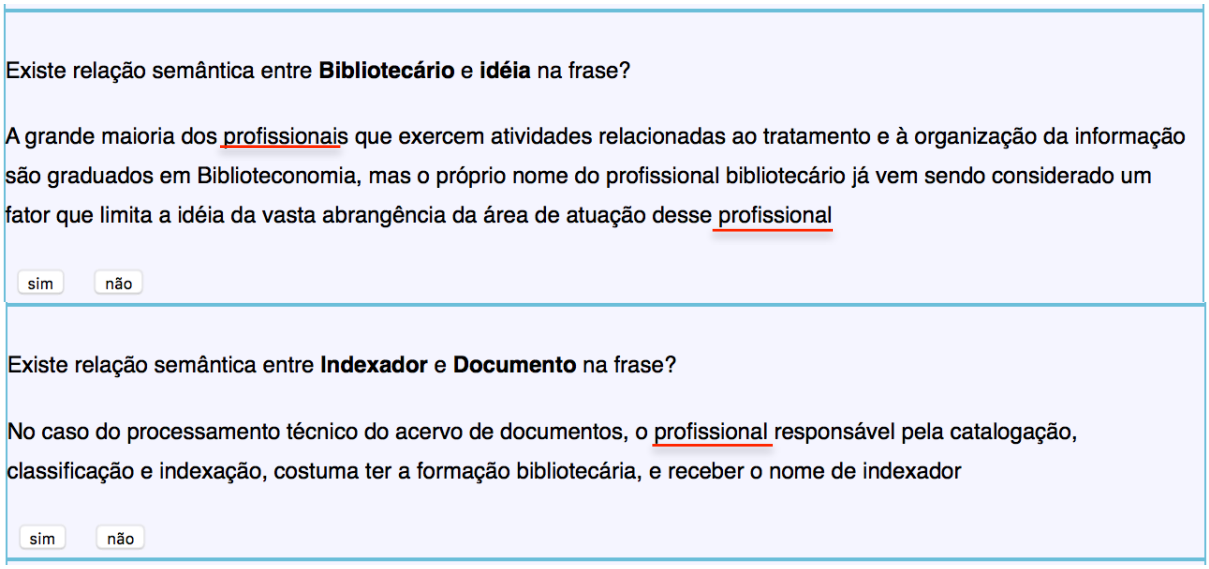
6.4 Considerações sobre a amostra

Percebeu-se, pela análise dos estudo de caso, que alguns cuidados deveriam ter sido tomados com relação à amostra. Nesse sentido, percebeu-se que o recorte realizado a partir da tese de Naves (2000) deveria considerar os Capítulos 5, 6 e 7, pois os conceitos *indexador experiente*, *indexador pouco experiente* e *indexador novato*, não identificados pelo Semantizar, estavam predominantemente nesses capítulos. Logo, a exclusão desses capítulos na amostra impactou no resultado das relações semânticas explicitadas.

No que diz respeito aos conceitos *autores* e *narrativos*, como pode ser visto, eles foram determinados na estrutura facetada genuinamente no plural. Desse modo, quando esses conceitos ocorreram no singular no documento acadêmico, eles não foram detectados. Por exemplo, na frase "Outro aspecto que merece ser ressaltado é que o *autor* do *texto*, ao escrevê-lo, tem em mente um determinado leitor alvo para o qual direciona suas idéias […]" (NAVES, 2000, p. 53, grifos nossos), a relação entre *autor* e *texto* destacada não foi encontrada.

Ainda sobre a amostra, o conceito *profissional da informação* não foi representativo para a identificação de conceitos, pois, muitas vezes, apenas a palavra *profissional* cumpria a função de significar *profissional da informação*, devido ao contexto, conforme mostra os exemplos da Figura 106. Outrossim, caso essa palavra ocorresse no plural, ela não seria encontrada, pois seu plural não é constituído apenas da letra “s” no final (escreve-se *profissionais* e não “*profissionals*”). No Semantizar não foi implementada uma função para tratar esse erro.

Figura 106 – Exemplos em que a palavra profissional poderia ser detectada pelo Semantizar caso fosse tratada na amostra



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

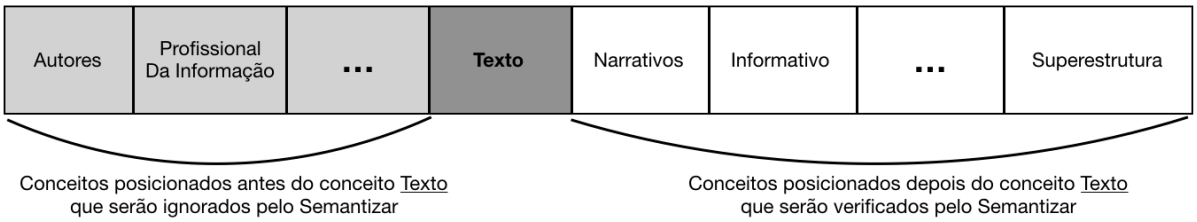
6.5 Considerações sobre o Semantizar

O Modelo de Extração de Relações Semânticas, implementado no Semantizar, evidencia no contexto da Biblioteconomia e Ciência da Informação, uma nova forma de determinar e explicitar relações semânticas, e, como tal, carece de aperfeiçoamentos, pois, como toda pesquisa, o Modelo não é inerte. Dessa forma, investigações serão realizadas tanto para melhorar o Modelo e o Semantizar quanto para obter contribuições que não tenham sido observadas nesta tese. Nesse sentido, esta seção trata especificamente de apresentar as correções, testes e aperfeiçoamentos que precisam ser realizados no Semantizar.

Como foi constatado, o Semantizar detectou inicialmente 48 pares de conceitos. Um número bem expressivo, dado a amostra com 22 conceitos. Porém, uma falha no código impediu que o Semantizar identificasse mais pares de conceitos. Conforme relatado na Seção 5.1, o algoritmo do Semantizar transforma a estrutura classificatória em um vetor. Quando um conceito do vetor é encontrado em uma frase, a posição do conceito no vetor é fixada e o algoritmo busca nas próximas posições desse mesmo vetor se existe outro conceito na mesma frase. Contudo, do modo como está implementado, o algoritmo não retorna ao início do vetor em uma frase em

que um conceito é encontrado. Por exemplo, a relação entre *autores* e *texto* foi encontrada; por outro lado, entre *texto* e *autores* não foi possível detectar a relação, pois os conceitos localizados antes de *texto* não são verificados quanto ao fato de existirem na mesma frase. Assim, como pode ser visto na Figura 107, que ilustra o vetor de conceitos da estrutura classificatória, ao identificar em uma frase o conceito *texto*, somente as próximas posições são verificadas. Essa falha indica que, se os conceitos estivessem em outra ordem, outras relações semânticas poderiam ser encontradas.

Figura 107 – Vetor de conceitos da estrutura classificatória apresentando a falha do algoritmo ao não verificar as posições anteriores do vetor



Fonte: Elaborada pela autora.

Com base nisso, um teste foi realizado com a estrutura classificatória organizada em ordem alfabética. Nesse teste, detectou-se a mesma quantidade de indícios de relações semânticas (199), mas alguns pares de conceitos identificados foram díspares e a quantidade de indícios de relações semânticas quando os conceitos estavam em ordem inversa (porque estavam em ordem alfabética na estrutura classificatória) também foram diferentes. O Quadro 23 mostra alguns exemplos de indícios de pares de conceitos encontrados quando a estrutura não estava em ordem alfabética (ou seja, conforme a estrutura facetada) e quando estava organizada em ordem alfabética.

Quadro 23 – Relações do conceito *autores* com outros conceitos em diferentes posições da amostra

Sujeito	Objeto	Sem ordem	Ordem alfabética
Conceito	Prática	0	1
Prática	Conceito	2	0
Conceito	Indexador	0	10
Indexador	Conceito	12	0
Ideia	Documento	1	0

Sujeito	Objeto	Sem ordem	Ordem alfabética
Documento	Ideia	0	2
Documento	Indexador	0	31
Indexador	Documento	28	0
Documento	Pensamento	0	1
Pensamento	Documento	3	0
Ideia	Indexador	0	4
Indexador	Ideia	3	0
Primário	Texto	0	1
Texto	Primário	2	0

Fonte: Elaborado pela autora.

O Quadro 24 mostra alguns pares de conceitos não identificados e as respectivas frases em que existia relação semântica entre eles.

Quadro 24 – Pares de conceitos não identificados pelo Semantizar

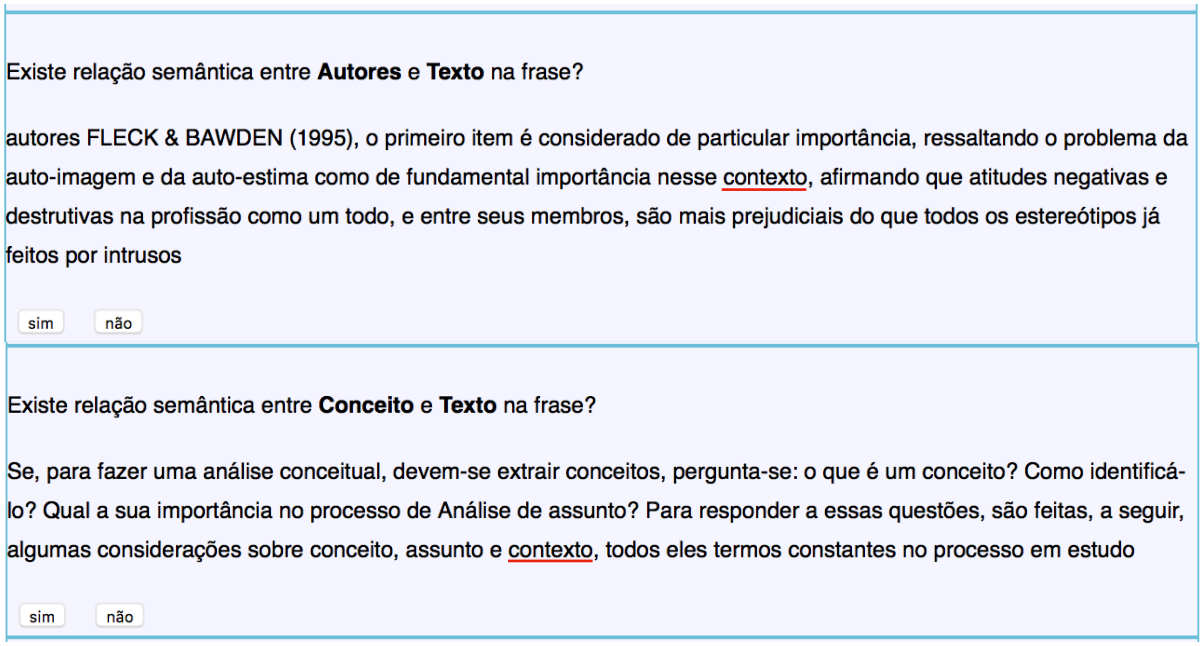
Par de conceitos	Frase
indexador e *texto*	“Concepção orientada para o conteúdo - envolve uma interpretação adicional do conteúdo, que vai além dos limites da estrutura léxica e gramatical, com o estabelecimento de assuntos que não estão explicitamente colocados no *texto*, mas que são facilmente identificados pelo *indexador*, envolvendo, portanto, uma abstração mais indireta do documento” (NAVES, 2000, p. 61, grifos nossos).
texto e *informativo*	“A tipologia mais comumente utilizada divide os *textos* em narrativos e *informativos*” (NAVES, 2000, p. 44, grifos nossos).
texto e *secundário*	“O *texto* original é chamado texto primário, um sumário ou resumo, texto *secundário*, e a expressão do texto primário numa linguagem documentária, texto terciário” (NAVES, 2000, p. 48, grifos nossos).
Indexador e *texto*	“Outro aspecto que merece ser ressaltado é que o autor do *texto*, ao escrevê-lo, tem em mente um determinado leitor alvo para o qual direciona suas idéias; suas intenções não são dirigidas para o leitor/*indexador* e não lhe interessa se esse vai ter capacidade para interpretar as informações que aquele texto está veiculando” (NAVES, 2000, p. 53).
conceito e *texto*	“Nesta fase da determinação da atinência para representar os *conceitos* extraídos do *texto*, inicia-se um processo lingüístico e o problema de descrever documentos para recuperação é, principalmente, o problema de como a linguagem é usada” (NAVES, 2000, p. 67).

Fonte: Elaborado pela autora.

O Semantizar empregou funções PHP desenvolvidas por comunidade de desenvolvedores *web* que utilizam o PHP. Uma dessas funções é a *srtipos*, que

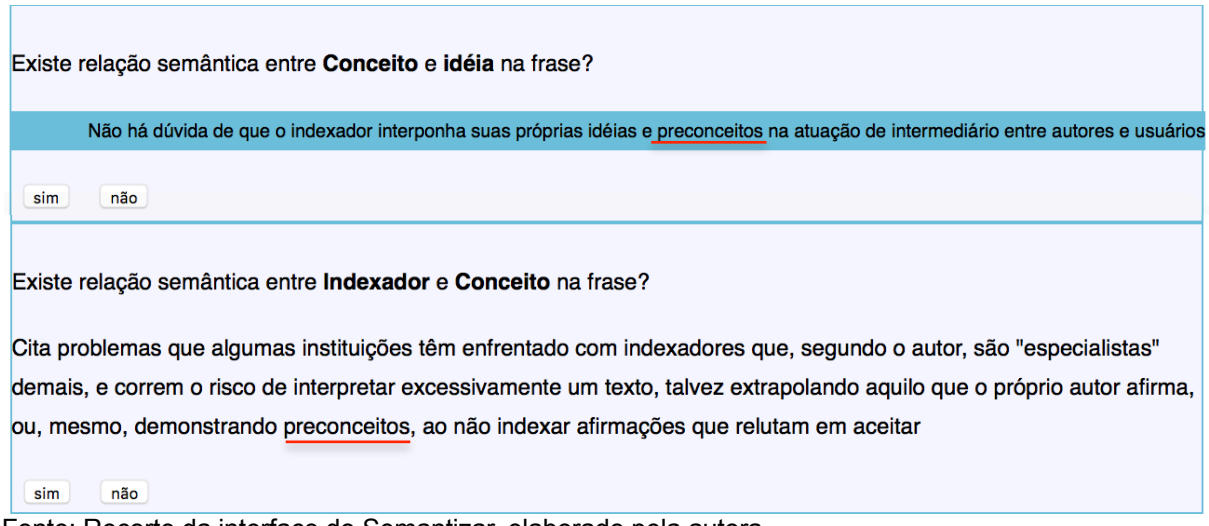
implementa a busca de *strings.* Contudo, nas buscas de *strings* realizadas, as palavras formadas pelos conceitos da amostra o Semantizar contava como o próprio conceito, como, por exemplo, as palavras con*texto* (que contém a palavra *texto*) e pre*conceito* (que contém a palavra *conceito*). Nessas situações, essas duas palavras foram detectadas pelo sistema mesmo sem constarem na amostra da estrutura classificatória. Outrossim, elas geraram um resultado falso-positivo, ou seja, o Semantizar detectou o indício, mas ele não existia de fato. As Figuras 108 e 109 mostram exemplos.

Figura 108 – Frases em que o Semantizar encontrou *texto*, mas a palavra na frase era *contexto*



Fonte: Recorte da interface do Semantizar, elaborado pela autora.

Figura 109 – Frases em que o Semantizar encontrou conceito, mas a palavra na frase era *preconceito*

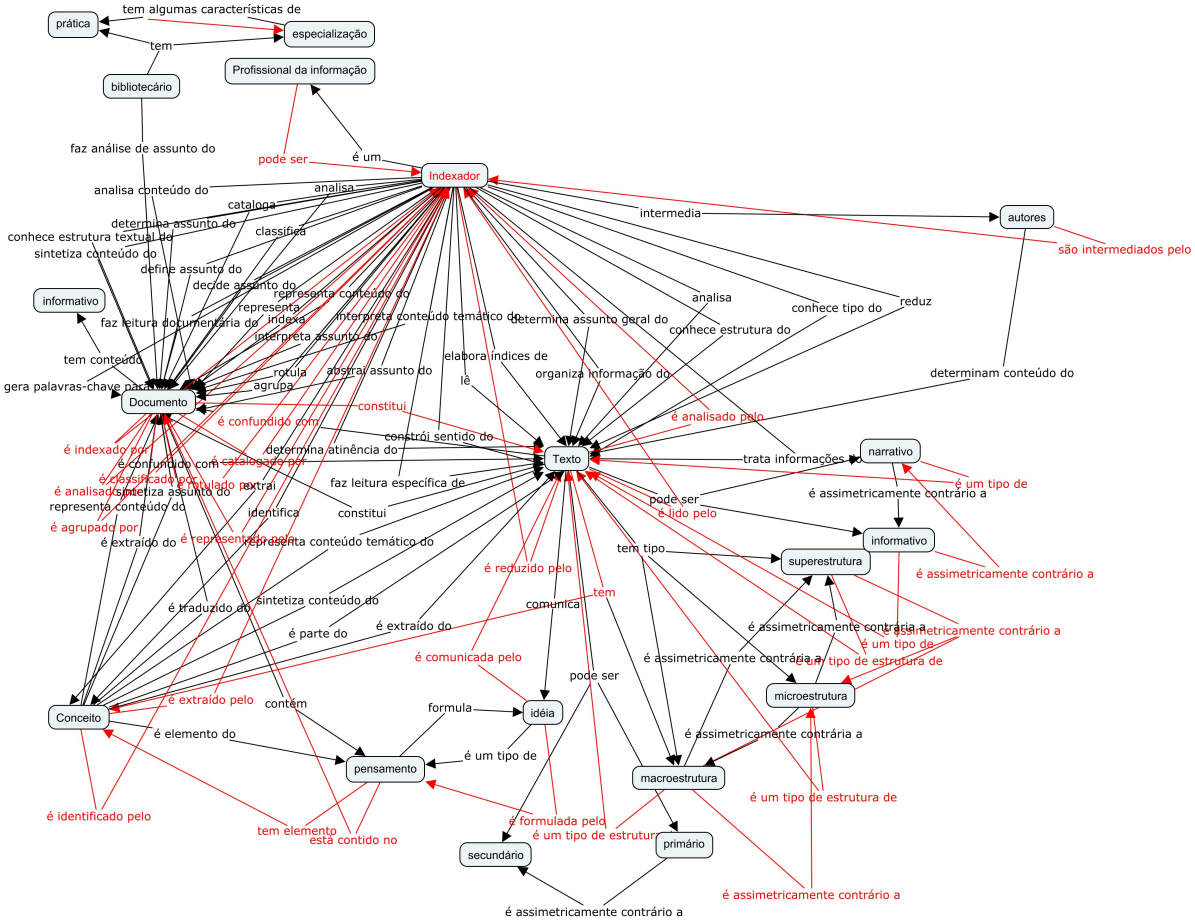


Fonte: Recorte da interface do Semantizar, elaborado pela autora.

7 RESULTADOS

O estudo de caso realizado elucidou importantes resultados para a extração e explicitação de relações semânticas. O primeiro deles é esboçado em um mapa conceitual (Figura 110) de compilação das relações semânticas explicitadas. Nesse mapa, as relações semânticas destacadas na cor vermelha são relações inversas. Devido à complexidade do mapa conceitual, pois 101 relações semânticas foram explicitadas, foi elaborado o Quadro 25 com todas as relações semânticas.

Figura 110 – Mapa conceitual das relações semânticas explicitadas no estudo de caso



Fonte: Elaborada pela autora.

Quadro 25 – Relações semânticas entre os conceitos

Sujeito	Predicado	Objeto
Indexador	intermedeia	autores
Autores	são intermediados pelo	indexador
Autores	determinam conteúdo do	texto

Sujeito	Predicado	Objeto
Texto	tem	conceito
Conceito	é parte do	texto
Conceito	é extraído do	texto
Conceito	sintetiza conteúdo do	texto
Conceito	representa conteúdo temático do	texto
Conceito	representa conteúdo do	documento
Documento	constitui	texto
Texto	constitui	documento
Texto	é confundido com	documento
Documento	é confundido com	texto
Conceito	sintetiza assunto do	documento
Conceito	é traduzido do	documento
Conceito	é extraído do	documento
Indexador	identifica	conceito
Conceito	é identificado pelo	indexador
Indexador	extrai	conceito
Conceito	é extraído pelo	indexador
Indexador	agrupa	documento
Documento	é agrupado por	indexador
Indexador	rotula	documento
Documento	é rotulado por	indexador
Indexador	representa	documento
Documento	é representado pelo	indexador
Indexador	indexa	documento
Documento	é indexado por	indexador
Indexador	analisa	documento
Documento	é analisado por	indexador
Indexador	cataloga	documento
Documento	é catalogado por	indexador
Indexador	classifica	documento
Documento	é classificado por	indexador
Documento	tem conteúdo	informativo
Informativo	é um tipo de	texto
Texto	pode ser	informativo

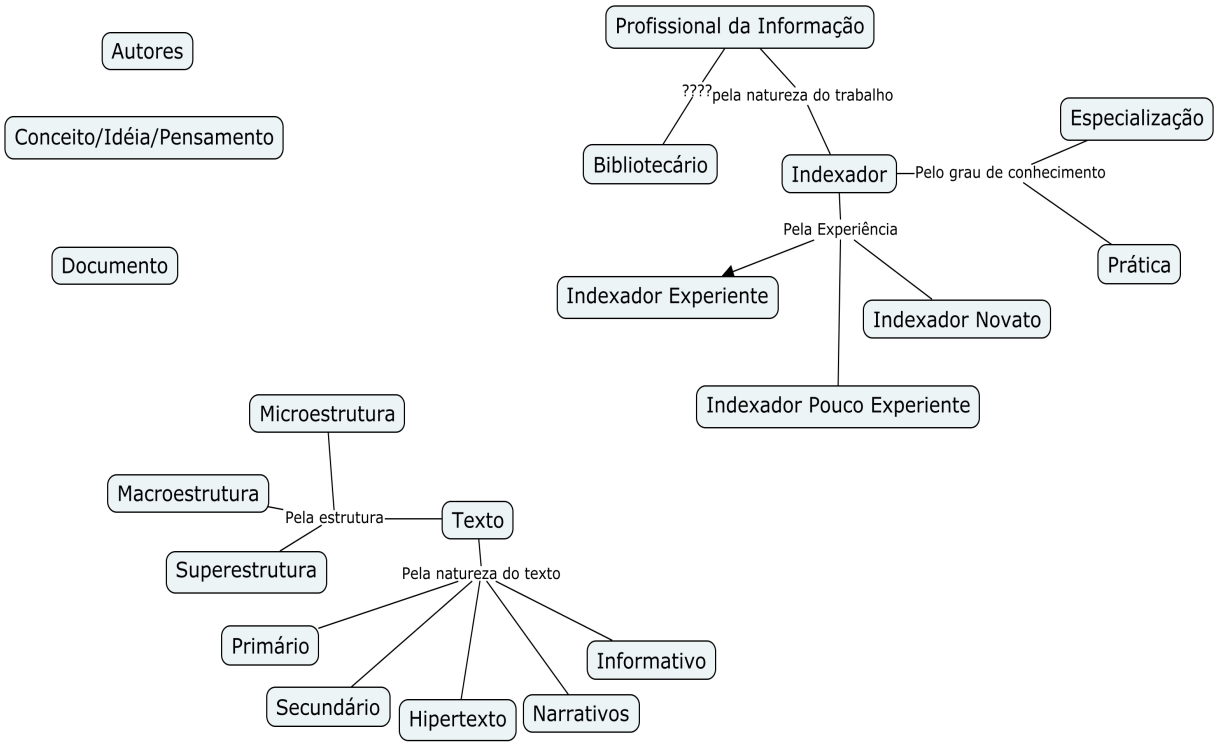
Sujeito	Predicado	Objeto
Texto	pode ser	narrativo
Narrativo	é um tipo de	texto
Informativo	é assimetricamente contrário a	narrativo
Narrativo	é assimetricamente contrário a	informativo
Documento	contém	pensamento
Pensamento	está contido no	documento
Conceito	é elemento do	pensamento
Pensamento	tem elemento	conceito
Pensamento	formula	ideia
Ideia	é formulada pelo	pensamento
Ideia	é um tipo de	pensamento
Texto	comunica	ideia
Ideia	é comunicada pelo	texto
Indexador	interpreta assunto do	documento
Indexador	define assunto do	documento
Indexador	analisa conteúdo do	documento
Indexador	determina assunto do	documento
Indexador	sintetiza conteúdo do	documento
Indexador	decide assunto do	documento
Indexador	gera palavras-chave para	documento
Indexador	representa conteúdo do	documento
Indexador	abstrai assunto do	documento
Indexador	conhece estrutura textual do	documento
Indexador	faz leitura documentária do	documento
Indexador	interpreta conteúdo temático do	documento
Indexador	determina assunto geral do	texto
Indexador	constrói sentido do	texto
Indexador	determina atinência do	texto
Indexador	trata informações do	texto
Indexador	conhece estrutura do	texto
Indexador	conhece tipo do	texto
Indexador	elabora índices de	texto
Indexador	reduz	texto
Texto	é reduzido	indexador

Sujeito	Predicado	Objeto
Indexador	analisa	texto
Texto	é analisado pelo	indexador
Indexador	lê	texto
Texto	é lido pelo	indexador
Indexador	faz leitura específica de	texto
Indexador	organiza informação do	texto
Texto	tem tipo	microestrutura
Microestrutura	é um tipo de estrutura de	texto
Texto	tem tipo	macroestrutura
Macroestrutura	é um tipo de estrutura de	texto
Texto	tem tipo	superestrutura
Superestrutura	é um tipo de estrutura de	texto
Macroestrutura	é assimetricamente contrário a	microestrutura
Macroestrutura	é assimetricamente contrária a	superestrutura
Microestrutura	é assimetricamente contrária a	macroestrutura
Microestrutura	é assimetricamente contrária a	superestrutura
Superestrutura	é assimetricamente contrário a	microestrutura
Superestrutura	é assimetricamente contrário a	macroestrutura
Texto	tem atributo de	macroestrutura
Texto	pode ser	primário
Texto	pode ser	secundário
Primário	é assimetricamente contrário a	secundário
Secundário	é assimetricamente contrário a	primário
Bibliotecário	tem	especialização
Bibliotecário	tem	prática
Especialização	tem algumas características de	prática
Prática	tem algumas características de	especialização
Bibliotecário	faz análise de assunto do	documento
Profissional da informação	pode ser	indexador
Indexador	é um	profissional da informação

Fonte: Elaborado pela autora.

A Figura 111 é um mapa conceitual gerado a partir da estrutura classificatória sem o intermédio do Semantizar. O Quadro 26, detalha as 15 relações semânticas existentes *a priori* na estrutura classificatória original.

Figura 111 – Mapa conceitual da estrutura classificatória



Fonte: Elaborada pela autora.

Quadro 26 – Relações semânticas da estrutura classificatória original

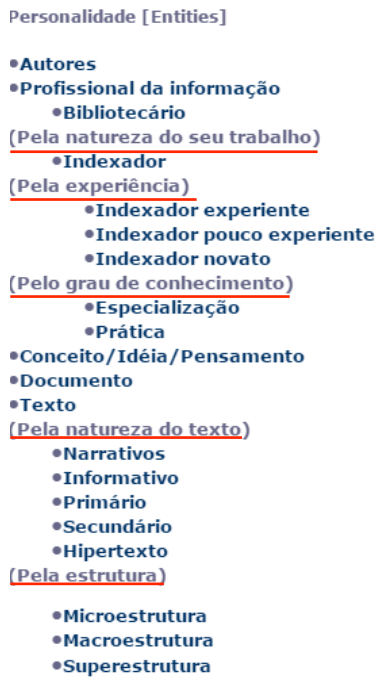
Sujeito	Predicado	Objeto
Indexador	Pelo grau de conhecimento	Especialização
Indexador	Pelo grau de conhecimento	Prática
Indexador	Pela Experiência	Indexador Experiente
Indexador	Pela Experiência	Indexador Novato
Indexador	Pela Experiência	Indexador Pouco Experiente
Profissional da Informação	????	Bibliotecário
Profissional da Informação	Pela natureza do trabalho	Indexador
Texto	Pela natureza do texto	Hipertexto
Texto	Pela natureza do texto	Informativo
Texto	Pela natureza do texto	Narrativos
Texto	Pela natureza do texto	Primário
Texto	Pela natureza do texto	Secundário
Texto	Pela estrutura	Macroestrutura

Sujeito	Predicado	Objeto
Texto	Pela estrutura	Microestrutura
Texto	Pela estrutura	Superestrutura

Fonte: Elaborado pela autora.

Ao observar a Figura 111 e o Quadro 26, constata-se que as relações semânticas explicitadas são decorrentes da nomeação das subfacetas utilizadas por Lima (2004) (Veja as subfacetas em destaque na Figura 112). Observa-se também que existe uma relação semântica entre *profissional da informação* e *bibliotecário,*; inerente da indentação que denota um tipo de hierarquia. Contudo pela estrutura facetada não foi possível especificar qual é de fato a relação semântica.

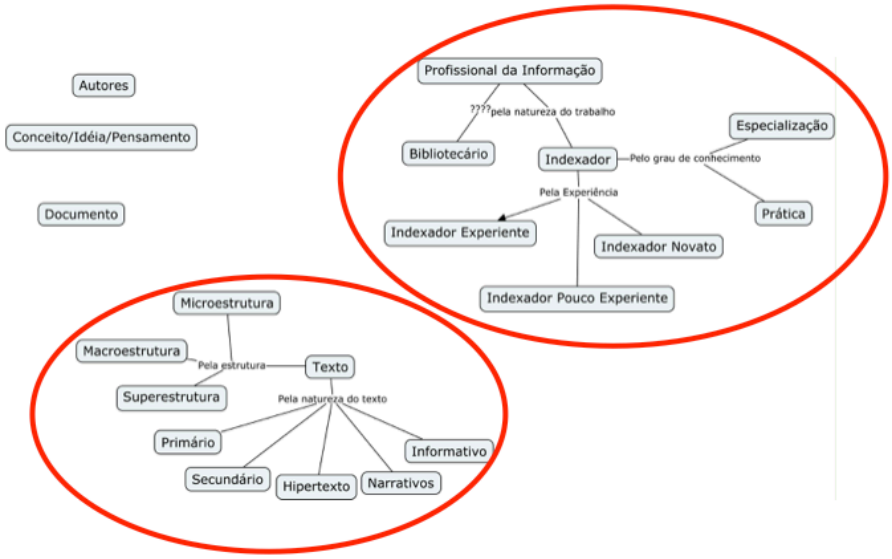
Figura 112 – Estrutura facetada utilizada na amostra



Fonte: Disponível em <http://www.gercinalima.com/mhtx/pages/prototipo-btdeci/teses/naves-mml/estrutura-facetada.php>. Acesso em 20 mai. 2018.

Na organização da estrutura facetada no mapa conceitual, verifica-se a presença de dois *clusters*, conforme destacado na Figura 113. O primeiro grupo é composto por conceitos relacionados a *texto* e o outro a *indexador*.

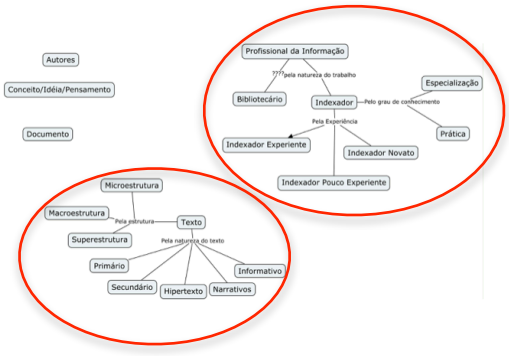
Figura 113 – *Clusters* resultantes do mapa conceitual da estrutura classificatória



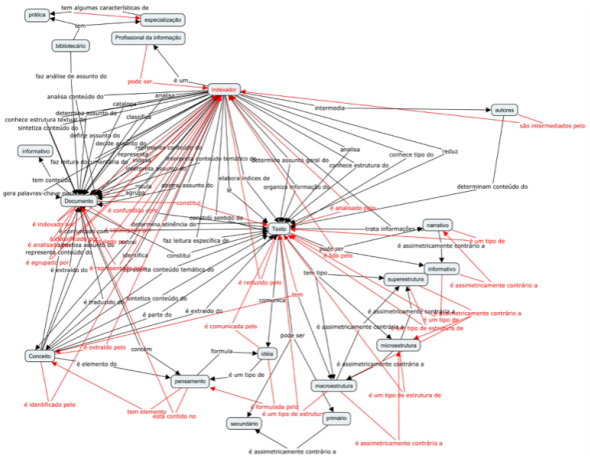
Fonte: Elaborada pela autora.

Nos *clusters* os objetos pertencentes a um grupo são relacionados entre si, porém, eles não se relacionam com os conceitos que estão fora de seu grupo. Portanto, como pode ser visto na Figura 114, no mapa conceitual gerado a partir das relações semânticas encontradas no Semantizar, há uma coesão entre todos os conceitos de tal maneira que esses *clusters* não são possíveis de serem obtidos, ou seja, os conceitos relacionam-se todos entre si.

Figura 114 – Mapa conceitual gerado sem/com as relações semânticas extraídas pelo Semantizar



Mapa conceitual criado **sem** as relações semânticas extraídas pelo Semantizar



Mapa conceitual criado **com** as relações semânticas extraídas pelo Semantizar

Fonte: Elaborada pela autora.

O relacionamento entre todos os conceitos só foi possível com o apoio do Semantizar, que permitiu a criação de uma representação que cobre todos – ou quase todos, devido aos conceitos *indexador novato*, *indexador pouco experiente* e *indexador experiente* – os conceitos de uma estrutura estrutura classificatória semanticamente relacionados de tal forma que o usuário consiga vislumbrar todas as relações possíveis entre os conceitos. Nesse sentido, há um ganho de informação para o usuário. Desse modo, o Semantizar realizou seu objetivo de enriquecer semanticamente uma estrutura classificatória.

7.1 Contribuições para estudos sobre extração de relações semânticas

Esta seção faz uma compilação das contribuições do estudo de caso consideradas importantes para as pesquisas sobre extração de relações semânticas. Elas são enumeradas logo a seguir.

1) A presença de dois conceitos em uma frase é um indício de existência de relação semântica entre esses dois conceitos. Isso pode ser constatado pelos resultados obtidos pela amostra ao apontar que, em 53% das vezes em que ocorreu a presença de dois conceitos da estrutura classificatória em uma frase do documento acadêmico, existia de fato uma relação semântica entre esses conceitos. Nesse caso, alguns dos indícios falsos resultantes foram causados por problemas na amostra ou no Semantizar.

2) Um par de conceitos pode ter mais de uma relação semântica diferente. Constatou-se nos resultados do estudo de caso que alguns pares de conceitos, como o par *indexador* e *documento*, tiveram diferentes relações semânticas explicitadas no mesmo domínio. Desse modo, com várias relações semânticas para o mesmo par de conceitos, nota-se um ganho de informação, pois a representação resultante é semanticamente enriquecida com diferentes aspectos de ligação entre os conceitos.

3) Um par de conceitos pode ter a mesma relação semântica, mesmo em contextos diferentes. Algumas relações semânticas se repetiram para o

mesmo par de conceitos. Essa repetição foi entendida como uma corroboração da relação semântica específica desse par de conceitos. Outra constatação importante diz respeito ao contexto dessas relações semânticas inerentes a esse caso; nas atribuições dos tipos, subtipos e propriedades das relações semânticas não houve alterações quando elas estavam em outros contextos.

4) O contexto, e o conhecimento sobre ele, é fundamental para a determinação do tipo e/ou subtipo da relação semântica para o mesmo par de conceitos com relações diferentes. Isso foi constatado com os pares de conceitos *ideia* e *pensamento* e *macroestrutura* e *texto*. Para o primeiro par, em um contexto ele foi classificado como uma relação associativa (ideia *é formulada pelo* pensamento) e em outro ele foi classificado como uma relação de equivalência (ideia *é um tipo de* pensamento). O outro par de conceitos, *texto* e *macroestrutura*, teve subtipos de relações semânticas diferentes em seus contextos. Em uma situação (texto *tem atributo* macroestrutura), a relação entre eles era do tipo associativa e subtipo atributo convidado e em outro contexto (texto *tem tipo de* macroestrutura) considerou-se que ela era associativa, porém de subtipo conceito-propriedade.

5) A determinação das relações semânticas, na maneira como foi tratada neste tese, depende da interpretação humana. O fator humano foi essencial para que as relações semânticas entre os conceitos fossem estabelecidas coerentemente. Do mesmo modo, a interpretação humana foi fundamental para determinar relações semânticas mesmo quando elas não eram sugeridas pelo Semantizar, como nas relações semânticas entre os conceitos irmãos na hierarquia *microestrutura*, *macroestrutura* e *superestrutura*, assim como entre *primário* e *secundário* e entre *narrativos* e *informativo*. Além disso, o embasamento teórico sobre relações semânticas permitiu identificar, por meio dos parênteses, uma relação de sinonímia entre os conceitos *pensamento* e *ideia*.

6) Os verbos são a principal classe gramatical para definir uma relação semântica. *Contudo, outras classes gramaticais, como os substantivos e as preposições, podem ser utilizadas na formação das relações semânticas.* Essa constatação refutou a ideia que considerava que as relações semânticas

eram definidas apenas pelos verbos. No início da pesquisa, acreditava-se que, ao encontrar dois conceitos em uma frase, o verbo existente nessa frase denotaria a relação semântica entre eles. Contudo, como observou-se nas análises, os verbos são sim a principal classe gramatical para determinar relações semânticas, mas eles não são o único modo de identificar relações semânticas. Além disso, os parênteses, como referido logo acima, a organização de sequências e a análise dos conceitos que pertencem à mesma hierarquia podem indicar a existência de relações semânticas.

7) Nem todas as relações semânticas podem ser explicitadas. Observou-se que, algumas vezes, a relação semântica revelada poderia não ser fidedigna ao que o autor do documento acadêmico pretendia transmitir, por isso, nesses casos, a relação semântica entre os conceitos existia, mas ela não podia ser determinada para não limitar a ideia do autor. Em alguns desses casos, a presença dos advérbios “se” e “não” e alguns adjetivos impediram a determinação das relações semânticas, pois eles qualificavam os substantivos que denotavam os conceitos da amostra, o que dificultou o estabelecimento da relação semântica.

Com relação aos tipos das relações semânticas no contexto em que elas estavam, observou-se que a maioria delas eram associativas. Nesse sentido, prevaleceram as relações causais de subtipo agente-objeto. Logicamente, esse subtipo caracterizou os pares de relações que mais ocorreram, que são: *indexador* e *documento* e *indexador* e *texto*.

Constatou-se sobre as propriedades das relações semânticas que a propriedade de simetria é independente do tipo da relação semântica. Nesse sentido, percebeu-se que, na determinação da simetria, o tipo básico da relação semântica (hierárquica, equivalência e associativa) não foi o fator resolutivo, pois tanto para as relações simétricas quanto para as assimétricas ocorreram todos os tipos de relações semânticas, não sendo possível estabelecer um padrão para essa propriedade.

Ainda sobre a propriedade de simetria, percebeu-se que nenhum subtipo de relação semântica ou par de conceitos foi classificado ora simétrico, ora assimétrico. Isso indica que a simetria não dependeu do contexto em que os pares de conceitos se

encontravam no documento acadêmico e que houve uma consistência na determinação dessa propriedade nas relações semânticas.

Além do mais, constatou-se que a simetria foi fator determinante para o estabelecimento de relações inversas. Conforme mencionou-se, as relações inversas não são possíveis para relações semânticas indiretas; contudo, em todos os casos em que ocorreram relações semânticas diretas classificadas como simétricas, naturalmente podia-se definir as relações inversas devido à própria essência da simetria.

Enfim, observou-se que a propriedade de reflexividade depende do contexto. Diferentemente da simetria, a reflexividade dos tipos de relações semânticas e dos pares de conceitos dependeu do contexto em que o par de conceitos se encontrava. Um exemplo é o par de conceitos *ideia* e *pensamento*, que em um momento foi classificado como irreflexivo e em outro contexto como reflexivo.

7.2 Contribuições do Semantizar

A realização do estudo de caso, dentro do escopo apresentado, foi possível devido ao suporte computacional do Semantizar, pois o enriquecimento semântico de estruturas classificatórias para a representação do conhecimento realizado manualmente pode demandar tempo e esforço por parte do profissional que o realiza. Portanto, o Semantizar facilitou a extração e a explicitação de relações semânticas essenciais para a representação do conhecimento, semiautomatizando essa tarefa. Desse modo, o Semantizar contribuiu para a representação do conhecimento a partir de uma estrutura classificatória, mostrando-se objetivo ao detectar dois conceitos em uma frase de um texto extenso. Entende-se que essa identificação dos pares de conceitos em cada frase é uma das etapas mais laboriosa no contexto que o Semantizar foi criado para atuar.

Logo, por meio do Semantizar, foi possível explicitar 101 relações semânticas, incluindo as inversas, em 53 diferentes pares de conceitos, também considerando as relações inversas, a partir de uma estrutura classificatória contendo 22 conceitos. Com isso, foi possível melhorar a semântica da amostra.

Outra contribuição importante do Semantizar foi que, de certa forma, o Semantizar atuou como um agente de validação da estrutura classificatória, pois ele indicou 199 indícios de relações semânticas em 131 frases do recorte do documento acadêmico ao qual a estrutura classificatória representou. Acredita-se que essa quantidade de indícios é um indicador importante para sugerir que a estrutura relevantemente representa o documento acadêmico. Outro fator importante nesse aspecto diz respeito às frases do documento acadêmico, pois algumas delas abarcavam mais de um indício de relação semântica, o que reforçou a confiança na estrutura classificatória.

Além disso, foi possível, por meio do Semantizar, indicar os pares de conceitos mais importantes para o documento acadêmico. Verificou-se durante a análise dos dados que os pares de conceitos que mais ocorreram denotaram ser os mais relevantes no domínio representado. No caso da amostra, destacou-se os pares: *indexador* e *documento* e *indexador* e *texto*. Tomando como referência que esses pares foram extraídos da tese intitulada *Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores*, de Naves (2000), naturalmente pode-se dizer, mesmo sem ler todo o conteúdo do documento acadêmico, que tais pares são os mais importantes do domínio. Consequentemente, o Semantizar também indicou os conceitos mais importantes para o documento acadêmico. Do mesmo modo como para os pares de conceitos, tomando como referência os conceitos que mais ocorreram nas relações semânticas, pode-se afirmar que os conceitos mais importantes da tese de Naves (2000) foram *indexador*, *texto* e *documento*, sendo que *texto* e *documento* foram classificados como quase sinônimos, o que também indica coerência na determinação dessa relação semântica.

Constatou-se também durante as análises que, muitas vezes, quando as relações inversas não eram possíveis porque as relações eram indiretas, existiam os conceitos *assunto*, *conteúdo* e *informação*. Logo, apurou-se que esses conceitos deveriam compor a estrutura classificatória pelo fato de que eles repetidamente ocorreram e que eles são representativos para o domínio. De maneira análoga, verificou-se que os conceitos com mais indícios falsos no contexto em que eles estavam eram aqueles que, no tipo do texto do documento acadêmico, são utilizados corriqueiramente, como *conceito*, *ideia* e *autores*. Nesse sentido, sugere-se que uma revisão da estrutura classificatória para indicar se eles permanecem para representar

o documento acadêmico. Desse modo, ao apontar essas sugestões, pode-se afirmar que o Semantizar pode atuar no refinamento da estrutura classificatória.

Por fim, notou-se, que a extração dos conceitos pode ser realizada a partir de listas de termos, pois a hierarquia inerente da estrutura classificatória não foi determinante no Semantizar para a indicação da existência de uma relação semântica.

Este capítulo evidenciou os resultados do Semantizar e as suas contribuições para as pesquisas sobre extração de relações semânticas. O próximo capítulo apresenta as considerações finais da tese apontando suas contribuições e trabalhos futuros.

8 CONSIDERAÇÕES FINAIS

Os documentos acadêmicos, como as dissertações e teses, são fontes de conhecimento explícito. Contudo, algumas vezes, tal conhecimento pode ser de difícil compreensão devido à complexidade que envolve algumas pesquisas. Desse modo, a representação do conhecimento pode facilitar o entendimento desses documentos. Na representação do conhecimento, as relações semânticas entre os conceitos contribuem com o enriquecimento semântico do domínio a partir de (ou compondo) bases conceituais, como as estruturas classificatórias, uma vez que as relações semânticas permitem a compreensão da natureza que envolve a ligação entre os conceitos em determinado contexto. Desse modo, esta tese apresentou um Modelo de Extração de Relações Semânticas implementado em um sistema *web* chamado Semantizar.

O Semantizar, com toda a pesquisa necessária para a sua elaboração, considerou a resolução do seguinte problema: *Dado dois conceitos A e B, retirados de uma estrutura classificatória, uma relação semântica pode ser descoberta a partir da análise de A e B nas frases de um texto?* Pelos resultados obtidos nesta tese, pode-se afirmar que sim, dois conceitos de uma estrutura classificatória em uma frase é um forte indício de que existe uma relação semântica entre eles.

Do mesmo modo, o objetivo dessa tese, qual seja, *propor um modelo de extração de relações semânticas para a representação do conhecimento de documentos acadêmicos no contexto do idioma português brasileiro*, foi alcançado com êxito. Através dos resultados do estudo de caso foi possível perceber, com a compilação das relações semânticas em um mapa conceitual, que o modelo implementado pelo Semantizar refletiu o conhecimento embutido no documento acadêmico dentro do escopo dos conceitos da amostra.

Constata-se também que os objetivos específicos da tese foram atingidos. Conforme proposto, (1) o Semantizar facilitou a representação do conhecimento contido em documentos acadêmicos em meio digital. Um mapa conceitual com todas as relações entre os conceitos revelou relações semânticas que *a priori* a estrutura classificatória não sustentava. Desse modo, ele mostrou que é possível apoiar a representação do conhecimento de textos em linguagem natural sugerindo relações

semânticas entre os conceitos em seus contextos. Além disso, (2) o Semantizar permitiu a explicitação das relações semânticas existentes em estruturas classificatórias a partir da extração de suas fontes de informações. Ainda, (3) constatou-se que o Semantizar contribuiu com estudos sobre a extração de relações semânticas em português brasileiro, pois, com base na literatura pesquisada, este trabalho é pioneiro na extração de relações a partir de uma estrutura classificatória, provendo um objeto de estudos futuros para a extração automática de relações semânticas em português brasileiro. Por fim, (4) esta tese colaborou com os estudos acerca de relações semânticas no idioma português brasileiro no cenário da Biblioteconomia e Ciência da Informação. Nesse quesito, propôs-se uma taxonomia que compila todos os tipos de relações semânticas resultantes da bibliografia selecionada. Nessa compilação, arranjos foram feitos para contemplar as visões dos autores pesquisados em relação aos seus pares. Considera-se que essa taxonomia será útil para a Biblioteconomia e Ciência da Informação, sobretudo no Brasil, pois ela oferece as relações semânticas em português. Acredita-se ainda que ela poderá servir de ponto de partida para outras classificações de relações semânticas no contexto nacional de tal forma que lapidações possam ser realizadas ao longo do tempo. Além disso, considera-se a possibilidade de classificar as relações semânticas de acordo com cada Sistema de Organização do Conhecimento.

Sobre a metodologia de pesquisa adotada, concluiu-se que ela foi satisfatória. Nela determinou-se a realização de uma pesquisa aplicada que empregou técnicas de pesquisa exploratória, como pesquisa bibliográfica e estudo de caso.

A pesquisa bibliográfica foi necessária, sobretudo, em dois momentos: na fundamentação teórico-metodológica e na revisão de literatura. Sendo que no primeiro momento, houve duas contribuições, quais sejam: (1) a taxonomia das relações semânticas, mencionada anteriormente; (2) a corroboração de algumas constatações, como dos autores Arnold e Rahm (2014) sobre o uso dos parênteses nas construções frasais como recomendação de conceitos sinônimos; de Stock (2010) sobre as relações ternárias na formação de relações binárias; e de Broughton (2008), que afirmou que os conceitos irmãos na mesma relação hierárquica podem indicar a existência de relações semânticas associativas.

No que tange à revisão de literatura, buscou-se conhecer pesquisas que tratavam da extração de relações semânticas, a fim de compreender o estado da arte

dessa temática no intervalo de tempo de 2013 a 2017, seguindo uma metodologia específica que apoiou a seleção dos trabalhos encontrados. Nessa seleção, apurou-se as técnicas utilizadas pelos pesquisadores para a extração de relações semânticas e os contextos aos quais elas foram aplicadas. Desse modo, constatou-se a escassez de pesquisas sobre extração de relações semânticas em português brasileiro realizadas recentemente. Notou-se ainda que a aplicação de técnicas de extração de relações semânticas não é facilmente adaptada para outros idiomas devido às suas características próprias. Logo, cada idioma exige um estudo apurado para a extração de relações semânticas.

O estudo de caso mostrou-se adequado para a proposta. A metodologia empregada permitiu análises que apontaram importantes contribuições para pesquisas acerca da extração e explicitação de relações semânticas quais foram: a presença de dois conceitos em uma frase é um indício de existência de relação semântica entre esses dois conceitos; um par de conceitos pode ter mais de uma relação semântica diferente; um par de conceitos pode ter a mesma relação semântica, mesmo em contextos diferentes; o contexto, e o conhecimento sobre ele, é fundamental para a determinação do tipo e/ou subtipo da relação semântica para o mesmo par de conceitos com relações diferentes; a determinação das relações semânticas, na maneira como foi tratada neste tese, depende da interpretação humana; os verbos são a principal classe gramatical para definir uma relação semântica, mas não a única e; nem todas as relações semânticas podem ser explicitadas.

A respeito do Semantizar, ele foi desenvolvido em PHP e MySQL de tal forma que possa ser utilizado no ambiente Web. Ele foi implementado para processar e armazenar uma estrutura classificatória e o documento acadêmico ao qual essa estrutura foi criada para representar. No processamento da estrutura classificatória, o Semantizar colhe todos os conceitos e os armazena no banco de dados, caso eles ainda não estejam nele, e indica relações semânticas para que o usuário possa validá-las e armazená-las, ratificando seus respectivos tipos e subtipos, propriedades e relações inversas. Destaca-se, portanto, que o Semantizar permitiu manipular os conceitos de uma estrutura classificatória e encontrá-los nas frases do documento acadêmico ao qual ela representava.

Por fim, uma das principais contribuições desta tese foi a determinação de um *novo subtipo de relação semântica associativa*, chamado *agente subordinado*. Esse subtipo foi definido para suprir as relações entre *autores* e *indexador* nos contextos em que eles se apresentavam ao serem analisados no estudo de caso.

8.1 Trabalhos futuros

Pesquisas futuras para aperfeiçoar o Semantizar deverão incluir uma verificação se a relação semântica entre dois conceitos extraídos pelo sistema já existe na base de dados. Desse modo, acredita-se que à medida que a base de dados for ampliada, ela poderá dar suporte para as pesquisas sobre automatização da extração das relações semânticas nesse contexto. Estas pesquisas deverão incluir técnicas de mineração de dados para descoberta de padrões; reconhecimento de entidades nomeadas (com o uso do DBPedia), em que cada conceito será tratado como uma URI; POS *tags* (com o emprego do *software* PALAVRAS) e análise de árvores sintáticas. Ainda, estudos sobre pontuações *f-score* deverão ser feitos para avaliar formalmente a eficiência do Semantizar na extração de relações semânticas.

Outra alternativa de automatização da extração de relações semânticas é a criação de *synsets,* tanto de conceitos quanto de relações semânticas. Os *synsets* de conceitos seguem o modelo já existente, como o WordNet[60], que naturalmente suportam os conceitos sinônimos, tais como *documento* e *texto*, por exemplo. Já os *synsets* de relações semânticas apoiarão casos de sinônimos de relações semânticas explicitadas anteriormente, como, por exemplo: *faz indexação e indexa.* Porém, essa implementação carecerá de pesquisas e profissionais que entendam do domínio para validar os sinônimos. Como visto, existem algumas classificações de sinônimos que devem ser respeitadas.

Notou-se que a validação dos indícios de relações semânticas e a determinação e caracterização delas, por vezes são subjetivas, ou seja, essas atividades dependem da concepção do usuário que está realizando-as. Desse modo, sugere-se efetuar estudos com usuários de modo que possa ser apurado o fator de

[60] Disponível em <https://wordnet.princeton.edu>. Acesso em 20 set. 2017.

interferência deles na determinação das relações semânticas. Ainda neste aspecto, deverá ser determinado o perfil do usuário adequado para utilizar o Semantizar para explicitar relações semânticas.

Conforme apontado na seção 6.5, deverão ser realizadas melhorias no algoritmo do Semantizar no que tange à manipulação do vetor de conceitos e à função que implementa buscas de *strings.* Do mesmo modo, pesquisas sobre a utilização de XML na marcação da estrutura classificatória no arquivo de texto deverão ser feitas para indicar de antemão as relações hierárquicas inerentes da estrutura.

REFERÊNCIAS

ABNT NBR 14724. ABNT NBR 14724: Informação e documentação – Trabalhos acadêmicos – Apresentação. Rio de Janeiro: Associação Brasileira de Normas Técnicas, 2011.

ARNOLD, P.; RAHM, E. Extracting Semantic Concept Relations from Wikipedia. *4th International Conference on Web Intelligence*, Mining and Semantics, 2014.

AUGENSTEIN, I.; MAYNARD, D.; CIRAVEGNA, F. Distantly supervised web relation extraction for knowledge base population. *Semantic Web*, v. 7, n. 4, p. 335-349, 2016.

AZEREDO, J. C. DE. *Fundamentos de Gramática do Português*. [S.l.]: Jorge Zahar Editor Ltda, 2000.

BACH, N.; BADASKAR, S. A review of relation extraction. *Literature review for Language and Statistics II*, 2007. Disponível em: <http://orb.essex.ac.uk/CE/CE807/Readings/A-survey-on-Relation-Extraction.pdf>. Acesso em: 1 abr. 2017.

BAILEY, K. D. *Typologies and Taxonomies*: An Introduction to Classification Techniques. [S.l.]: SAGE, 1994.

BATISTA, D. S. *et al.* Extracçao de Relaçoes Semânticas de Textos em Português Explorando a DBpédia e a Wikipédia. *linguamatica*, v. 5, n. 1, p. 41-57, 2013.

BEAN, A. C.; GREEN, R. *Relationships in the Organization of Knowledge*. Boston/Dordrecht/London: Kluwer Academic Publishers, 2001. v. 2. Disponível em: <https://books.google.com.br/books?hl=en&lr=&id=PLzAzBZYcegC&oi=fnd&pg=PR7&dq=Relationships+in+the+organization+of+knowledge&ots=wRpS5vuWsQ&sig=nj0W-N1G0-Erby4dLtStydAxnjs>. Acesso em: 1 nov. 2016.

BEGHTOL, C. Relationships in Classificatory Structure and Meaning. *Relationships in the Organization of Knowledge*. 1ª ed. Boston/Dordrecht/London: Kluwer Academic Publishers, v. 2, 2001, p. 99-113.

BLANCO, E.; MOLDOVAN, D. Composition of semantic relations: Theoretical framework and case study. *ACM Transactions on Speech and Language Processing (TSLP)*, v. 10, n. 4, p. 17, 2013. Disponível em: <http://dl.acm.org/citation.cfm?id=2513146>. Acesso em: 14 nov. 2016.

BRÄSCHER, M. Semantic Relations in Knowledge Organization Systems. *Knowledge Organization*, v. 41, n. 2, p. 175-180, abr. 2014. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=95782279&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid>. Acesso em: 17 set. 2015.

BROUGHTON, V. A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification. *Axiomathes*, v. 18, n. 2, p. 193-210, 2008.

BROUGHTON, V. *et al.* Knowledge Organization. *European Curriculum Reflections on Library and Information Science Education*. [S.l.]: Royal School of Library and Information Science, Copenhagen, 2005. . Disponível em: <http://arizona.openrepository.com/arizona/handle/10150/105851>. Acesso em: 16 abr. 2016.

BUENO, S. *Dicionário Global Escolar Silveira Bueno da Língua Portuguesa*. [S.l.]: Global Editora e Distribuidora Ltda, 2017.

CABRÉ CASTELLVÍ, M. T. C. *Terminology*: Theory, methods and applications. [S.l.]: John Benjamins Publishing, 1999.

CAFÉ, L. M. A.; BRASCHER, M. *Organização do Conhecimento*: teorias semânticas como base para estudo e representação de conceitos. 2011. Disponível em: <http://repositorio.unb.br/handle/10482/12894>. Acesso em: 31 out. 2016.

CAFÉ, L.; MENDES, F. Estudo sobre a estrutura definitória para desenvolvimento de ontologias. *Informação & Sociedade*, v. 19, n. 2, 2009. Disponível em: <http://search.proquest.com/openview/fedd043473ed063d8f598cb6498d804a/1?pq-origsite=gscholar>. Acesso em: 22 maio 2016.

CAMPOS, M. L. DE A. Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. *Ciência da Informação*, v. 33, n. 1, p. 22-32, abr. 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-19652004000100003&lng=en&nrm=iso&tlng=pt>. Acesso em: 29 abr. 2016.

CARVALHO, D. S.; FREITAS, A.; DA SILVA, J. C. P. Graphia: Extracting Contextual Relation Graphs from Text. In: CIMIANO, P. *et al.* (Org.). . *Semantic Web: Eswc 2013 Satellite Events*. Berlin: Springer-Verlag Berlin, v. 7955, 2013, p. 236-241.

CHAFFIN, R.; HERRMANN, D. J. The similarity and diversity of semantic relations. *Memory & Cognition*, v. 12, n. 2, p. 134-141, 1984. Disponível em: <http://link.springer.com/article/10.3758/BF03198427>. Acesso em: 20 out. 2016.

CHUA, C. E. H.; STOREY, V. C.; CHIANG, R. H. Knowledge representation: a conceptual modeling approach. *Journal of Database Management (JDM)*, v. 23, n. 1, p. 1-30, 2012. Disponível em: <http://www.igi-global.com/article/content/62030>. Acesso em: 1 out. 2015.

COOKE, N. J. Eliciting semantic relations for empirically derived networks. *International Journal of Man-Machine Studies*, v. 37, n. 6, p. 721-750, 1992.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, v. 13, n. 1, p. 21-27, 1967.

CRUSE, D. A. Hyponymy and Its Varieties. *The Semantics of Relationships*: An Interdisciplinary Perspective. [S.l.]: Springer-Science+Business Media, B.V, 2002, p. 3-22.

CUNNINGHAM, H. Information extraction, automatic. *Encyclopedia of language and linguistics,* p. 665-677, 2005.

DA SILVA, E. L.; MENEZES, E. M. *Metodologia da pesquisa e elaboração de dissertação*. 4. ed. rev. atual. ed. Florianópolis: UFSC, 2005.

DAHLBERG, I. Teoria do conceito. *Ciência da Informação*, v. 7, n. 2, 30 dez. 1978a. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/view/1680>. Acesso em: 21 jul. 2015.

DAHLBERG, I. A referent-oriented, analytical concept theory of Interconcept. *International Classification*, v. 5, n. 3, p. 122-151, 1978b.

DAHLBERG, I. Fundamentos teórico-conceituais da classificação. *Revista de Biblioteconomia de Brasília*, Brasília, v. 6, n. 1, p. 9-21, jan./jun. 1978c.

DAVIS, R.; SHROBE, H.; SZOLOVITS, P. What is a knowledge representation? *AI magazine*, v. 14, n. 1, p. 17, 1993. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1029>. Acesso em: 22 jul. 2015.

DE ABREU, S. C.; BONAMIGO, T. L.; VIEIRA, R. A review on Relation Extraction with an eye on Portuguese. *Journal of the Brazilian Computer Society*, v. 19, n. 4, p. 553-571, 2013.

DODEBEI, V. L. D. Tesauro: linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002.

DOLK, D. R.; KONSYNSKI, B. R. Knowledge representation for model management systems. *Software Engineering, IEEE Transactions on*, n. 6, p. 619–628, 1984. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5010291>. Acesso em: 22 jan. 2016.

FIGUEIREDO, N. Da importância dos artigos de revisão da literatura. *Revista Brasileira de biblioteconomia e documentação, São Paulo*, v. 23, n. 1/4, p. 131-135, 1990.

FINK, A. *Conducting Research Literature Reviews: From the Internet to Paper*. [S.l.]: SAGE Publications, 2013.

GARSHOL, L. M. Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. *Journal of information science* 30, n. 4, 2004, p. 378-391.

GERHARDT, T. E.; SOUZA, A. C. DE. Unidade 1 – Aspectos teóricos e conceituais. *Métodos de pesquisa*. Porto Alegre: Editora da UFRGS, 2009, p. 11-30. Disponível em: <http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf>. Acesso em: 17 jun. 2016.

GERSTL, P.; PRIBBENOW, S. A conceptual theory of part-whole relations and its applications. *Data & Knowledge Engineering*, v. 20, n. 3, p. 305-322, 1996. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169023X96000146>. Acesso em: 13 dez. 2016.

GIL, A. C. Como elaborar projetos de pesquisa. *São Paulo*, v. 5, p. 61, 2002. Disponível em: <http://www.academia.edu/download/31110279/9482_lista_de_revisao_1%C2%BA_bimestre_com_respostas_direito.pdf>. Acesso em: 17 jun. 2016.

GIL, A. C. *Métodos e Técnicas de Pesquisa Social*. 2 ed. São Paulo: Editora Atlas S.A, 1989. Disponível em: <https://www.google.com.br/imgres?imgurl=http://d19qz1cqidnnhq.cloudfront.net/imagens/capas/_8c333c5afc5be79beaa5fa74a563676a518e4555.jpg&imgrefurl=http://www.estantevirtual.com.br/b/antonio-carlos-gil/metodos-e-tecnicas-de-pesquisa-social/3452947768&h=314&w=223&tbnid=ejpUIY4NQO3h-M:&tbnh=160&tbnw=113&docid=dljACHZFDNvNzM&itg=1&client=firefox-b-ab&usg=__tYLSr67X2j3ekBfcAqPdRpvq7DY=>. Acesso em: 12 jul. 2016.

GIUNCHIGLIA, F.; DUTTA, B.; MALTESE, V. From Knowledge Organization to Knowledge Representation. *Knowledge Organization*, v. 41, n. 1, p. 44-56, 2014.

GREEN, R. Relationships in knowledge organization. *Knowledge organization*, v. 35, n. 2-3, p. 150-159, 2008.

GREEN, R. Relationships in the Organization of Knowledge: An Overview. *Relationships in the Organization of Knowledge*. 1ª ed. Boston/Dordrecht/London: Kluwer Academic Publishers, v. 2, 2001, p. 3-18.

GREEN, R.; BEAN, C. A.; MYAENG, S. H. *The Semantics of Relationships: An Interdisciplinary Perspective*. [S.l.]: Springer Science & Business Media, 2013.

GRIGOROVA, D.; NIKOLOV, N. Knowledge representation in systems with natural language interface. 2007, [S.l.]: ACM, 2007, p. 68. Disponível em: <http://dl.acm.org/citation.cfm?id=1330670>. Acesso em: 26 jan. 2016.

GRISHMAN, R.; SUNDHEIM, B. *Message Understanding Conference-6*: A Brief History. 1996, [S.l: s.n.], 1996, p. 466-471. Disponível em: <http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/C96-1079.pdf>. Acesso em: 1 abr. 2017.

HJØRLAND, B. *Information seeking and subject representation:* An Activity-theoretical Approach to Information Science. [S.l.]: ABC-CLIO/Greenwood, 1997.

HJØRLAND, B. Semantics and knowledge organization. *Annual review of information science and technology*, v. 41, n. 1, p. 367-405, 2007. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/aris.2007.1440410115/full>. Acesso em: 27 out. 2016.

HUHNS, M. N.; STEPHENS, L. M. Plausible Inferencing Using Extended Composition. In: *IJCAI.* 1989. p. 1420-1425.

ISO 25964-1. *ISO 25964-1: Information and documentation* — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval. [S.l.]: International Standart, 2011.

KHOO, C. S.; NA, J.-C. Semantic relations in information science. *Annual review of information science and technology*, v. 40, p. 157, 2006. Disponível em: <https://dr.ntu.edu.sg/bitstream/handle/10220/18367/1440400112_ftp_acc.pdf?sequence=3>. Acesso em: 25 out. 2016.

KLIEGR, T. Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery. *Journal of Web Semantics*, v. 31, p. 59-69, mar. 2015.

KONSTANTINOVA, N. Review of Relation Extraction Methods: What Is New Out There? 2014, [S.l.]: Springer, 2014, p. 15-28. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-319-12580-0_2>. Acesso em: 29 mar. 2017.

KUCZORA, P. W.; COSBY, S. J. Implementation of meronymic (part-whole) inheritance for semantic networks. *Knowledge-Based Systems*, v. 2, n. 4, p. 219-227, 1989. Disponível em: <http://www.sciencedirect.com/science/article/pii/095070518990066X>. Acesso em: 1 dez. 2016.

KUMAR, E. *Natural Language Processing*. [S.l.]: I. K. International Pvt Ltd, 2011.

LA BARRE, K. Facet analysis. *Annual review of information science and technology*, v. 44, n. 1, p. 243-284, 2010.

LI, H. *et al.* A relation extraction method of Chinese named entities based on location and semantic features. *Applied Intelligence*, v. 38, n. 1, p. 1-15, jan. 2013.

LIMA, G. Â. B. DE O. *MAPA HIPERTEXTUAL (MHTX): Um modelo para organização hipertextual de documento*. 2004. 207 f. Tese de Doutorado – Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

LUFT, C. P. *Moderna gramática brasileira*. Edição Reimpressão. Globo Livros, 2002, 265 p.

MACULAN, B. C. M. DOS S. *Estudo e aplicação de metodologia para reengenharia de tesauro: remodelagem do THESAGRO*. 2015. 343 f. Tese de Doutorado – Universidade Federal de Minas Gerais, Belo Horizonte, 2015. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/BUBD-9ZKMUV>. Acesso em: 23 maio 2016.

MANI, I.; MAYBURY, M. T. *Advances in Automatic Text Summarization*. [S.l.]: MIT Press, 1999.

MARINOV, M. Using frames for knowledge representation in a CORBA-based distributed environment. *Knowledge-Based Systems*, v. 21, n. 5, p. 391-397, 2008.

MAZZOCCHI, F. Relations in KOS: is it possible to couple a common nature with different roles? *Journal of Documentation*, v. 73, n. 2, p. 368-383, 13 mar. 2017.

MOREHOUSE, S. *LibGuides: Research Skills Tutorial*: Current and Retrospective Information Sources. Disponível em: <http://subjectguides.esc.edu/researchskillstutorial/currency>. Acesso em: 26 fev. 2017.

MUKUL, G.; DEEPA; GUPTA. *Research Methodology*. [S.l.]: PHI Learning Pvt. Ltd., [S.d.].

MURPHY, M. L. *Semantic Relations and the Lexicon*: Antonymy, Synonymy and other Paradigms. [S.l.]: Cambridge University Press, 2003.

NANARD, J.; NANARD, M. *Should anchors be typed too?*: an experiment with MacWeb. 1993, [S.l.]: ACM, 1993, p. 51-62. Disponível em: <http://dl.acm.org/citation.cfm?id=168767>. Acesso em: 15 jan. 2017.

NAVES, M. M. L. *Fatores interferentes no processo de análise de assunto*: estudo de caso de indexadores. 2000. 283 f. Tese de Doutorado – Universidade Federal de Minas Gerais, Belo Horizonte, 2000. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/BUOS-A4RGHM>. Acesso em: 16 jun. 2016.

NEBHI, K. *A rule-based relation extraction system using DBpedia and syntactic parsing*. 2013, [S.l.]: CEUR-WS.org, 2013, p. 74-79. Disponível em: <http://dl.acm.org/citation.cfm?id=2874487>. Acesso em: 4 abr. 2017.

OKOLI, C.; SCHABRAM, K. A guide to conducting a systematic literature review of information systems research. *Sprouts Work. Pap. Inf. Syst*, v. 10, n. 26, 2010. Disponível em: <http://www.academia.edu/download/3250666/OkoliSchabram2010SproutsLitReviewGuide.pdf>. Acesso em: 24 fev. 2017.

PAIVA, L. *et al.* Discovering Semantic Relations from Unstructured Data for Ontology Enrichment Asssociation rules based approach. In: ROCHA, A. *et al.* (Org.). *Proceedings of the 2014 9th Iberian Conference on Information Systems and Technologies (cisti 2014)*. New York: Ieee, 2014.

PETERS, I. P.; WELLER, K. Paradigmatic and syntagmatic relations in knowledge organization systems. v. 59(1), p. 100-107, 2008. Disponível em: <http://www.phil.hhu.de/fileadmin/Redaktion/Institute/Informationswissenschaft/peters/1204547334paradigmat.pdf>. Acesso em: 31 out. 2016.

PINHEIRO, Lena Vania Ribeiro; FERREZ, Helena Dodd. *Tesauro Brasileiro de Ciência da Informação*. Rio de Janeiro; Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), 2014. Disponível em: <http://www.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1/copy_of_TESAUROCOMPLETOFINALCOMCAPA24102014.pdf>. Acesso em 31 de out. 2017.

PRADONOV, C. C.; FREITAS, E. C. DE. *Metodologia do Trabalho Científico:* Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico - 2ª Edição. 2 ed. Novo Hamburgo, Rio Grande do Sul: Editora Feevale, 2013.

PRESSMAN, R.; MAXIN, B. Engenharia de Software. 8 ed. McGraw Hill Brasil, 2016, 968 p.

PRIETO-DÍAZ, R. *A faceted approach to building ontologies*. 2003, [S.l.]: IEEE, 2003. p. 458-465. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1251451>. Acesso em: 25 maio 2016.

RODRIGUES, M.; TEIXEIRA, A. *Advanced Applications of Natural Language Processing for Performing Information Extraction*. [S.l.]: Springer, 2015.

SALES, L. F. Modelo triádico de relações para aplicação em ontologias. *SEMINÁRIO BRASILEIRO DE ONTOLOGIAS*, v. 1, 2010. Disponível em: <http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=18732>. Acesso em: 22 maio 2016.

SALES, L. F.; SAYÃO, L. F.; DA MOTTA, D. F. *Modelagem de Relações Conceituais para a Área Nuclear*. 2012, [S.l.]: Citeseer, 2012. p. 182-187. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.8828&rep=rep1&type=pdf#page=182>. Acesso em: 22 maio 2016.

SALES, L. F.; SAYÃO, L. F.; OTHERS. *Modelo triádico de relações*: um protótipo de modelagem conceitual para a área nuclear. 2012. Disponível em: <http://carpedien.ien.gov.br/bitstream/ien/650/1/artigo_enancib2012_sales_say%C3%A3o.pdf>. Acesso em: 19 set. 2016.

SARAWAGI, S. Information extraction. *Foundations and Trends® in Databases*, v. 1, n. 3, p. 261-377, 2008.

SELIGMANN-SILVA, M. *Leituras de Walter Benjamin*. [S.l.]: Annablume, 2007.

SILVEIRA, D. T.; CÓRDOVA, F. P. Unidade 2 - A pesquisa científica. *Métodos de pesquisa*. Porto Alegre: Editora da UFRGS, 2009, p. 31-42. Disponível em: <http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf>. Acesso em: 17 jun. 2016.

SOERGEL, D. *Knowledge Organization Systems. Overview*. [S.l: s.n.], [S.d.]. Disponível em: <http://www.dsoergel.com/SoergelKOSOverview.pdf>.

SOERGEL, D. The rise of ontologies or the reinvention of classification. *Journal of the Association for Information Science and Technology*, v. 50, n. 12, p. 1119, 1999.

STOCK, W. G. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, v. 61, n. 10, p. 1951-1969, 2010. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.21382/full>. Acesso em: 27 out. 2016.

STOREY, V. C. Understanding semantic relationships. *The VLDB Journal*, v. 2, n. 4, p. 455-488, 1993. Disponível em: <http://link.springer.com/article/10.1007/BF01263048>. Acesso em: 24 out. 2016.

SVENONIUS, E. *The intellectual foundation of information organization*. Cambridge, MA: MIT Press, 2000. Disponível em: <https://books.google.com.br/books?

hl=en&lr=lang_en|lang_pt|
lang_es&id=r0iBW7fygu8C&oi=fnd&pg=PR3&dq=The+intellectual+foundations+of+in
formation+organization&ots=QP7vFeermU&sig=55IiwTSCQ5T8f1MtEW8e_lhf050>.
Acesso em: 15 jul. 2017.

SZOSTAK, R. Classifying Relationships. *Knowledge Organization*, v. 39, n. 3, 2012.

TRISTÃO, A. M. D.; FACHIN, G. R. B.; ALARCON, O. E. Sistema de classificação facetada e tesauros: instrumentos para organização do conhecimento. *Ciência da Informação*, Brasília, v. 33, n. 2, p. 161-171, ago. 2004.

VAN HARMELEN, F.; LIFSCHITZ, V.; PORTER, B. *Handbook of knowledge representation*. [S.l.]: Elsevier, 2008. v. 1. Disponível em: <https://
books.google.com.br/books?hl=pt-
BR&lr=&id=xwBDylHhJhYC&oi=fnd&pg=PP1&dq=knowledge+representation&ots=W
SIKnmABQZ&sig=SM-OlxPCk0KsKC71ppGU1PNaqKU>. Acesso em: 22 jan. 2016.

VICKERY, B. Knowledge representation: a brief review. *Journal of Documentation*, v. 42, n. 32, 1986.

VICKERY, B. C. On Knwoledge Organisation. *Facets of Knowledge Organization:* Proceedings of the ISKO UK Second Biennial Conference, 4th-5th July, 2011, London. London: Emerald Group Publishing, 2008.

VICKERY, B. C. Thesaurus-A new word in documentation. *Journal of documentation*, v. 16, n. 4, p. 181-189, 1960.

WELLER, K.; STOCK, W. G. Transitive Meronymy. v. 59, n. 3, p. 165-170, 2008. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?
doi=10.1.1.218.9555&rep=rep1&type=pdf>. Acesso em: 12 dez. 2016.

WINSTON, M. E.; CHAFFIN, R.; HERRMANN, D. A taxonomy of part-whole relations. *Cognitive science*, v. 11, n. 4, p. 417-444, 1987. Disponível em: <http://
onlinelibrary.wiley.com/doi/10.1207/s15516709cog1104_2/abstract>. Acesso em: 12 dez. 2016.

WOHLIN, C. *et al. Experimentation in Software Engineering*. [S.l.]: Kluwer Academic Publishers Boston/Dordrecht/London, 2000.

XU, M. *et al.* Discovering Missing Semantic Relations between Entities in Wikipedia. In: ALANI, H. *et al.* (Org.). *Semantic Web - Iswc 2013, Part I*. Berlin: Springer-Verlag Berlin, 2013. v. 8218, p. 673-686.

ZELENKO, D.; AONE, C.; RICHARDELLA, A. Kernel methods for relation extraction. *Journal of machine learning research*, v. 3, n. Feb, p. 1083-1106, 2003.

ZENG, M. L. Knowledge organization systems (KOS). *Knowledge organization*, v. 35, n. 2-3, p. 160-182, 2008. Disponível em: <http://cat.inist.fr/?
aModele=afficheN&cpsidt=20613691>. Acesso em: 15 set. 2015.

ZHU, X. *et al.* Detecting concept relations in clinical text: Insights from a state-of-the-art model. *Journal of Biomedical Informatics*, v. 46, n. 2, p. 275-285, abr. 2013.

APÊNDICE A – DICIONÁRIO DE DADOS DO SEMANTIZAR

tbPublicacao

Coluna	Tipo	Nulo	Padrão	Links para
IDPublicacao *(Primária)*	int(11)	Não		
TituloPublicacao	varchar(200)	Não		
AutorPublicacao	varchar(100)	Não		
OrientadorPublicacao	varchar(100)	Não		
CoorientadorPublicacao	varchar(100)	Não		
AnoPublicacao	year(4)	Não		
DiretorioPublicacao	mediumblob	Não		
DiretorioEstruturaPublicacao	mediumblob	Não		
TipoPublicacao	varchar(30)	Não		
SiglaUniversidade	varchar(30)	Sim	*NULL*	tbUniversidade -> SiglaUniversidade

Índices

Nome da chave	Único	Coluna	Nulo
PRIMARY	Sim	IDPublicacao	Não
UniquetituloPublicacao	Sim	TituloPublicacao	Não
fk_SiglaUniversidade	Não	SiglaUniversidade	Sim

tbPublicacaoRelacao

Coluna	Tipo	Nulo	Padrão	Links para
IDRelacao *(Primária)*	int(11)	Não		tbRelacao -> IDRelacao
IDPublicacao *(Primária)*	int(11)	Não		tbPublicacao -> IDPublicacao

Índices

Nome da chave	Único	Coluna	Nulo
PRIMARY	Sim	IDRelacao	Não
		IDPublicacao	Não
fk_Pub	Não	IDPublicacao	Não

tbRelacao

Coluna	Tipo	Nulo	Padrão	Links para
IDRelacao *(Primária)*	int(11)	Não		
IDPalavraSujeito	int(11)	Não		tbSubstantivo -> IDSubstantivo
Predicado	varchar(100)	Não		
IDPalavraObjeto	int(11)	Não		tbSubstantivo -> IDSubstantivo
IDTipoRelacao	int(11)	Não		tbTipoRelacao -> IDTipoRelacao
RelacaoInversa	varchar(100)	Sim	*NULL*	
Simetrica	tinyint(1)	Sim	*NULL*	
Reflexiva	tinyint(1)	Sim	*NULL*	
Transitiva	tinyint(1)	Sim	*NULL*	

Índices

Nome da chave	Único	Coluna	Nulo
PRIMARY	Sim	IDRelacao	Não
fk_IDPalavraSujeito	Não	IDPalavraSujeito	Não
fk_IDPalavraObjeto	Não	IDPalavraObjeto	Não
fk_TipoRelacao	Não	IDTipoRelacao	Não

tbSubstantivo

Coluna	Tipo	Nulo	Links para
IDSubstantivo *(Primária)*	int(11)	Não	
NomeSubstantivo	varchar(100)	Não	

Índices

Nome da chave	Único	Coluna	Nulo
PRIMARY	Sim	IDSubstantivo	Não
NomePalavra	Sim	NomeSubstantivo	Não

tbTipoRelacao

Coluna	Tipo	Nulo	Padrão	Links para
IDTipoRelacao *(Primária)*	int(11)	Não		
NomeTipoRelacao	varchar(100)	Não		
subTipoRelacaoDe	int(11)	Sim	*NULL*	tbTipoRelacao -> IDTipoRelacao

Índices

Nome da chave	Único	Coluna	Nulo
PRIMARY	Sim	IDTipoRelacao	Não
fk_subTipoRelacao	Não	subTipoRelacaoDe	Sim

tbUniversidade

Coluna	Tipo	Nulo	Padrão	Links para
SiglaUniversidade *(Primária)*	varchar(30)	Não		
NomeUniversidade	varchar(100)	Sim	*NULL*	

Índices

Nome da chave	Único	Coluna	Nulo
PRIMARY	Sim	SiglaUniversidade	Não

APÊNDICE B - INDÍCIOS DE RELAÇÕES SEMÂNTICAS FALSOS

Indexador e Documento

1. “Citam um experimento de Zunde & Dexter em que o mesmo documento foi dado para oito indexadores em um blind test , para determinar a inconsistência entre eles” (NAVES, 2000, p. 32)

2. “Os resultados mostram que nenhum termo de indexação foi selecionado por todos os oito indexadores e que nem dois indexadores selecionaram a mesma lista de termos para o documento em questão” (NAVES, 2000, p. 32)

3. “A indexação baseada nos títulos dos documentos, ao mesmo tempo em que pode ser considerada rápida e fácil, por outro lado pode tornar-se uma armadilha para indexadores e usuários, na medida em que as palavras de títulos muitas vezes fogem ao verdadeiro conteúdo do documento” (NAVES, 2000, p. 38)

4. “Questões relevantes são levantadas por ALBRECHTSEN (1993), como: De que estamos falando quando falamos sobre "assuntos" de livros e outros documentos? Existem diferentes concepções de assuntos e daí, de análise de assuntos; então, são essas concepções interconcectadas com métodos aplicados pela indexação? Em seu artigo, o autor apresenta um modelo alternativo para discutir Análise de assunto e indexação, com a intenção de tentar colocar a indexação num contexto social mais amplo, além dos métodos de evolução mecânica, e apontar novos desafios para indexadores” (NAVES, 2000, p. 61)

5. “É o caso, por exemplo, da Norma ISO 5963, de 1985, em que são estabelecidos métodos para examinar documentos, determinar seus assuntos e selecionar termos indexadores” (NAVES, 2000, p. 69)

Conceito e Texto

1. “Cita problemas que algumas instituições têm enfrentado com indexadores que, segundo o autor, são "especialistas" demais, e correm o risco de interpretar excessivamente um texto, talvez extrapolando aquilo que o próprio autor afirma, ou, mesmo, demonstrando preconceitos, ao não indexar afirmações que relutam em aceitar” (NAVES, 2000, p. 23). Nesse caso, o Semantizar detectou preconceitos como conceito.

2. “Van Dijk afirma que texto é um conceito abstrato que se manifesta ou realiza em discursos concretos (BÉRNARDEZ, 1982)” (NAVES, 2000, p. 41)

3. “O texto é considerado a unidade da análise do discurso e o que caracteriza a relação entre discurso e texto, para ORLANDI (1987), é o seguinte: eles se eqüivalem, mas em níveis conceituais diferentes, ou seja, o discurso é tomado como conceito teórico e metodológico, e o texto, em contrapartida, como o conceito analítico correspondente, sendo o texto considerado uma unidade complexa de significação, tendo em vista as condições de sua realização” (NAVES, 2000, p. 41)

4. “O conceito de coesão textual é apresentado por HALLIDAY & HASAN (citados por KOCK, 1991) como um conceito semântico, que se refere às relações de sentido existentes no interior do texto e que o definem como um texto” (NAVES, 2000, p. 45)

5. “Sua solução, às vezes, depende da compreensão de um conceito ou de uma teoria; problemas ontológicos - são provenientes das crenças ideológicas do leitor, são construídos pelo leitor e geralmente resultam em comentários críticos sobre o texto” (NAVES, 2000, p. 53)

6. “Se, para fazer uma análise conceitual, devem-se extrair conceitos, pergunta-se: o que é um conceito? Como identificá-lo? Qual a sua importância no processo de Análise de assunto? Para responder a essas questões, são feitas, a seguir, algumas considerações sobre conceito, assunto

e contexto, todos eles termos constantes no processo em estudo" (NAVES, 2000, p. 54). Nesse caso, o Semantizar detectou contexto como texto.

7. "Os dois conceitos - assunto e contexto - estão sempre juntos nas teorias tradicionais de indexação" (NAVES, 2000, p. 62). Idêntico ao caso anterior.

8. "É o caso do conhecimento humano materializado em textos científicos e do "mundo" dos fatos lógicos (conceitos universais) arranjados numa estrutura lógica" (NAVES, 2000, p. 77)

9. "Para analisar os processos cognitivos da geração e compreensão dos sistemas de linguagem (textos ou discursos), BEGHTOL (1986) considera que os textos lingüísticos tomam emprestado da Psicologia Cognitiva os conceitos de dois tipos de processamento mental da informação: top-down e bottom-up, que ocorrem durante a leitura" (NAVES, 2000, p. 80)

Indexador e Texto

1. "Diante das novas tecnologias da informação e da evolução de softwares desenvolvidos para tornar a indexação mais eficaz e rápida, certas questões estão sempre presentes nas discussões dos profissionais que lidam com a informação: Qual será o futuro do indexador? Diante das bibliotecas virtuais e de textos digitalizados em sofisticados sistemas e bases de dados, haverá lugar para o indexador humano? Se este permanecer, qual será o seu papel? Observa-se, pelas discussões sobre o tema, que não se conseguiu, pelo menos até hoje, transferir para a máquina o tipo de tarefa em que estão presentes elementos como a abstração, a percepção, a interpretação e outros processos inerentes ao funcionamento da mente humana" (NAVES, 2000, p. 17)

2. "Outros autores, como CHU & O`BRIEN (1993), já consideram quatro estágios no processo de indexação, a saber: (1) análise de assunto do texto; (2) expressão do conteúdo do assunto nas palavras dos indexadores (linguagem natural); (3) tradução para um vocabulário de indexação; (4) expressão do assunto em termos do índice" (NAVES, 2000, p. 27)

Conceito e Documento

1. "Também é usado o conceito de "inconsistência inter-indexadores", o que significa que diferentes indexadores determinam diferentes termos para o mesmo documento" (NAVES, 2000, p. 31)

2. "Essas considerações sobre conceitos permitem que se introduza a noção de "análise conceitual", que nada mais é do que a identificação dos tópicos estudados num documento e, para LANCASTER (1993), o processo de reconhecer o "de que trata" um documento se reveste de interesse para uma determinada comunidade, pelo fato de contribuir para nossa compreensão de tópicos" (NAVES, 2000, p. 58)

3. "Ele oferece um novo conceito de assunto como a totalidade dos potenciais epistemológicos de documentos" (NAVES, 2000, p. 60)

4. "Após essas considerações sobre conceito, assunto e contexto, pode-se afirmar que esses são aspectos interdependentes que ocorrem durante todo o processo de Análise de assunto, até o momento de se afirmar sobre o que trata um documento, próxima fase do processo a ser tratado" (NAVES, 2000, p. 64)

5. "Em seu trabalho sobre atinência, MARON (1977) dá um enfoque probabilístico ao estudo do termo e, para ele, o conceito chave da teoria da indexação parece ser o conceito about, porque a decisão chave no procedimento de indexação refere-se à questão sobre de qual assunto trata o documento" (NAVES, 2000, p. 64)

6. “Nesta fase da determinação da atinência para representar os conceitos extraídos do texto, inicia-se um processo lingüístico e o problema de descrever documentos para recuperação é, principalmente, o problema de como a linguagem é usada” (NAVES, 2000, p. 67)

7. “De acordo com essa norma, os métodos são especialmente planejados para sistemas de indexação nos quais os assuntos dos documentos são expressos na forma de Sumário e onde conceitos são registrados nos termos de uma linguagem de indexação controlada” (NAVES, 2000, p. 69)

Idéia e Texto

1. “Para os textos informativos, a classificação mais conhecida, segundo GIASSON (1993) é a de Meyer, e compreende: descrição (dá informações de um sujeito e especifica alguns de seus atributos e características), enumeração (o texto apresenta uma lista de elementos ligados entre si por um ponto comum), comparação (texto que serve para comparar objetos, pessoas ou acontecimentos entre sí, tendo em conta suas semelhanças e diferenças), causa-efeito (é possível identificar no texto uma relação causal entre as idéias) e problema-solução (pergunta/resposta - o problema antecede à solução)” (NAVES, 2000, p. 44)

2. “Outro aspecto que merece ser ressaltado é que o autor do texto, ao escrevê-lo, tem em mente um determinado leitor alvo para o qual direciona suas idéias; suas intenções não são dirigidas para o leitor/indexador e não lhe interessa se esse vai ter capacidade para interpretar as informações que aquele texto está veiculando” (NAVES, 2000, p. 53)

3. “Essa é a idéia de LURIA (1994) e, segundo ele, precisamente, esse aspecto é o mais complexo e importante problema da compreensão de textos” (NAVES, 2000, p. 80)

4. “Um outro tipo de processamento, ou modelo, é o interativo, que se fundamenta na idéia de que a produção da leitura se dá a partir da interação, ou troca, contínua entre leitor e texto” (NAVES, 2000, p. 82).

5. “Compreender implica, então, extrair de algum objeto alguma forma de significado e relacioná-lo a outras experiências ou idéias, parafrasear, abstrair conteúdos, responder a questões sobre o material escrito, ou criticar um texto (WEINER & CROMER, citados por DELL'ISOLA, 1999)” (NAVES, 2000, p. 84).

Documento e Texto

1. “julgamento:deve-se tomar decisões sobre documentos e a abordagem a ser dada a eles e, se o documento for aceito, deve-se decidir o grau de exaustividade a ser aplicado na sua indexação; leitura: esta significa incorporação mental da estrutura e do conteúdo de um texto, não para si mesmo, mas como valor para o usuário do texto” (NAVES, 2000, p. 18).

2. “Concepção orientada para o conteúdo - envolve uma interpretação adicional do conteúdo, que vai além dos limites da estrutura léxica e gramatical, com o estabelecimento de assuntos que não estão explicitamente colocados no texto, mas que são facilmente identificados pelo indexador, envolvendo, portanto, uma abstração mais indireta do documento” (NAVES, 2000, p. 36).

3. “Existem textos orais mas, no caso em estudo, a atenção fica restrita a textos escritos, maioria esmagadora na produção científica de documentos publicados” (NAVES, 2000, p. 41).

4. “Por isso se afirma que a análise documentária tem como objetivo recuperar e disseminar informação, e não textos, discursos ou documentos - suportes materiais da informação” (NAVES, 2000, p. 45).

5. “Uma segunda leitura é necessária, devendo-se concentrar nos cabeçalhos do documento, seções (objetivos, metodologia, resultados e conclusões, no caso do texto científico), pois essas contêm as estruturas do texto” (NAVES, 2000, p. 54).

6. “Questões relevantes são levantadas por ALBRECHTSEN (1993), como: De que estamos falando quando falamos sobre "assuntos" de livros e outros documentos? Existem diferentes concepções de assuntos e daí, de análise de assuntos; então, são essas concepções interconcectadas com métodos aplicados pela indexação? Em seu artigo, o autor apresenta um modelo alternativo para discutir Análise de assunto e indexação, com a intenção de tentar colocar a indexação num contexto social mais amplo, além dos métodos de evolução mecânica, e apontar novos desafios para indexadores” (NAVES, 2000, p. 61). Nesse caso, o Semantizar detectou contexto entendendo ser texto.

7. “Na determinação do assunto, é preciso que se verifique o contexto no qual o documento é produzido e para o qual ele existe, em determinado momento” (NAVES, 2000, p. 61). Idem ao caso anterior.

8. “Após essas considerações sobre conceito, assunto e contexto, pode-se afirmar que esses são aspectos interdependentes que ocorrem durante todo o processo de Análise de assunto, até o momento de se afirmar sobre o que trata um documento, próxima fase do processo a ser tratado” (NAVES, 2000, p. 64). Idem ao caso anterior.

9. “Nesta fase da determinação da atinência para representar os conceitos extraídos do texto, inicia-se um processo lingüístico e o problema de descrever documentos para recuperação é, principalmente, o problema de como a linguagem é usada” (NAVES, 2000, p. 67).

10. “Há, entretanto, segundo FARROW (1995), algumas diferenças quanto aos objetivos a serem alcançados: para o classificador, o objetivo é a representação da atinência do documento, traduzida para a linguagem do sistema de classificação; para o elaborador de resumos, o que importa é representar em prosa contínua os argumentos principais do documento; para o indexador, há dois tipos de propósitos: (a) para o indexador de livros, a tarefa é ler o texto, distinguindo entre informações relevantes e periféricas, e empregando uma mistura dos processamentos top-down (conceitual) e bottom-up (perceptivo) obtidos na leitura fluente normal, e (b) para o indexador acadêmico, a indexação é menos exaustiva e usa predominantemente a abordagem top-down” (NAVES, 2000, p. 85).

Indexador e Conceito

1. “Uma distinção dos tipos de indexação é feita no terceiro capítulo deste estudo, para diferenciar a indexação de livros da indexação acadêmica, no sentido em que estas são aqui compreendidas, o que, conseqüentemente, altera também o conceito de indexador” (NAVES, 2000, p. 16).

2. “Não há dúvida de que o indexador interponha suas próprias idéias e preconceitos na atuação de intermediário entre autores e usuários” (NAVES, 2000, p. 18). O Semantizar detectou preconceito ao invés de conceito

3. “Cita problemas que algumas instituições têm enfrentado com indexadores que, segundo o autor, são "especialistas" demais, e correm o risco de interpretar excessivamente um texto, talvez extrapolando aquilo que o próprio autor afirma, ou, mesmo, demonstrando preconceitos, ao não indexar afirmações que relutam em aceitar” (NAVES, 2000, p. 23). Idem ao caso anterior.

4. “Alguns aspectos e conceitos abordados neste capítulo voltam a ser comentados quando se fala na atividade desempenhada pelo indexador, que é a indexação, e no principal processo desenvolvido por ele, a Análise de assunto, que é a primeira etapa dessa atividade, como pode ser visto nos próximos capítulos” (NAVES, 2000, p. 25).

5. “Também é usado o conceito de "inconsistência inter-indexadores", o que significa que diferentes indexadores determinam diferentes termos para o mesmo documento” (NAVES, 2000, p. 31).

6. “O experimento consistiu na produção de duas listas de palavras para cada livro analisado: uma de conceitos e uma de termos de indexação selecionados da linguagem Rameau […], baseada na lista de cabeçalhos de assunto da Library of Congress (USA), e foram gravadas fitas durante a execução da tarefa de oito indexadores, para observação da sua atividade mental” (NAVES, 2000, p. 31).

7. “Para propósitos documentários, PINTO MOLINA (1995) considera os seguintes tipos de inferência: lógica - usada para estabelecer causas, motivações e condições que permitem fatos específicos; evolutiva - na qual analistas aplicam suas crenças às situações descritas; integrativa - executada no momento da compreensão e baseada nos conceitos e propriedades da organização hierárquica; construtiva - baseada no conhecimento do indexador” (NAVES, 2000, p. 85).

Conceito e Idéia

1. “Não há dúvida de que o indexador interponha suas próprias idéias e preconceitos na atuação de intermediário entre autores e usuários” (NAVES, 2000, p. 19). Nesse caso, o Semantizar detectou preconceito entendendo ser conceito.

2. “Definir um conceito é uma operação verbal e lógica bem clara, na qual se usa uma série de idéias logicamente subordinadas para chegar a uma conclusão geral” (NAVES, 2000, p. 56).

3. “Já com relação ao conceito objetivo-idealístico, o autor pensa que esse não considera o assunto como subjetivo e, sim, tende a enfatizar aspectos da análise teórica e os torna absolutos e, para ele, as idéias existem independentemente da consciência humana” NAVES, 2000, p. 59).

4. “Este último conceito encontra-se expresso em diversos vocábulos tais como mensagem do autor, visão de conjunto, elementos importantes, ponto de vista principal, idéia central do texto, havendo a respeito diversidade de concepções” (NAVES, 2000, p. 60).

5. “Essa informação importante pode variar segundo os tipos de texto: nos textos narrativos, a idéia principal tem a ver com os acontecimentos e a sua interpretação, enquanto nos textos informativos o que é importante pode ser um conceito, uma generalização, uma regra” (NAVES, 2000, p. 61).

6. “É necessário incluir considerações de como o pensamento é produzido, isto é, de que maneira estímulos externos ou internos resultam em atividade mental ou registro mental, e como os pensamentos são retidos na memória, ou dela desaparecem, e podem ser diferenciados de outros pensamentos (percepção, conceitos, idéias)” (NAVES, 2000, p. 86).

Texto e Superestrutura

1. “Ela afirma que pesquisas têm demonstrado que leitores com conhecimento prévio específico sobre superestruturas textuais executam a tarefa de ler de forma mais fácil que leitores que não vêm essa superestrutura e, por isso, são obrigados a um maior apoio na leitura palavra por palavra, dificultando a integração das informações no texto como um todo” (NAVES, 2000, p. 43).
2. “O leitor que domina as superestruturas textuais capta com mais facilidade as idéias centrais do texto, pois tem como parâmetro a identificação dos constituintes básicos” (NAVES, 2000, p. 43).

Autores e Texto

1. “autores FLECK & BAWDEN (1995), o primeiro item é considerado de particular importância, ressaltando o problema da auto-imagem e da auto-estima como de fundamental importância nesse contexto, afirmando que atitudes negativas e destrutivas na profissão como um todo, e

entre seus membros, são mais prejudiciais do que todos os estereótipos já feitos por intrusos” (NAVES, 2000, p. 16).

2. “Por exemplo, cita o texto *Prefaces* de George Bernard Shaw, valioso por sua informação autobiográfica, que pode ser analisado por sua crítica social, ou pelos diferentes comentários dos autores” (NAVES, 2000, p. 39).

3. “E o último componente definido pelos autores acima é a intertextualidade, que concerne aos fatores que fazem a utilização de um texto dependente do conhecimento de outro(s) texto(s)” (NAVES, 2000, p. 47).

4. “O modelo desses autores defende a visão de que o processamento de texto se baseia em redução proposicional, isto é, a redução do texto através de suas proposições para se chegar à macroestrutura (estrutura semântica global)” (NAVES, 2000, p. 80).

Conceito e Pensamento

1. “Vale ressaltar a importância de dois conceitos presentes nos processos cognitivos: o pensamento e a memória” (NAVES, 2000, p. 85).

Autores e Conceito

1. “Imagem profissional é um conceito difícil de ser definido com precisão e os autores acima mencionados citam, sobre o assunto, a abordagem dada por KOREN, que considera, como um ponto de partida razoável para a definição da imagem de um profissional, os elementos título profissional, papel profissional (definido), capacidade, comportamento e status social” (NAVES, 2000, p. 15).

2. “Verifica-se, no exame da literatura especializada em Biblioteconomia e Ciência da Informação, que o termo "Análise de assunto" é o mais comumente utilizado, mas que grande parte dos autores que tratam do tema o consideram ou como a etapa de tradução dos conceitos extraídos dos documentos para um vocabulário controlado, ou até mesmo do processo de indexação como um todo” (NAVES, 2000, p. 38).

3. “Esses autores descrevem problemas ocorridos na área de padronização e escolha do melhor termo pnara representar o conceito” (NAVES, 2000, p. 39).

4. “Alguns autores tratam do processo de formação de conceitos” (NAVES, 2000, p. 56).

Autores e Documento

1. “Essas dificuldades se dão por motivos como diferenças de interpretação, existência de várias maneiras de se pensar sobre os tópicos, sobre os argumentos dos autores ao se basearem em pontos de vista diversos, e tornam essa noção de indexação de documentos interpretada como "sobre o que eles tratam" difícil de sustentar” (NAVES, 2000, p. 65).

2. “Segundo os autores, é evidente que nessa área não há uma maneira simples de caracterizar documentos e, por esse motivo, propuseram uma "abordagem multi-modal” (NAVES, 2000, p. 65).

3. “De acordo com os autores, o conteúdo de assunto de um documento se refere, algumas vezes, ao que eles chamam de atinência intrínseca” (NAVES, 2000, p. 66).

4. “Já as questões de "como" o documento pode ser usado, por que ele foi adquirido, e outras variadas considerações externas, referem-se ao que os autores chamam de atinência extrínseca” (NAVES, 2000, p. 66).

Indexador e Prática

1. “Cita a abordagem prática de Preschel e, para este, "conceito" significa "matéria indexável", e "análise conceitual" é a percepção, pelo indexador, de matéria indexável” (NAVES, 2000, p. 58).

2. “SMIT (1978) afirma que, salvo algumas exceções, nessa época, os trabalhos de síntese ainda não haviam sido efetuados, e aponta algumas razões para essa contradição: (1) as pesquisas lingüísticas desenvolveram-se primordialmente, até então, no campo da sintaxe, cujo interesse é importante mas não suficiente numa perspectiva documentária e (2) a prática da análise documentária baseou-se muito tempo em operações implícitas, circunscritas à "cultura" individual do indexador (indexação manual)” (NAVES, 2000, p. 71).

Ideia e Pensamento

1. “Ele cita autores como Cutter (que define assunto como tema ou tópico, podendo ou não estar no título do documento), Kaiser (que toma "assuntos" como "coisas em geral", reais ou imaginárias, e as condições para designá-las, que são chamadas "concretos e processo"), Ranganathan (que fala sobre o pensamento contido no documento), Coates (que identifica assunto como abstração da idéia global corporificada no conteúdo de uma unidade literária dada), e Vickery (que se refere ao tema a partir do qual livros, parte de livros, artigos ou parte de artigos são escritos)” (NAVES, 2000, p. 59).

Texto e Informativo

1. “É um tipo de leitura bem racional e rápido, em que o leitor técnico não tem chances de aproveitar a leitura, já que seu propósito é o de extrair o conteúdo informativo do texto, tendo em vista a sua posterior recuperação por um leitor interessado” (NAVES, 2000, p. 53).

Autores e Indexador

1. “Outros autores, como CHU & O`BRIEN (1993), já consideram quatro estágios no processo de indexação, a saber: (1) análise de assunto do texto; (2) expressão do conteúdo do assunto nas palavras dos indexadores (linguagem natural); (3) tradução para um vocabulário de indexação; (4) expressão do assunto em termos do índice” (NAVES, 2000, p. 27).

2. “De Mey acha que a tarefa de recuperação da informação é explicitar estruturas cognitivas de autores, analistas de sistemas e indexadores, de acordo com a necessidade real dos usuários” (NAVES, 2000, p. 78).

Pensamento e Documento

1. “Ele se vale de um pensamento de Foskett, quando diz que a operação chave da indexação, que é a decisão sobre o que o documento é, ainda é menos discutida e a menos redutível a regras” (NAVES, 2000, p. 27).

2. “Para interpretar o conteúdo de um documento, é certo que o indexador passa por um processo cognitivo, e o autor acima citado acredita que a Ciência da Informação deve incluir estudos sobre seres humanos e, em particular, sobre os processos de pensamento e comportamento em todas as ações relativas à comunicação, se são origem dela, se são seus manipuladores ou são recipientes de comunicações” (NAVES, 2000, p. 86).

Indexador e Idéia

1. “Essa falsa idéia ocorre por absoluto desconhecimento da complexidade do processo, que exige esforços por parte do profissional, no caso o indexador, no sentido de seguir uma metodologia adequada para obter resultados satisfatórios em seu trabalho, e sabe-se que uma das formas de

avaliar a eficácia desse pode ser através dos resultados obtidos no momento da recuperação da informação" (NAVES, 2000, p. 36).

2. "Outro aspecto que merece ser ressaltado é que o autor do texto, ao escrevê-lo, tem em mente um determinado leitor alvo para o qual direciona suas idéias; suas intenções não são dirigidas para o leitor/indexador e não lhe interessa se esse vai ter capacidade para interpretar as informações que aquele texto está veiculando" (NAVES, 2000, p. 53).

3. "Concorda-se com essas idéias e defende-se a necessidade da presença do indexador humano nesse tipo de atividade" (NAVES, 2000, p. 62).

Profissional da Informação e Indexador

1. "Este capítulo discorre, inicialmente, sobre características do profissional da informação e, sobre o indexador, mostra a importância do papel por ele exercido, e de certos aspectos relativos à interferência de seu modo de ser no seu próprio trabalho, como subjetividade, conhecimento prévio, formação e experiência anterior" (NAVES, 2000, p. 14).

Texto e Primário

1. "A estrutura semântica é caracterizada em dois níveis: microestrutura (nível primário, no qual se designa, por um lado, a estrutura das proposições individuais e, por outro lado, as relações entre as seqüências das frases no texto) e a macroestrutura (que representa as relações entre os grupos de frases, ou a organização geral do texto" (NAVES, 2000, p. 83).

Prática e Conceito

1. "Cita a abordagem prática de Preschel e, para este, "conceito" significa "matéria indexável", e "análise conceitual" é a percepção, pelo indexador, de matéria indexável" (NAVES, 2000, p. 58).

2. "Alguns autores como Blair, Hjorland, Weinberg e Soërgel apontam novas direções para a indexação, restabelecento o conceito de "assunto" numa parte principal da prática e da teoria da indexação" (NAVES, 2000, p. 61).

Prática e Texto

1. "No estudo desse processo, verificam-se duas abordagens para a identificação do conteúdo de um texto: a abordagem teórica e a abordagem prática" (NAVES, 2000, p. 24).

2. "Ao se considerar a leitura como uma prática de dar sentido perceptivo e intelectual a um texto, está sendo inserido, nesse contexto, o sujeito e toda a sua capacidade subjetiva de interpretar" (NAVES, 2000, p. 48).

Autores e Prática

1. "Alguns autores como Blair, Hjorland, Weinberg e Soërgel apontam novas direções para a indexação, restabelecento o conceito de "assunto" numa parte principal da prática e da teoria da indexação" (NAVES, 2000, p. 61).

Autores e Idéia

1. "Ele cita autores como Cutter (que define assunto como tema ou tópico, podendo ou não estar no título do documento), Kaiser (que toma "assuntos" como "coisas em geral", reais ou imaginárias, e as condições para designá-las, que são chamadas "concretos e processo"), Ranganathan (que fala sobre o pensamento contido no documento), Coates (que identifica assunto como abstração

da idéia global corporificada no conteúdo de uma unidade literária dada), e Vickery (que se refere ao tema a partir do qual livros, parte de livros, artigos ou parte de artigos são escritos)" (NAVES, 2000, p. 59).

Bibliotecário e Indexador

1. "Esse é o ponto de vista de LANCASTER (1993), que recomenda ao indexador não apenas permanecer "nos bastidores", mas, também, procurar desempenhar outras atividades, inclusive a de bibliotecário de referência" (NAVES, 2000, p. 25).

Bibliotecário e Prática

1. "Quanto à prática, somente após longa experiência, é que, provavelmente, o bibliotecário desenvolverá métodos de trabalho eficientes" (NAVES, 2000, p. 22).

Bibliotecário e Idéia

1. "A grande maioria dos profissionais que exercem atividades relacionadas ao tratamento e à organização da informação são graduados em Biblioteconomia, mas o próprio nome do profissional bibliotecário já vem sendo considerado um fator que limita a idéia da vasta abrangência da área de atuação desse profissional" (NAVES, 2000, p. 14).

Especialização e Texto

1. "A especialização do profissional é destacada também por GINEZ DE LARA (1993) e, para ela, na indexação realizada diretamente a partir do texto original, a perda de referenciais concretos é minimizada quando o profissional envolvido com a tarefa é um especialista de área (em condições, portanto, de reconhecer a arquitetura conceitual do campo em questão e identificar e interpretar ocorrências típicas da linguagem técnico-científica) ou quando a experiência faz do documentalista um especialista" (NAVES, 2000, p. 23).

Conceito e Secundário

1. "Já para OLIVEIRA (1997), conceito é visto como uma imagem subjetiva do mundo objetivo e revela aspectos essenciais, universais do objeto, abstraindo-se dos aspectos secundários" (NAVES, 2000, p. 56).

Idéia e Documento

1. "O modelo acima citado se assemelha à idéia de CESARINO & PINTO (1980), citada no início deste capítulo, de que há dois momentos em que se faz Análise de assunto: o momento em que se analisa o documento e o momento em que se analisam as questões do usuário" (NAVES, 2000, p. 66).

Documento e Secundário

1. "Isso significa que há sérias dificuldades na escolha de um assunto que seja considerado o principal de um documento, deixando outros, que ele inclua, em plano secundário" (NAVES, 2000, p. 60).

Texto e Secundário

1. “Segundo essa autora, podem surgir problemas durante todo um processo de leitura, como os apontados a seguir: problemas contingentes: são ligados, por exemplo, ao significado de uma palavra desconhecida; problemas táticos: são relacionados à organização do texto, por exemplo, a confusão entre os planos principal e secundário do discurso, devido a falha do leitor ou do escritor; problemas modais - devem ser resolvidos pelo próprio leitor” (NAVES, 2000, p. 52).

Primário e Microestrutura

1. “A estrutura semântica é caracterizada em dois níveis: microestrutura (nível primário, no qual se designa, por um lado, a estrutura das proposições individuais e, por outro lado, as relações entre as seqüências das frases no texto) e a macroestrutura (que representa as relações entre os grupos de frases, ou a organização geral do texto” (NAVES, 2000, p. 83).

APÊNDICE C - RELAÇÕES SEMÂNTICAS EXPLICITADAS

Relação Semântica 01

Conceitos	Autores e Texto
Frase	"Dentre os textos informativos, pode-se reconhecer o texto científico e, nesse tipo de texto, o conteúdo é quase inteiramente determinado pelo autor; geralmente, relatórios de pesquisa são altamente informativos e os autores os constroem numa estrutura convencional com introdução, metodologia, resultados e discussão, o chamado modelo clássico" (NAVES, 2000, p. 44, grifo nosso).
Relação Semântica	autores *determinam conteúdo do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível

Relação Semântica 02

Conceitos	Autores e Indexador
Frase	"Não há dúvida de que o indexador interponha suas próprias idéias e preconceitos na atuação de intermediário entre autores e usuários" (NAVES, 2000, p. 19, grifo nosso)
Relação Semântica	indexador *intermedia* autores
Tipo da Relação	Associativa - Agente Subordinado (Nova relação semântica)
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	autores *são intermediados pelo* indexador

Relação Semântica 03

Conceitos	Indexador e Documento
Frase	“Os processos de produção e compreensão são similares, no trabalho de indexadores, classificadores, elaboradores de resumos (abstractors), de condensações e sinopses: o processo comum entre eles é a Análise de assunto e seu propósito comum, a representação do documento.” (NAVES, 2000, p. 85, grifo nosso)
Relação Semântica	indexador *representa* documento
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	documento *é representado pelo* indexador.

Relação Semântica 04

Conceitos	Indexador e Documento
Frase	“Mas pouco é encontrado sobre como indexadores decidem qual é o assunto de um documento, como eles decidem "sobre o que trata", em que classes de assuntos se insere” (NAVES, 2000, p. 28, grifo nosso)
Relação Semântica	indexador *decide assunto do* documento
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 05

Conceitos	Indexador e Documento
Frase	“No ato de pensar, quando faz abstrações, interpreta e define o assunto de um documento, o indexador sofre influência de diversos fatores pertencentes a vários campos, principalmente oriundos da Lingüística, da Ciência Cognitiva e da Lógica” (NAVES, 2000, p. 70, grifo nosso).
Relação Semântica	indexador *abstrai assunto do* documento
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.
Observação	Nesse caso ainda existem outras duas relações semânticas entre indexador e documento. Os próximos quadros tratam dessas relações.

Relação Semântica 06

Conceitos	Indexador e Documento
Frase	“No ato de pensar, quando faz abstrações, interpreta e define o assunto de um documento, o indexador sofre influência de diversos fatores pertencentes a vários campos, principalmente oriundos da Lingüística, da Ciência Cognitiva e da Lógica” (NAVES, 2000, p. 70, grifo nosso).
Relação Semântica	indexador *interpreta assunto do* documento
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 07

Conceitos	Indexador e Documento
Frase	"No ato de pensar, quando faz abstrações, interpreta e define o assunto de um documento, o indexador sofre influência de diversos fatores pertencentes a vários campos, principalmente oriundos da Lingüística, da Ciência Cognitiva e da Lógica" (NAVES, 2000, p. 70, grifo nosso).
Relação Semântica	indexador *define assunto do* documento
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 08

Conceitos	Indexador e Profissional da Informação
Frase	"Apesar de este estudo tratar especificamente da pessoa do indexador, termo adotado para designar o profissional que faz a indexação, algumas considerações devem ser feitas, inicialmente, sobre o profissional da informação" (NAVES, 2000, p. 14, grifo nosso).
Relação Semântica	indexador *é um* profissional da informação
Tipo da Relação	Hierárquica - Hipônimo-Hiperônimo - Inclusão de Classe - Funcionalmente subordinado
Propriedades	Assimétrica (Não pode ser afirmado que todo profissional da informação é indexador)
	Reflexiva (Indexador *é um* indexador)
Relação Inversa	profissional da informação *pode ser* indexador.

Relação Semântica 09

Conceitos	Indexador e Conceito
Frase	"Em síntese, este capítulo mostra o processo em que o indexador faz a leitura de um texto, empreende a extração de conceitos e determina a sua atinência". (NAVES, 2000, p. 69, grifo nosso)
Relação Semântica	Indexador *extrai* conceito.
Tipo da Relação	Associativa - Ação subordinada (Extrair conceitos é parte da atividade do indexador)
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Conceito *é extraído pelo* indexador.
Observação	O Semantizar também encontrou o indício de relação entre conceito e texto nessa frase.
	Nessa frase, também existe uma relação entre indexador e texto, contudo, devido a uma falha do Semantizar, ela não foi detectada automaticamente. Logo, a relação semântica entre indexador e texto nesse contexto é: indexador *lê* texto. Essa relação foi detectada na frase que está no **Quadro de Relações Semânticas 11.** Ainda nessa frase tem a relação: indexador *determina a atinência do* texto, como pode ser visto no **Quadro 22.**
	Essa relação também foi encontrada na frase: "A leitura do texto é, então, a primeira fase do processo de Análise de assunto e, a partir dela, o indexador parte para a próxima fase, que é a da extração, do texto, de conceitos que representem seu conteúdo" (NAVES, 2000, p. 54, grifo nosso). Contudo, essa relação configura em uma relação ternária entre texto-indexador-conceito. Nesse caso a relação seria a seguinte: conceito *é extraído do* texto *pelo* indexador. Ou indexador *extrai* conceito *do* texto. Como mencionado, uma relação ternária pode desmembrar-se em relações binárias. Logo outra relação surge: conceito *é extraído do* texto. Essa relação está detalhada no **Quadro Relações Semânticas 30.**

Relação Semântica 10

Conceitos	Indexador e Conceito
Frase	"É proposta, ainda, uma *check-list* que orienta o indexador através de uma lista de questões a serem formuladas no momento da identificação de conceitos" (NAVES, 2000, p. 69, grifo nosso).
Relação Semântica	Indexador *identifica* conceito.
Tipo da Relação	Associativa - Ação subordinada (Identificar conceitos é parte da atividade do indexador)
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Conceito *é identificado pelo* indexador.

Relação Semântica 11

Conceitos	Indexador e Texto
Frase	"A esse respeito, FARROW (1995) afirma que a indexação back-of-book permite ao leitor localizar informação sobre um tópico dentro do livro; a tarefa do indexador é ler o texto, distinguir entre informação relevante e periférica e empregar os tipos de processamento de informação presentes na leitura" (NAVES, 2000, p. 26, grifo nosso).
Relação Semântica	Indexador *lê o* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Texto *é lido pelo* indexador.
	Essa relação semântica também foi detectada na frase: "Para ter uma competência textual é preciso que, além de conhecer o texto que tem em mãos para análise sob todos os aspectos até aqui abordados, o indexador faça dele uma leitura adequada, e sabe-se que um texto pode gerar muitas leituras, interessando mais, neste estudo, a leitura para fins documentários" (NAVES, 2000, p. 48, grifo nosso).

Relação Semântica 11

Observação	Nessa frase o Semantizar detectou a relação entre indexador e documento, contudo, ao analisa-la, verificou-se que existe a relação indexador *lê* texto. "para o indexador, há dois tipos de propósitos: (a) para o indexador de livros, a tarefa é ler o texto, distinguindo entre informações relevantes e periféricas, e empregando uma mistura dos processamentos top-down (conceitual) e bottom-up (perceptivo) obtidos na leitura fluente normal, e (b) para o indexador acadêmico, a indexação é menos exaustiva e usa predominantemente a abordagem top-down (NAVES, 2000, p. 85, grifo nosso).

Relação Semântica 12

Conceitos	**Indexador e Texto**
Frase	"Nas literaturas inglesa e americana, o termo indexador é aplicado tanto àquela pessoa que elabora índices de textos ou livros quanto àquela que faz a indexação acadêmica" (NAVES, 2000, p. 16, grifo nosso).
Relação Semântica	Indexador *elabora índices de* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 13

Conceitos	Indexador e Texto
Frase	"É mediante a interação de diversos níveis de conhecimento, como o conhecimento lingüístico, o textual, o conhecimento de mundo, que o leitor, no caso o indexador, consegue construir o sentido do texto" (NAVES, 2000, p. 21, grifo nosso).
Relação Semântica	Indexador *constrói sentido do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 14

Conceitos	Indexador e Texto
Frase	"Os resultados mostram que os indexadores demonstraram dificuldade menor em determinar o assunto geral de textos em Ciências Sociais e Ciências do que em textos de Humanidades" (NAVES, 2000, p. 33, grifo nosso).
Relação Semântica	Indexador *determina assunto geral do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 15

Conceitos	Indexador e Texto
Frase	“Para facilitar seu trabalho, é necessário que o indexador conheça estruturas e tipos de textos”. (NAVES, 2000, p. 43, grifo nosso)
Relação Semântica	Indexador *conhece estrutura do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 16

Conceitos	Indexador e Texto
Frase	“Para facilitar seu trabalho, é necessário que o indexador conheça estruturas e tipos de textos”. (NAVES, 2000, p. 43, grifo nosso)
Relação Semântica	Indexador *conhece tipo do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 17

Conceitos	Indexador e Texto
Frase	"O autor acima acredita que as teorias de Van Dijk e Kintsh têm uma atração imediata para indexadores: indexação é a redução do texto e tem-se, aqui, uma explanação teórica desse processo" (NAVES, 2000, p. 85, grifo nosso).
Relação Semântica	Indexador *reduz* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Texto *é reduzido pelo* indexador

Relação Semântica 18

Conceitos	Indexador e Texto
Frase	"Este item mostra a importância do texto como matéria prima para a Análise de assunto, de estruturas e de tipos de textos, bem como a forma específica da leitura que deve ser feita pelo indexador" (NAVES, 2000, p. 41, grifo nosso).
Relação Semântica	Indexador *faz leitura específica de* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 19

Conceitos	Indexador e Texto
Frase	"Como se vê, a nova concepção de texto deixa confuso o indexador, responsável pela organização e tratamento das informações que nele estão contidas" (NAVES, 2000, p. 48, grifo nosso).
Relação Semântica	Indexador *organiza informação do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.
Observação	Nesse caso, outra relação é possível: Indexador *trata informações do* texto. Essa relação será tratada no próximo Quadro.

Relação Semântica 20

Conceitos	Indexador e Texto
Frase	"Como se vê, a nova concepção de texto deixa confuso o indexador, responsável pela organização e tratamento das informações que nele estão contidas" (NAVES, 2000, p. 48, grifo nosso).
Relação Semântica	Indexador *trata informações do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 21

Conceitos	Indexador e Texto
Frase	"Ao analisar um texto, o indexador recorre automaticamente a esse quadro de referência que, quando rico de informações estocadas sobre tema da área específica tratada, certamente vai auxiliá-lo no seu trabalho de compreensão, o mesmo não ocorrendo no caso de o objeto de análise vir de uma área totalmente desconhecida para ele" (NAVES, 2000, p. 89, grifo nosso).
Relação Semântica	Indexador *analisa* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Texto *é analisado por* indexador.

Relação Semântica 22 (Relação não identificada pelo Semantizar)

Conceitos	Indexador e Texto
Frase	"Em síntese, este capítulo mostra o processo em que o indexador faz a leitura de um texto, empreende a extração de conceitos e determina a sua atinência". (NAVES, 2000, p. 69, grifo nosso)
Relação Semântica	Indexador *determina atinência do* texto.
Tipo da Relação	Associativa - Causal - Agente-Objeto
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível
Observação	Para essa frase, o Semantizar detectou a presença dos pares de conceitos: indexador e conceito e, conceito e texto. Contudo, ao realizar a análise dessa frase no Quadro 9, percebeu-se que nela também poderia existir a relação entre indexador e texto. E mais, constatou-se que duas relações semânticas poderiam existir entre indexador e texto, que é indexador *lê* texto (Quadro 11) e a relação explicitada nesse Quadro: Indexador *determina atinência do* texto.

Relação Semântica 23

Conceitos	Conceito e Pensamento
Frase	"FARRADANE (1980), em seu estudo do escopo da Ciência da Informação, faz uma distinção entre conceitos e palavras individuais, e acha que *conceitos individuais* são elementos de pensamentos e *palavras individuais* são somente rótulos para os conceitos tratados na mente" (NAVES, 2000, p. 57, grifo nosso).
Relação Semântica	Conceito *é elemento do* pensamento.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Pensamento *tem elemento* conceito.

Relação Semântica 24

Conceitos	Conceito e Documento
Frase	"Este estudo pretende concentrar-se nos aspectos do segundo sentido acima citado e, neste caso, a indexação acadêmica compreende duas etapas distintas: a Análise de assunto, quando ocorre a extração de conceitos que possam representar o conteúdo de um documento, […]" (NAVES, 2000, p. 26, grifo nosso).
Relação Semântica	Conceito *representa conteúdo do* documento.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 25

Conceitos	Conceito e Documento
Frase	“O processo de ler um documento para extrair conceitos que traduzam a sua essência é conhecido como "Análise de assunto", para alguns, como análise temática, para outros, ou, ainda, como análise documentária, análise conceitual ou, mesmo, análise de conteúdo” (NAVES, 2000, p. 35, grifo nosso).
Relação Semântica	Conceito *é extraído do* documento.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.
Observação	Essa relação também foi encontrada na frase: “No entanto, a concepção orientada pela demanda já pode ser vista como uma fase posterior à Análise de assunto propriamente dita, considerando ser essa a etapa em que a preocupação é traduzir os conceitos extraídos do documento para os termos de uma linguagem de indexação” (NAVES, 2000, p. 35, grifo nosso).
	Idem para essa frase: “Verifica-se, no exame da literatura especializada em Biblioteconomia e Ciência da Informação, que o termo "Análise de assunto" é o mais comumente utilizado, mas que grande parte dos autores que tratam do tema o consideram ou como a etapa de tradução dos conceitos extraídos dos documentos para um vocabulário controlado, ou até mesmo do processo de indexação como um todo” (NAVES, 2000, p. 38, grifo nosso).

Relação Semântica 26

Conceitos	Conceito e Documento
Frase	“No entanto, a concepção orientada pela demanda já pode ser vista como uma fase posterior à Análise de assunto propriamente dita, considerando ser essa a etapa em que a preocupação é traduzir os conceitos extraídos do documento para os termos de uma linguagem de indexação” (NAVES, 2000, p. 35, grifo nosso).
Relação Semântica	Conceito *é traduzido do* documento.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 27

Conceitos	Conceito e Documento
Frase	“Terminada a fase de extração dos conceitos, é necessário que se faça uma seleção daqueles que realmente sintetizem o assunto do documento […],” (NAVES, 2000, p. 64, grifo nosso).
Relação Semântica	Conceito *sintetiza assunto do* documento.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 28

Conceitos	Conceito e Texto
Frase	"Após essa leitura, passa-se à fase da extração de conceitos que possam representar o conteúdo temático do texto, para se chegar ao momento da fase de representação da atinência, […]" (NAVES, 2000, p. 40, grifo nosso).
Relação Semântica	Conceito *representa conteúdo temático do* texto.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 29

Conceitos	Conceito e Texto
Frase	"Esta pesquisa estuda a Análise de assunto que se realiza em textos escritos e, para indexá-los, é preciso que seja feita uma leitura, não uma leitura na íntegra, mas uma leitura que possibilite a extração de conceitos que sintetizem o conteúdo desses textos" (NAVES, 2000, p. 41, grifo nosso).
Relação Semântica	Conceito *sintetiza conteúdo do* texto.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.

Relação Semântica 30

Conceitos	Conceito e Texto
Frase	“Para definir em termos adequados o assunto de um texto é necessário que primeiro se extraiam os conceitos que nele estão contidos” (NAVES, 2000, p. 54, grifo nosso).
Relação Semântica	Conceito *é extraído do* texto.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível.
Observação	Essa relação semântica também foi detectada nas frases: “No processo de extrair conceitos de textos para definir seu assunto, o silogismo ocorre sempre, pois tanto a dedução quanto a inferência nele estão presentes” (NAVES, 2000, p. 87, grifo nosso) “Em síntese, este capítulo mostra o processo em que o indexador faz a leitura de um texto, empreende a extração de conceitos e determina a sua atinência” (NAVES, 2000, p. 69, grifo nosso)

Relação Semântica 31

Conceitos	Conceito e Texto
Frase	“diz respeito ao modo como os componentes do universo textual, ou seja, os conceitos e relações subjacentes ao texto de superfície são mutuamente acessíveis e relevantes entre si, entrando numa configuração veiculadora de sentidos” (NAVES, 2000, p. 46, grifo nosso).
Relação Semântica	Conceito *é parte do* texto.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Texto *tem* conceito.

Relação Semântica 31

Observação	Essa relação foi encontrada também na frase: “Manifesta-se, em grande parte, macrotextualmente, referindo-se aos modos como os componentes do universo textual, isto é, os conceitos e as relações subjacentes ao texto de superfície, se unem numa configuração, de maneira reciprocamente acessível e relevante” (NAVES, 2000, p. 46, grifo nosso)

Relação Semântica 32

Conceitos	**Texto e Idéia**
Frase	“Neste estudo, texto é considerado como o meio, o veículo que permite a comunicação de idéias entre o sujeito que cria e dissemina informação (emissor, no caso, o autor) e o sujeito que necessita e adquire informação (receptor, no caso, o leitor)” (NAVES, 2000, p. 41, grifo nosso).
Relação Semântica	Texto *comunica* idéia.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Idéia *é comunicada pelo* texto.

Relação Semântica 33

Conceitos	**Pensamento e Documento**
Frase	“Ele cita autores como Cutter (que define assunto como tema ou tópico, podendo ou não estar no título do documento), Kaiser (que toma "assuntos" como "coisas em geral", reais ou imaginárias, e as condições para designá-las, que são chamadas "concretos e processo"), Ranganathan (que fala sobre o pensamento contido no documento), […]” (NAVES, 2000, p. 59, grifo nosso).
Relação Semântica	Documento *contém* pensamento.
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo

Relação Semântica 33

Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Pensamento *está contido no* documento.

Relação Semântica 34

Conceitos	**Texto e Documento**
Frase	"A palavra texto é, muitas vezes, confundida com documento, discurso, informação, dado, conhecimento ou até mesmo como literatura, que é adotado por KAISER, citado por SVENONIUS (1978)" (NAVES, 2000, p. 41, grifo nosso).
Relação Semântica	Texto *é confundido com* documento.
Tipo da Relação	Equivalência - Quase sinônimo
Propriedades	Simétrica
	Irreflexiva
Relação Inversa	Documento *é confundido com* texto.

Relação Semântica 35

Conceitos	Texto e Documento
Frase	"No seu estudo de análise documentária, GINEZ DE LARA (1993) afirma que os textos veiculam informações de diversas naturezas e, para a documentação, tradicionalmente, eles constituem os documentos, que são a matéria prima da análise documentária" (NAVES, 2000, p. 42, grifo nosso).
Relação Semântica	Texto *constitui* documento.
Tipo da Relação	Equivalência - Quase sinônimo
Propriedades	Simétrica
	Irreflexiva
Relação Inversa	Documento *constitui* texto.

Relação Semântica 36

Conceitos	Documento e Informativo
Frase	"Fazem uma análise com o objetivo de determinar o conteúdo informativo do documento, tendo em vista o objetivo do sistema e as necessidades dos usuários; […]" (NAVES, 2000, p. 80, grifo nosso).
Relação Semântica	Documento *tem conteúdo* informativo
Tipo da Relação	Hierárquica - Merônimo-Holônimo - Objeto Estruturado - Componente-Complexo
Propriedades	Assimétrica
	Irreflexiva
Relação Inversa	Não é possível

APÊNDICE D - RELAÇÕES SEMÂNTICAS QUE NÃO PUDERAM SER EXPLICITADAS

Indexador e Documento

1. “o grau de concordância na representação da informação essencial do conteúdo do documento por certos grupos de termos de indexação, selecionados individualmente e independentemente, por cada indexador do grupo” (NAVES, 2000, p. 30).

2. “Um dos problemas que têm sido observados nos índices de assuntos, e que colocam em xeque o trabalho desenvolvido por indexadores, é o da variedade de termos que são definidos para representar o conteúdo de um mesmo documento num mesmo sistema, o que levanta a questão da consistência na indexação” (NAVES, 2000, p. 30).

3. “Para que ocorra a consistência, é necessário que haja uma coincidência de termos adotados para um mesmo documento, ou seja, é preciso que diferentes indexadores definam, para um determinado item indexado, o mesmo assunto, ou assuntos” (NAVES, 2000, p. 30).

4. “Quando é feita uma comparação entre os termos de indexação definidos para o mesmo documento, por dois ou mais indexadores, certamente são detectadas diferenças de julgamento a respeito de quais termos seriam os mais adequados” (NAVES, 2000, p. 30)

5. “Rolling afirma que Desde que consistência de indexação manifesta-se na similaridade dos termos de indexação determinados a um dado documento por diferentes indexadores, e desde que a seleção de termos de indexação por um indexador reflita seu julgamento quanto à informação contida no documento e sua representação, consistência da indexação é essencialmente a medida de similaridade de reação de diferentes seres humanos processando a mesma informação” (NAVES, 2000, p. 30)

6. “Assim, mais precisamente, consistência de indexação é definida como o grau de concordância na representação do conteúdo de informação (essencial) do documento por certas listas de termos de indexação selecionados individualmente e independentemente por cada indexador” (NAVES, 2000, p. 31)

7. “Ao analisar um documento, o indexador não deve limitar-se a representar ou resumir apenas a informação explícita no documento; mais do que isso, deve perguntar-se: Como eu poderia tornar esse conteúdo, ou parte dele, visível para o usuário potencial? Que termos deverei utilizar para levar esse conhecimento até o leitor interessado? (ALBRECHTSEN, 1993)” (NAVES, 2000, p. 37)

8. “Concepção orientada para o conteúdo - envolve uma interpretação adicional do conteúdo, que vai além dos limites da estrutura léxica e gramatical, com o estabelecimento de assuntos que não estão explicitamente colocados no texto, mas que são facilmente identificados pelo indexador, envolvendo, portanto, uma abstração mais indireta do documento;” (NAVES, 2000, p. 61).

Conceito e Texto

1. “A extração de conceitos tem como produto um assunto, que representa o conteúdo informacional de um texto” (NAVES, 2000, p. 58)

Indexador e Texto

1. “tradução: é requerida do indexador quando o texto é em língua estrangeira, adicionando pressões de um vocabulário, de uma gramática e de uma cultura, bem como as da mensagem do autor e das necessidades dos usuários; análise do texto: o trabalho de análise tem três componentes que envolvem o ato da comparação: (1) comparação do texto com outros textos indexados anteriormente, o que envolve o trabalho da memória; (2) comparação da amplitude de cada texto com a experiência pessoal do indexador como um todo; (3) ato de comparação entre o texto e as conhecidas necessidades e expectativas dos usuários” (NAVES, 2000, p. 18)

2. “Certamente, se indexadores abordam um texto apenas com a intenção de decidir a questão de assunto em sistema de vocabulário, eles podem perder algumas nuances que poderiam acrescentar aos subseqüentes termos do índice” (NAVES, 2000, p. 27)

3. “O processo de Análise de assunto, do ponto de vista do indexador, é iniciado com a fase de leitura do texto” (NAVES, 2000, p. 40)

4. Apesar da complexidade do estudo do significado e da sua relação direta com a atinência, acredita-se que ambos estejam relacionados exatamente ao momento em que o indexador diz: "O texto X trata de tal assunto” (NAVES, 2000, p. 68)

Conceito e Documento

1. “Pode-se dizer que o que coloca em risco a coerência da indexação é que distintos indexadores (ou até mesmo um indexador em momentos distintos) percebem de forma diferente: - o conteúdo real do documento; - a parte desse conteúdo que será suscetível de responder realmente às necessidades (inevitavelmente futuras) dos usuários; - os conceitos importantes que devem ser conservados para representar esse conteúdo; - os descritores definidos para representar esses conceitos” (NAVES, 2000, p. 29)

2. “Fazem uma análise com o objetivo de determinar o conteúdo informativo do documento, tendo em vista o objetivo do sistema e as necessidades dos usuários; (b) ao receberem um pedido de informação, quando, então, devem fazer uma análise deste com o objetivo de compreender a necessidade de informação transmitida pelo usuário, identificar os conceitos existentes no pedido e traduzi-los para a linguagem adotada pelo sistema” (NAVES, 2000, p. 35)

Ideia e Texto

1. “O leitor que domina as superestruturas textuais capta com mais facilidade as idéias centrais do texto, pois tem como parâmetro a identificação dos constituintes básicos” (NAVES, 2000, p. 43).

2. “A atividade de identificar a(s) idéia(s) principal(ais) do texto exige a capacidade de compreensão de seu conteúdo, o que está ligado a processos cognitivos” (NAVES, 2000, p. 60).

3. “A questão levantada pela autora é quanto à idéia principal de um texto e de como se pode definir essa noção” (NAVES, 2000, p. 60).

4. “Este último conceito encontra-se expresso em diversos vocábulos tais como mensagem do autor, visão de conjunto, elementos importantes, ponto de vista principal, idéia central do texto, havendo a respeito diversidade de concepções” (NAVES, 2000, p. 60).

5. “Essa informação importante pode variar segundo os tipos de texto: nos textos narrativos, a idéia principal tem a ver com os acontecimentos e a sua interpretação, enquanto nos textos

informativos o que é importante pode ser um conceito, uma generalização, uma regra" (NAVES, 2000, p. 61).

6. "O primeiro pode ser descoberto quando se pergunta de que trata um artigo, e o segundo quando se pergunta sobre qual é a coisa mais importante que o autor nos quer dizer, no texto, esperando-se, aí, obter, como resposta, uma idéia principal" (NAVES, 2000, p. 61).

Indexador e Conceito

1. "Pode-se dizer que o que coloca em risco a coerência da indexação é que distintos indexadores (ou até mesmo um indexador em momentos distintos) percebem de forma diferente: - o conteúdo real do documento; - a parte desse conteúdo que será suscetível de responder realmente às necessidades (inevitavelmente futuras) dos usuários; - os conceitos importantes que devem ser conservados para representar esse conteúdo; - os descritores definidos para representar esses conceitos" (NAVES, 2000, p. 29).

2. "SVENONIUS (1981) concorda com esse ponto de vista, e lembra que a aparente inconsistência, mostrada em estudos dos anos 60, resultou do fato de que os conceitos selecionados para indexação foram verbalizados diferentemente, por diferentes indexadores, em diferentes níveis de especificidade" (NAVES, 2000, p. 31).

Texto e Superestrutura

1. "Para CINTRA (1987), os constituintes básicos de um determinado tipo de texto é que definem a sua superestrutura" (NAVES, 2000, p. 43).

2. "Segundo Van Dijk (citado por KOBASHI, 1996), a superestrutura é considerada um elemento fundamental para a apreensão do significado do texto" (NAVES, 2000, p. 43).

3. "A superestrutura é, por definição, uma estrutura convencional, uma organização paradigmática, e o processo de compreensão supõe a transferência das unidades semânticas identificadas no texto para esse esqueleto conceitual" (NAVES, 2000, p. 43).

Conceito e Pensamento

1. "Conceito é a representação dum objeto pelo pensamento, por meio de suas características gerais" (NAVES, 2000, p. 55).

2. " Conceitos são conectados na mente por regras de pensamento, mas essas regras não devem ser confundidas com gramática e sintaxe, ou as regras da linguagem, que variam de uma linguagem para outra" (NAVES, 2000, p. 57).

Indexador e Prática

1. "Já no estudo de Bertrand & Cellier foram consideradas duas categorias de indexadores: indexadores com anos de experiência na indexação de livros e indexadores iniciantes, com conhecimento básico de regras de indexação, mas com pouca prática" (NAVES, 2000, p. 33).

2. "A prática profissional do indexador não exige que ele estoque na memória apenas a seqüência de eventos, como ocorre, freqüentemente, com os relatos da vida diária" (NAVES, 2000, p. 89).

Bibliotecário e Especialização

1. “Parece que, no caso brasileiro, pouco ou nada tem sido feito no sentido de melhorar a formação geral do bibliotecário, exigindo-se dele uma "especialização forçada" numa determinada área que ele não domina e, pior ainda, avaliando e criticando seu trabalho sem que se investiguem as possíveis causas de erros que possam ocorrer” (NAVES, 2000, p. 22).

Texto e Macroestrutura

1. “O modelo desses autores defende a visão de que o processamento de texto se baseia em redução proposicional, isto é, a redução do texto através de suas proposições para se chegar à macroestrutura (estrutura semântica global)” (NAVES, 2000, p. 80).

Profissional da Informação e Bibliotecário

1. “O profissional da informação tem uma imagem pública pobre e FLECK (citado por FLECK & BAWDEN, 1995) dá exemplos, formais e informais, do "infortunado estereótipo do bibliotecário” (NAVES, 2000, p. 15).

Indexador e Pensamento

1. “Para interpretar o conteúdo de um documento, é certo que o indexador passa por um processo cognitivo, e o autor acima citado acredita que a Ciência da Informação deve incluir estudos sobre seres humanos e, em particular, sobre os processos de pensamento e comportamento em todas as ações relativas à comunicação, se são origem dela, se são seus manipuladores ou são recipientes de comunicações” (NAVES, 2000, p. 86).

Texto e Hipertexto

1. “É o hipertexto, uma matriz de textos potenciais que se encontra num suporte digital e, por isso, permite novos tipos de leituras, nas quais os leitores podem criar, através de elos de ligação, seus próprios textos, modificando e acrescentando novos textos, imagens.” (NAVES, 2000, p. 48).