Carlos Duarte de Oliveira Junior

EXTRAÇÃO AUTOMÁTICA DE CONTEXTOS DEFINITÓRIOS EM TEXTOS ACADÊMICOS DA CIÊNCIA DA INFORMAÇÃO

Brasília

março de 2012

Carlos Duarte de Oliveira Junior

EXTRAÇÃO AUTOMÁTICA DE CONTEXTOS DEFINITÓRIOS EM TEXTOS ACADÊMICOS DA CIÊNCIA DA INFORMAÇÃO

Dissertação apresentada à Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre.

Orientadora: Prof^a Dr^a Marisa Bräscher Basílio Medeiros

UNIVERSIDADE DE BRASÍLIA – UNB
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO
CADUARTEJR@GMAIL.COM

Brasília

março de 2012

FOLHA DE APROVAÇÃO

Título: "Extração automática de contextos definitórios em textos acadêmicos da Ciência da Informação"

Autor (a): Carlos Duarte de Oliveira Junior

Área de concentração: Transferência da Informação

Linha de pesquisa: Arquitetura da Informação

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Mestre** em Ciência da Informação.

Dissertação aprovada em 29 de março de 2012.

Aprovado por:

Prof. Dr. Marisa Bräscher Basílio Medeiros
Presidente – (UnB/PPGCINF)

Prof. Dr. Cláudio Chauke Nehme
Membro Externo – (UCB)

Prof. Dr. Rogério Henrique de Araújo Júnior
Membro Interno – (UnB/PPGCINF)

Prof. Dr. Jorge Henrique Cabral Fernandes
Suplente – (UnB/PPGCINF)

Agradecimentos

A Deus, inteligência suprema, causa primária de todas as coisas, que, sem sua permissão, nada poderia ter sido alcançado.

Aos professores da Faculdade de Ciência da Informação da Universidade de Brasília, Mamede, Claudio Duque, Jorge Fernandes, Dulce e André, que me estimularam ao estudo na área.

Aos membros da banca, professor Rogério e o colega de trabalho, Claudio Chauke, por muito contribuírem para o aprimoramento da pesquisa.

Às meninas da secretaria, Jucilene e Martha, sempre muito prestativas e alegres.

Aos colegas de mestrado, por compartilharem seus conhecimentos e desafios e, em especial, aos amigos que fiz, Georgia, Leonardo, Thalita e Raphael.

Ao professor Jaime Robredo, meu primeiro orientador, que, com sua experiência e tranquilidade, muito acrescentou à pesquisa e ao pesquisador. Foi chamado por Deus e espero que continue seu trabalho em outra esfera.

À professora e amiga Marisa Bräscher, que, desde o início, me cativou com sua simplicidade e capacidade, me orientando durante o curso e na fase final desta dissertação.

Aos amigos de infância, colegas de trabalho e parentes, em especial a meu querido amigo/irmão André Siqueira, que me auxiliou e socorreu em momentos difíceis desta caminhada.

Ao meu sogro Jorge e minha sogra Lourdes, pela guarida e incentivo na fase final da pesquisa.

Aos meus amados Pai, Carlos Duarte, e Mãe, Cristina, por me darem amor e me ensinarem as duas coisas mais importantes da vida, que são confiar em Deus e procurar ser uma pessoa de bem. Aos meus irmãos, Thiago e Lucas, e irmãs, Meriele e Nayara, por me compreenderem e me aceitarem, e pelo carinho de sempre. Aos agregados também, pois fazem parte da minha família, Jorge e Déborah.

Em especial, ao meu querido filho Mateus, que me deu um novo sentido para a vida, me faz ver, todos os dias, a beleza da existência. E a minha querida e amada esposa Fabiane, pelo amor, amizade e companheirismo durante a busca deste objetivo.

Muito obrigado!

“Mas não se chegará ao progresso da humanidade se não se atacar o mal pela raiz, ou seja, pela educação. Não essa educação que tende a fazer homens instruídos, mas a que tende a fazer homens de bem. A educação, se for bem compreendida, será a chave do progresso moral."

(Allan Kardec, Livro dos Espíritos, Q. 685)

Resumo

O trabalho apresenta estudo sobre o papel da Ciência da Informação, sua interdisciplinaridade e interseção com a Linguística e a Ciência da Computação no que se refere à utilização dos textos como fonte de informação e conhecimento a ser organizado ou reorganizado, nos grandes repositórios de informação já existentes, com a finalidade de recuperação. A ênfase é na extração automática de Contextos Definitórios (CD) em textos, o que se entende como qualquer fragmento textual que introduz e associa um termo a uma definição. Cita teorias de Organização da Informação como Classificação Facetada de Ranganathan, a teoria do Conceito de Dahlberg e as teorias da terminologia, tais como a Teoria Geral da Terminologia de Wüster e a Teoria Comunicativa da Terminologia de Cabré. Todas as teorias são abordadas com enfoque na importância do termo e principalmente da definição como elemento primordial para o mapeamento semântico de um documento e de um domínio do conhecimento. Enfatiza a visão da definição como elemento de ligação entre os objetos e seus conceitos, identifica tipos de definições, cita estudos anteriores de identificação e extração automática de enunciados definitórios em inglês, espanhol e francês. Menciona as técnicas de Processamento de Linguagem Natural e Descoberta de Conhecimento em Textos como ferramentas para o processamento e extração de informação em documentos escritos em língua natural. Por fim, propõe um método de extração automática de Contextos Definitórios em textos acadêmicos da Ciência da Informação, a partir de uma gramática de padrões definitórios em língua portuguesa criada no âmbito da pesquisa. Entende-se gramática de padrões definitórios como um conjunto de expressões linguísticas capazes de identificar um CD em um texto. A gramática foi validada comparando uma extração manual com uma automática. O método foi aplicado nas teses e dissertações da Faculdade de Ciência da Informação da Universidade de Brasília - UNB, disponibilizadas a partir de seu repositório RIUnb, de 2006 a 2011.

Palavras-chave: Contexto ou Enunciado definitório; Definição terminológica; Organização da Informação e do Conhecimento; Processamento de Linguagem Natural PLN; Descoberta de Conhecimento em Textos DCT; Métodos linguísticos na Ciência da Informação.

Abstract

The paper presents a study on the role of Information Science, and its interdisciplinary intersection with Linguistics and Computer Science with regard to the use of texts as a source of information and knowledge to be organized or reorganized, in large information repositories existing, with recovery purposes. The emphasis is on automatic extraction of Definitory Context (DC) in texts, which is understood as any fragment of text that introduces and associate a term with a definition. It makes reference to the theories of Information Organization and Faceted Classification of Ranganathan's theory of concept Dahlberg and theories of terminology, such as the General Theory of Terminology of Wüster's and the Communicative Theory of Terminology of Cabré's. All theories are discussed with emphasis on the importance of the term and the definition as a major element for the semantic mapping of a document and a domain of knowledge. It emphasizes the view of the definition as a liaison between the objects and their concepts, identifies types of settings, make reference to previous studies of automatic identification and extraction of Definitory Enunciation in English, Spanish and French. It mentions techniques of Natural Language Processing and Knowledge Discovery in Texts as tools for processing and extraction of information in documents written in natural language. Finally, it proposes a method for automatic extraction of Contexts in academic texts of Information Science, from a Definitory grammar patters in Portuguese established within the research. It is understood that Definitories grammar patters as a set of that linguistic expressions can identify a DC in a text. The grammar was validated by comparing an automatic with a manual extraction. The method was applied in thesis and dissertations at the Faculty of Information Science at the University of Brasilia - UNB, available from it repository RIUnB, from 2006 to 2011.

Key words: Context or Enunciation Definitory; Terminological Definition; Organization of Information and Knowledge; Natural Language Processing NLP, Knowledge Discovery in Texts DCT; linguistic methods on Information Science.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A Ciência da Informação para Robredo (2003, p.105) é "o estudo, com critérios, princípios e métodos científicos, da informação", sendo, portanto, a informação o objeto de estudo desta Ciência. Contudo, esse termo, tem sido usado para as mais diversas situações, além disso, a sua relação com o dado e o conhecimento também gera dúvida quanto aos limites de cada um, o que solicita a definição adotada por esta pesquisa para estes elementos.

Siqueira (2008, p.92), ao analisar a tecnologia e a natureza da informação propõe a seguinte hierarquia :

1. **a informação sintática** – aquela que não possui significado contextual. É um signo sintático cuja forma é objeto de observação;
2. **a informação semântica** – aquela que possui significado contextual para um sujeito;
3. **a informação pragmática** – aquela que está codificada e preparada para uso. É uma informação manipulada por um sujeito com fins de utilidade planejada.

Esta pesquisa adota este entendimento sobre o dado (informação sintática), a informação (informação semântica) e o conhecimento (informação pragmática) proposto por Siqueira (2008), pois considera mais adequada essa caracterização para o estudo da informação registrada.

A Ciência da Informação tem sua origem na biblioteconomia, em especial nas áreas de documentação e recuperação da informação, e seu surgimento está intimamente ligado à revolução científica e técnica que se seguiu à II Grande Guerra (SARACEVIC, 1996a; SOUSA, 2007). Para Saracevic (1996b), um marco histórico quanto a origem é o artigo de Bush (1945) que identificou o problema da explosão informacional, o crescimento exponencial da informação e de seus registros, particularmente em ciência e tecnologia e a dificuldade de acesso rápido à informação relevante e propôs usar as tecnologias de informação para combater o problema.

Nos dias de hoje o problema é basicamente o mesmo, muita informação sendo gerada e registrada, muitos documentos sendo armazenados em repositórios, mas a dificuldade de acesso à informação contida nos documentos continua. Em artigo recente Saracevic (2009) aborda três questões que mapeiam e norteiam os estudos em Ciência da Informação:

1. a questão física: quais são as características e as leis do universo de informações registradas?

2. a questão social: como as pessoas podem relacionar-se, buscar, e fazer uso da informação?

3. a questão de *design*: como é possível tornar mais rápido e eficaz os acessos aos registros da informação?

O terceiro item, *design*, se caracteriza como um dos maiores desafios e nos remete ao processo anterior a recuperação que é a organização. Taylor A. G.; Joudrey (2009, p.02), ao introduzirem a tema da Organização da Informação, em seu livro, apresentam uma intrigante questão: O que estamos organizando em nossas bibliotecas, museus, arquivos e semelhantes, informação ou conhecimento? Um questionamento anterior se faz pertinente, aumentando a avaliação sobre nossas instituições e sobre os processos de estruturação da informação. Estamos organizando em nossos sistemas: informação, conhecimento ou documentos?

Capurro R.; Hjorland (2007) e Frei (1996) questionam a Recuperação da Informação (RI) e afirmam que, na maioria das vezes, a RI faz uma recuperação de documentos ou referências, e não de informação. O problema está no tipo de organização dos repositórios que, a princípio, tinham como objetivo organizar os documentos, não favorecendo a recuperação efetiva da informação. Nessa direção encontram-se os repositórios institucionais acadêmicos, que representam sua produção científica e acadêmica em textos e armazenam em documentos (artigos, teses e dissertações) que podem ser acessados por meio dos seus sistemas de busca.

Para a efetiva recuperação da informação é necessário utilizar os métodos e técnicas para organização da informação e do conhecimento. Várias teorias a respeito de como adentrar nos documentos e representar a informação e o conhecimento ali contidos têm sido estudadas e utilizadas. Nesse sentido, podemos citar a Classificação Facetada de Ranganathan e a Teoria do Conceito de Dahlberg (1978b), bem como as teorias terminológicas que, assim como as teorias de Ranganathan e Dahlberg, estudam a função dos termos e a necessidade de obter suas definições para representar os conceitos tratados em determinada área do saber, facilitando a comunicação entre seus especialistas.

Para os repositórios já constituídos, com um volume grande de documentos já catalogados, a reorganização manual dos seus sistemas se torna muito difícil devido a falta de recursos humanos para executar essa atividade. Assim, mecanismos de automação de processos que auxiliem essa reorganização são objeto de estudo da Ciência da Informação como Araújo Jr. (2007), Schiessl (2007), Câmara Jr. (2007) e Capuano (2010).

Nesse contexto, a presente pesquisa propõe um método para adentrar nos documentos do Repositório Institucional da UNB, o RIUnb, extrair informações que auxiliem a reorganização do seu repositório, e criar novas visões sobre o conhecimento ali representado. As teorias citadas acima para representação da informação e do conhecimento identificam os termos e as definições como elementos fundamentais para organizar a informação e o conhecimento contido em um documento e de um domínio do conhecimento (DAHLBERG, 1978b; CAMPOS, 2001; CABRÉ, 2003; LARA, 2004; ALMEIDA; ALUÍSIO; OLIVEIRA, 2007; FRANCELIN, 2010). Assim sendo, a presente pesquisa escolheu a extração de definições contidas em textos de especialista para o estudo.

Vários pesquisadore tem se debruçado sobre a extração de definições ou parte de definições em textos; Podemos mencionar em língua inglesa, as pesquisas de Pearson (1998), Meyer (2001) e Rodriguez (2004). Em língua francesa, temos os trabalhos de Auger (1997) e Marshman (2003) e em língua espanhola, os estudos de Sierra e Alarcón (2003), Aguilar (2009), Alarcón (2009). Em português, poucos trabalhos foram identificados, podemos citar o trabalho das portuguesas Pinto e Oliveira (2004), que analisam um córpus em português de portugual.

Sierra e Alarcón (2003), a partir dos estudos de Meyer (2001) e Rodriguez (1999) e no âmbito do projeto coordenador pelo professor Gerardo Sierra, do Grupo de Engenharia Linguística, da Universidade Nacional do México – UNAM, propõem uma estrutura lingüística para identificação de uma definição, o que eles chamam de Contexto Definitório (CD).

Aguilar (2009) entende Contexto Definitório como qualquer fragmento textual onde se introduza e associe um termo a uma definição. Os CDs são compostos de um termo (T) e uma definição (D) que se encontram conectados mediantes a um padrão definitório (PD). Esses CDs podem apresentar outros tipos de informações metalingüísticas e pragmáticas referentes à forma, condições de uso ou alcance operativo, o que foi denominado por padrão pragmático (PPR) (SIERRA, 2009).

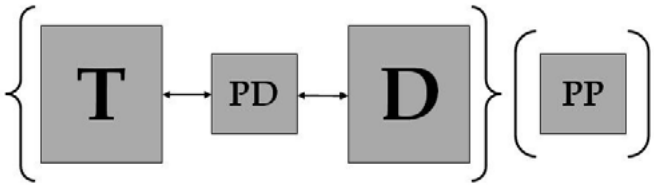

Figura 1: Estrutura de um Contexto Definitório
Fonte: (SIERRA, 2009)

Exemplo: <PPR> Tradicionalmente </PPR>, <T>la logística </T> <PD> se define como </PD><D> el arte militar que estudia el movimiento, transporte y estacionamiento de las tropas fuera del campo de batalla</D>.[1] (SIERRA, 2009, p.17)

Para a extração automática de CDs é necessária uma gramática de padrões definitórios que, segundo Sierra e Alarcón (2003), são expressões linguísticas capazes de identificar um contexto definitório em textos. Como não existe nenhuma gramática de padrões definitórios em língua portuguesa, um dos resultados desse estudo é a proposição de uma gramática para os documentos investigados no âmbito da pesquisa.

Sendo assim, coloca-se a seguinte questão: é possível identificar um padrão linguístico em textos acadêmicos que demarquem a presença de uma definição, possibilitando sua extração de forma automática? Outro ponto de questionamento seria: como validar esse padrão ou essa gramática?

A análise manual de textos acadêmicos e a comparação com uma identificação automática são um dos meios metodológicos utilizados na pesquisa para que o objetivo proposto seja alcançado.

[1] Tradução nossa: Tradicionalmente, a logística se define como a arte militar que estuda o movimento, transporte e estacionamento das tropas fora do campo de batalha.

Sobre o problema da pesquisa

2 Objetivos

2.1 Objetivo Geral

Propor um método de extração automática de contextos definitórios em textos acadêmicos por meio do uso de padrões da língua portuguesa observados em documentos contidos no repositório da Faculdade de Ciência da Informação da Universidade de Brasília - UNB.

2.2 Objetivos Específicos

1. Construir uma gramática de padrões definitórios para textos da Ciência da Informação em língua portuguesa a partir dos trabalhos de Sierra e Alarcón (2003) e Kamikawachi (2009).

2. Validar a gramática proposta através da comparação dos contextos definitórios extraídos de forma automática com grupo de contextos identificados de forma manual.

3. Identificar de forma automática os contextos definitórios (CDs) nas teses e dissertações da Faculdade de Ciência da Informação da Universidade de Brasília - UNB, contidas em seu repositório, RIUnB.

3 Justificativa

Obter conhecimento sempre foi um objetivo para os seres humanos, portanto, buscá-lo e comunicá-lo, tornou-se um fenômeno básico das sociedades em todas as épocas. Contudo, Capurro R.; Hjorland (2007) comentam que o surgimento da tecnologia da informação e seus impactos globais é que caracterizam a nossa sociedade como uma sociedade da informação. O autores dizem que:

> É lugar comum considerar-se a informação como condição básica para o desenvolvimento econômico juntamente com o capital, o trabalho e a matéria-prima, mas o que torna a informação especialmente significativa na atualidade é sua natureza digital (CAPURRO R.; HJORLAND, 2007, p. 02).

A facilidade de ferramentas de digitalização de textos, a crescente utilização dos meios de comunicação mediada por computador (CMC), como fóruns, *chats, blogs e e-mail*, como mecanismos de comunicação empresarial, além do registro de reuniões através de áudio e subsequente transcrição das gravações em texto, contribuem para o aumento significativo de informações produzidas em linguagem natural.

Segundo Santos (2001), a sociedade atual possui uma enorme quantidade de textos armazenados, porém não consegue acessar o conhecimento contido neles. Nessa direção, encontram-se também os repositórios institucionais acadêmicos. A produção científica e acadêmica das instituições de ensino superior (IES) é representada em textos e armazenada em documentos (artigos, teses e dissertações) que podem ser acessados através dos departamentos que têm o papel de organizar e disseminar essas informações, as bibliotecas e seus repositórios. Contudo, o modelo de organização e recuperação da informação, geralmente adotado pelas bibliotecas universitárias, possibilita o acesso aos documentos e não às informações neles expresso.

Capurro R.; Hjorland (2007) e Frei (1996) questionam a Recuperação da Informação (RI) e afirmam que, na maioria das vezes, a RI faz uma recuperação de documentos ou referências, e não de informação. O problema está no tipo de organização dos repositórios que, a princípio, tinham como objetivo organizar os documentos, não favorecendo a recuperação efetiva da informação.

A constatação desses fatos e o exponencial aumento de documentos portadores de informação armazenados nos repositórios têm gerado desafios no sentido da criação de modelos novos de organização da informação e de mecanismos automáticos para reorganizar os repositórios já existentes, a fim de recuperar e disponibilizar informação aos usuários, visto que os atuais processos manuais não conseguem suprir as demandas. A comunidade científica tem percebido isso e, nos dias 8 e 9 de maio de 2006, foi promovido um seminário pela Sociedade Brasileira de Computação (SBC) para identificar os desafios da computação para os próximos 10 anos, ou seja, de 2006 a 2016. Os cinco desafios propostos foram:

1. Gestão da Informação em grandes volumes de dados multimídia distribuídos;
2. Modelagem computacional de sistemas complexos artificiais, naturais e socioculturais e da interação homem-natureza;
3. Impactos para a área da computação da transição do silício para novas tecnologias;
4. Acesso participativo e universal do cidadão brasileiro ao conhecimento;
5. Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos.

O Processamento de Linguagem Natural (PLN), que visa propiciar à máquina "entender"a linguagem humana, está presente na maioria dos desafios identificados pela SBC para os próximos 5 anos, entretanto, é uma área de pesquisa ainda incipiente na língua portuguesa.

Poucos grupos de pesquisa no Brasil se dedicam a estudar e publicar artigos sobre esse tema. A Sociedade Brasileira de Computação criou uma Comissão Especial de Processamento de Linguagem Natural (CE-PLN) que, em 2009, fez um mapeamento da área no Brasil através de uma enquete *online* de ampla divulgação (NUNES, 2009). No total, apenas 148 pesquisadores responderam à enquete, sendo que, aproximadamente 2/3 desse grupo, considera o PLN como sua principal área de estudo.

Do universo de respondentes (148), destacamos ainda que apenas 01(um) tem formação em Ciência da Informação. A partir dessa premissa, observa-se a carência de pesquisadores na área, como demostra a Figura 2.

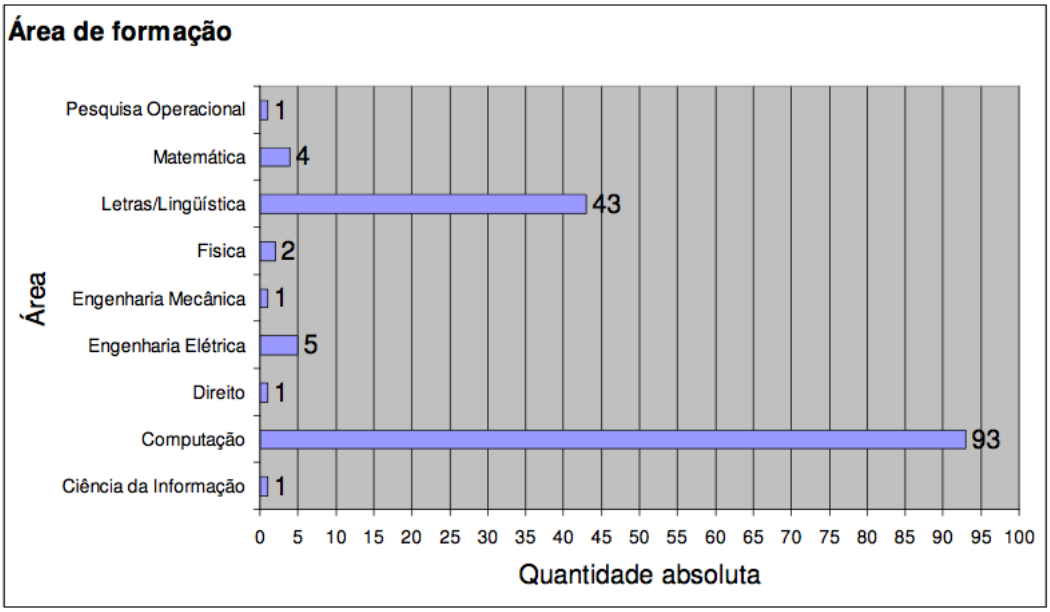

Figura 2: Distribuição de pesquisadores em PLN por área de formação
Fonte: Nunes (2009)

Saracevic (1996a, p.40), ao explicar sua abordagem do ponto de vista da problemática sobre a Ciência da Informação (CI), em importante artigo intitulado “Ciência da Informação; origem, evolução e relações”, cita um argumento de POPPER (1972) de que "... não somos estudantes de assuntos, mas estudantes de problemas. E os problemas constituem os recortes de qualquer assunto ou disciplina". Para Saracevic (1996a, p.41), então, “um campo é definido pelos problemas que são propostos”.

O problema da falta de recursos humanos frente ao volume de documentos gerados e já armazenados está posto e a Ciência da Informação, através de vários estudos, como de Araújo Jr. (2007), Schiessl (2007), Câmara Jr. (2007) e Capuano (2010) ,demonstra preocupação em estudar métodos de automação nos processos de organização, recuperação, descobrimento e indexação das informações contidas nesses repositórios textuais; Contudo, a carência de pesquisas nessa área ainda é muito grande.

A Ciência da Informação tem como um dos seus objetivos otimizar o acesso à informação relevante ou que venha ao encontro à necessidade do usuário (SARACEVIC, 2009). Para se recuperar informações relevantes é necessário organizar não apenas os termos descritos nos textos, mas os conceitos tratados. Para Dahlberg (1978b, p.01) na teoria do conceito, “Cada enunciado verdadeiro representa um elemento do conceito” e a definição trata de determinar ou fixar os limites de um conceito ou idéia. Francelin (2010) coloca que a identificação destes enunciados possibilita, mesmo sem um nome que designe um conceito, saber o que ele

é e formulá-lo pelo conjunto de suas características. Entende-se também que a pesquisa, ao criar um método de extração e armazenamento de contextos definitórios (SIERRA; ALARCÓN, 2003) ou exertos definitórios, segundo Almeida, Aluísio e Oliveira (2007), auxiliará a Ciência da Informação a reorganizar os repositórios, apoiar a criação de tesauros e ontologias, e consequentemente, possibilitará uma maior efetividade na Recuperação da Informação, um de seus principais objetivos.

Acredita-se ainda que o estudo aplicado no repositório da Faculdade de Ciência da Informação da UNB, que contêm uma memória riquíssima da produção técnica e acadêmica da área, trará grandes contribuições, pois um dos problemas básicos de qualquer Ciência é a organização de seus conceitos e a definição de seus princípios. Ao minerar a produção científica dos autores da área, seus pesquisadores, e extrair contextos definitórios, a pesquisa poderá trazer elementos significativos para o mapeamento da área.

A criação de uma gramática de padrões definitórios em língua portuguesa, que conforme Sierra e Alarcón (2003) são expressões linguísticas capazes de identificar um contexto definitório em textos, está estruturada no nível sintático, por isso, também é um resultado importante da pesquisa, visto que pode ser aplicada em diversos domínios do conhecimento registrado.

Por fim, entende-se que o estudo de padrões linguísticos na representação de definições em textos acadêmicos da Faculdade da Ciência da Informação da Universidade de Brasília pode contribuir com os desafios acima citados e auxiliar as pesquisas que visam a abordar o auxílio do computador na automação de processos de organização, recuperação, representação e extração da informação em língua portuguesa.

4 Metodologia

A proposta da pesquisa consiste na identificação automática de Contextos Definitórios (CD) nas teses e dissertações da Faculdade de Ciência da Informação da Universidade de Brasília – UNB, por intermédio de uma gramática de padrões definitórios para a língua portuguesa, criada no escopo desta investigação.

4.1 Classificação da Pesquisa

Segundo Gil (1999), é possível agrupar as pesquisas científicas em 3 grandes grupos: pesquisas descritivas, explicativas e exploratórias. As pesquisas descritivas objetivam a descrição das características de determinada população ou fenômeno, ou o estabelecimento de relações entre variáveis. As explicativas são aquelas que têm como preocupação central identificar os fatores que determinam para a ocorrência dos fenômenos. Por fim, as pesquisas exploratórias têm como finalidade desenvolver, esclarecer e modificar conceitos e idéias, tendo em vista a formulação de problemas mais precisos ou hipóteses verificáveis para estudos posteriores.

Esta pesquisa pode ser considerada exploratória por estudar o Processamento de linguagem natural (PLN) aplicado na descoberta de conhecimento em textos, especificamente na técnica de Extração da Informação para retirar Contextos Definitórios (CD) em língua portuguesa de maneira automática, algo muito pouco estudado, principalmente, no âmbito da Ciência da Informação. Porém, também é do tipo descritiva, pois visa conhecer uma realidade, quantificá-la e interpretar os fatos observados sem alterar o fenômeno estudado.

4.2 Fórmula para cálculo amostral

Para composição de amostra é importante conhecer o desvio padrão, contudo Cochran (1977) consideram que existem três situações possíveis para se determinar essa variável:

– Quando se pode estimar a variação populacional por meio de um levantamento pi-

loto.

– Quando a estimativa pode ser feita com o auxílio de pesquisas prévias.

– Quando não existe possibilidade de estimar. Neste caso é possível utilizar a fórmula: máximo de ocorrências possível, menos o mínimo de ocorrências, dividido por 4.

Sendo assim, a fórmula para a amostra é:

$$n_0 = \frac{\alpha^2 \sigma^2}{e_0^2}$$

Quando não podermos estimar o desvio padrão:

$$\frac{\alpha^2 (\frac{max-min}{4})^2}{e_0^2}$$

Porém, esta fórmula é adequada quando não se sabe o tamanho da população, quando se conhece, se faz uma correção por:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Sendo:

– α = valor da distribuição normal para o nível de confiança desejada. É uma constante, sendo mais usual os níveis de 90%, 95% ou 99% de confiança.

– σ = estimativa do desvio padrão.

– e_0 = erro amostral tolerável que é escolhido pelo pesquisador.

– N = Tamanho da população.

Gracio e Oliveira (2005) ao descreverem sobre o uso destas fórmulas aplicadas na área de Ciência da Informação, criaram vários exemplos de uso das diversas fórmulas conforme a variação das variáveis e das situações citadas acima.

4.3 Amostra

Para definição da amostra para a pesquisa, levou-se em consideração que o padrão sintático para identificação de um contexto definitório seria melhor percebido em uma base de especia-

listas, por conter um padrão de escrita formal e estruturado, além de possuir documentos com maior probabilidade de se identificar contextos definitórios, pois se trata de uma comunicação de especialista para profissionais da área ou de especialista para principiante (PEARSON, 1998). Optou-se então, por utilizar a base do Repositório Institucional da UNB.

A Faculdade de Ciência da Informação continha 378 documentos até o final do ano de 2011, divididos nas seguintes coleções: artigos publicados em periódicos, livros e capítulos de livros e trabalhos apresentados em eventos. Uma sub-comunidade, com nome de pós-graduação, armazena as teses e dissertações da faculdade. Essa sub-comunidade é o grupo amostral utilizado por essa pesquisa. O fluxo a seguir, figura 2, demonstra o processo para escolha dos documentos a serem utilizados no presente trabalho.

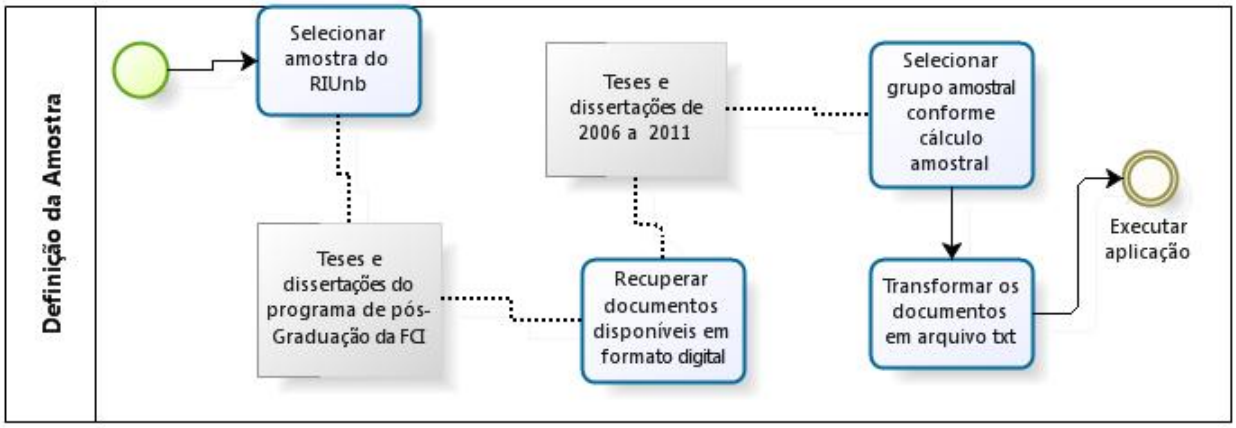

Figura 3: Fluxo de definição da amostra
Fonte: Elaboração do autor

A primeira tarefa foi identificar no repositório da UNB o grupo que seria fruto da amostra. Por se tratar de uma varredura no texto completo dos documentos, optou-se por trabalhar com um grupo não muito grande de documentos. Entre os departamentos e faculdades que compõem o repositório, o programa de pós-graduação da Faculdade de Ciência da Informação (FCI) foi selecionado e entre os tipos de documentos disponibilizados, as teses e dissertações serão o material utilizado na pesquisa.

Como atividade posterior, identificamos na FCI os documentos que estão em formato digital, o que possibilita serem processados automaticamente. Todos os documentos disponíveis no RIUnb estão em formato digital, ou seja, o estudo utiliza teses e dissertação com defesa a partir do ano de 2006 até 2011, em um total de 179 documentos.

Desse grupo, aplicou-se a fórmula apresentada na metodologia, item 4.2, com 95% de grau

de confiança, erro amostral considerado de 13, um máximo de 150 e um mínimo de 65 ocorrências em cada documento para uma população de 179 documentos. Identificou-se que seria necessária a analise manual de 10 documentos para criação da gramática, como será detalhado no percurso metodológico na próxima seção.

O terceiro e último passo do fluxo é a transformação dos documentos disponibilizados em formato .pdf para arquivos em formato texto e passíveis de manipulação por aplicações computacionais para investigação textual.

4.4 Percurso Metodológico

Para alcançar os objetivos propostos na pesquisa, os seguintes passos foram executados:

1. Levantamento do referencial teórico que fundamenta a extração automática de conteúdos em documentos textuais, em especial as definições;
2. Avaliação da gramática de padrões definitórios, segundo Sierra e Alarcón (2003), aplicada à língua espanhola;
3. Avaliação das expressões linguísticas identificadas em língua portuguesa na composição de definições, conforme Kamikawachi (2009);
4. Identificação das expressões linguísticas utilizadas para contextos definitórios em um grupo amostral de textos da base RIUnB;
5. Adaptação da gramática de padrões definitórios à língua portuguesa – um dos resultados desta dissertação – com base nas expressões linguísticas identificadas;
6. Proposição de um método de extração automática de contextos definitórios através do uso da gramática adptada, das expressões linguísticas identificadas por (KAMIKAWACHI, 2009) e pela análise dos documentos amostrais da base do RIUnB;
7. Identificação de contextos definitórios nos textos em língua portuguesa da Faculdade de Ciência da Informação da UNB – uma aplicação do resultado.

Na fase de levantamento bibliográfico da pesquisa, um estudo exploratório sobre as informações existentes a respeito da identificação automática de contextos definitórios foi realizado. Esse estudo permitiu identificar a multidisciplinalidade do problema proposto, visto que a maioria dos trabalhos correlatos, além de estarem em língua diferente do português, são da área da

Linguística, Terminologia e Ciência da Computação. Por esse motivo, foi acrescido à revisão de literatura um capítulo sobre a a Ciência da Informação e sua abordagem sobre os conceitos tratados na pesquisa.

4.5 Detalhamento do Percurso Metodológico

O detalhamento dos passos percorridos pela pesquisa visa delinear de maneira clara o que cada objetivo pretendia alcançar, o que permitiu maior precisão na medição dos resultados do estudo.

1. Levantamento bibliográfico que fundamenta a extração de conteúdo em textos. A fase de revisão de literatura, fundamentou-se em autores clássicos da Ciência da Informação, Linguistica, Terminologia e Ciência da Computação e trabalhos recentes sobre o tema. A revisão está dividida em três etapas:

    – A Ciência da Informação - tem como objetivo organizar, recuperar e disponibilizar o conhecimento registrado. Nesse sentido, foram realizados nesta etapa o levantamento bibliográfico sobre o conceito de informação e conhecimento, tipos de organização da informação e do conhecimento, teorias de análise de conteúdo para representação da informação e do conhecimento e o papel das definições nessas teorias.

    – Contexto Definitório - foram fruto desta etapa, um levantamento exaustivo de estudos sobre a caracterização de uma definição, seus conceitos e tipos, sua presença em textos de especialistas, além do conteúdo sobre o que é e como é composto um Contexto Definitório. Os trabalhos do grupo de Engenharia Linguística da Universidade Nacional do México – UNAM, Sierra e Alarcón (2003), Aguilar (2009) e Alarcón (2009), além dos trabalhos de Auger (1997), Pearson (1998), Pustejovsky (1991), Seppälä (2004) e Meyer (2001) foram relatados neste capítulo.

    – Extração de Contextos Definitórios - são apresentadas neste capítulo, as técnicas e métodos de manipulação automática da informação armazenada em textos como o Processamento de Linguagem Natural, com foco na manipulação de textos, e a Descoberta de Conhecimento em textos, com ênfase no método de Extração da Informação, além dos elementos utilizados em trabalhos anteriores de identificação e extração de estruturas definitórias como

padrões definitórios e gramáticas definitórias em língua espanhola Sierra e Alarcón (2003), Sierra (2009), em língua francesa Auger (1997), em inglês com Rodriguez (2004) e as expressões linguísticas em português encontradas no trabalho de Kamikawachi (2009) junto a uma base de definições do grupo Geterm, da Universidade de São Carlos.

2. Identificação de contextos definitórios nos textos em língua portuguesa da Faculdade de Ciência da Informação da UNB.

   – Identificação das expressões linguísticas utilizadas para contextos definitórios em um grupo amostral de textos da base RIUnB;

   – Adaptação da gramática de padrões definitórios à língua portuguesa com base nas expressões linguísticas identificadas;

   – Proposição de um método de extração automática de contextos definitórios através do uso da gramática adptada, das expressões linguísticas identificadas por (KAMIKAWACHI, 2009) e pela análise dos documentos amostrais da base do RIUnB;

   – Identificação de contextos definitórios nos textos em lingua portuguesa da Faculdade de Ciência da Informação da UNB – uma aplicação do resultado.

Para alcançar essa proposta as etapas seguiram exatamente a sequencia descrita abaixo:

(a) Identificação manual de contextos definitórios em uma amostra significativa do corpus estudado, o que segundo a fórmula apresentada no item 4.2 da metodologia, de Cochran (1977), com 95% de grau de confiança, erro amostral considerado de 13, um máximo de 150 e o mínimo de 65 ocorrências em cada documento e uma população de 179 documentos, identificou-se que seria necessário a analise de 10 documentos. Estes documentos foram analisados em dois blocos de 5.

(b) Após a analise do primeiro grupo de documentos, as expressões linguísticas identificadas, junto com a gramática proposta por Sierra e Alarcón (2003) , Aguilar (2009) e Alarcón (2009) em língua espanhola traduzida para o português, além das expressões identificadas no trabalho de (KAMIKAWACHI, 2009) comporam o padrão sintático implementado na ferramenta, construída no âmbito desta pesquisa para identificação automática dos contextos definitórios. Essa ferramenta é um dos resultados deste trabalho.

(c) Uma primeira rodada de identificação automática dos contextos definitórios foi executada nos primeiros 5 documentos analisados manualmente.

(d) Realizou-se uma validação da gramática através de métodos comparativos entre os contextos definitórios extraídos de forma automática e aqueles identificados de forma manual.

(e) Após a análise dos resultados e ajustes realizados na ferramenta a fim de melhorar a precisão na identificação dos contextos definitórios, os outros 5 documentos passaram pela análise automática e seus contextos definitórios foram identificados.

(f) Nesse momento havia uma lista de contextos definitórios identificados de forma automática do segundo grupo de 5 documentos. Após essa varredura, realizou-se a análise manual destes mesmos documentos para, enfim, realizar novamente uma comparação entre os contextos definitórios identificados pelos dois métodos.

(g) Apenas após esses dois caminhos percorridos é que se concluiu a gramática de padrões definitórios, outro resultado da pesquisa.

(h) Após a construção da gramática e sua implementação na ferramenta, realizou-se a extração de forma automática dos contextos definitórios de todos os documentos do córpus estudado, mas um resultado do estudo.

Um fluxo com as etapas da pesquisa consta no anexo do estudo, figura 27, anexo 01. Com esse percurso metodológico foi possível alcançar os 3 objetivos específicos propostos e, consequentemente, o objetivo geral dessa pesquisa.

Revisão de Literatura e Fundamentos

5 Ciência da Informação

Neste capítulo foi feita uma revisão das características da Ciência da Informação, seus objetivos, métodos e teorias que a apoiam, principalmente no que se refere ao papel da definição como um dos elementos fundamentais para organização, recuperação e comunicação da informação registrada. O capítulo é organizado da seguinte maneira:

- Uma breve descrição do aspecto interdisciplinar da Ciência da Informação e sua ligação com a Documentação, Linguística, Terminologia e Ciência da Informação, quanto a mecanismos automáticos de organização, descoberta e recuperação da informação registrada em língua natural, são descritas nesta seção. A importância do termo e da definição já são comentados, além da definição de informação e conhecimento adotada na pesquisa.
- Em seguida, elementos da organização da informação e do conhecimento, além da recuperação da informação, são destacados, caracterizando as diferenças de abordagens na organização e suas consequências na recuperação. O estudo propõe auxiliar na reorganização de repositórios textuais.
- Finalizando, métodos e teorias para representação da informação e do conhecimento registrado são detalhados, tanto da Ciência da Informacão, como da Classificação Facetada e a Teoria do Conceito, quanto da Terminologia. O enfoque é dado na busca destas teorias em representar os conceitos tratados nos documentos através de termos, suas definições e relações. Identifica o importante papel das definições ou de parte delas, os enunciados definitórios ou enxertos definitórios, na elaboração dos instrumentos de organização da informação e do conhecimento, tais como, a taxonomia, a ontologia ou uma base terminológica de uma área de domínio.

5.1 A interdisciplinaridade

A Ciência da Informação (CI) está em constante processo de autoavaliação, seus conceitos são revistos e refeitos para abordar as demandas que surgem. Zins (2007b), após realizar uma consulta à especialistas da área, obteve 50 definições sobre a Ciência da Informação e as relacionou com seis distintas e possíveis concepções para a área. O mesmo autor propõe (ZINS, 2007c) um mapa do conhecimento em CI em um modelo baseado em sete fatores de mediação entre usuários e fontes de informação e (ZINS, 2007a) descreve 28 possíveis esquemas de classificação da Ciência da Informação com o apoio de renomados pesquisadores da área que responderam sua consulta.

Saracevic (1996a) descreve três características gerais que constituem a razão da existência e da evolução da Ciência da Informação: a CI é interdisciplinar por natureza; o imperativo tecnológico determina a CI ; Terceira, a CI é , juntamente com muitas outras disciplinas, uma participante ativa e deliberada na evolução da sociedade da informação. Essas três características ou razões constituem o modelo para compreensão do passado, presente e futuro da CI e dos problemas e questões que ela enfrenta.

Em artigo mais recente Saracevic (2009) cita três questões que também mapeiam e norteiam os estudos em Ciência da Informação:

1. a questão física: quais são as características e as leis do universo de informações registradas?
2. a questão social: como as pessoas podem relacionar-se, buscar e fazer uso da informação?
3. a questão de *design*: como é possível tornar mais rápido e eficaz os acessos aos registros da informação?

5.1.1 Linguística e Terminologia

Um dos primeiros estudos a relacionar a Linguística e a documentação no Brasil foi de Wanderley (1973). Em sua pesquisa, o autor descreve o processo de desconstrução do texto para referenciar o contéudo contido nos documentos. Descreve sobre o percurso de transformar os documentos com informações em Linguagem Natural em uma representação do documento em uma Linguagem Formal ou Linguagem Documentária (LD), a fim de facilitar a recuperação da informação. Este processo necessário para a análise de conteúdo (AC), seja para classificar ou

indexar os documentos, para o autor, interliga a Linguística à Documentação, consequentemente à Ciência da Informação.

Mendonça (2000) em estudo bibliométrico sobre a produção de trabalhos relacionados à Linguística e à Ciências da Informação, analisa 42 artigos da revista Ciência da Informação (Brasília), no período de 1972, quando do lançamento da revista, até o ano de 1998. Após analisar os artigos em grupos temáticos como a abordagem textual (teórico), Linguística e Bibliometria (quantitativo), a representação da informação, abordagem semântica, conceitual e terminológica (temático), o estudo da indexação automática e da linguagem natural (aplicativo), as relações curriculares (ensino), as tecnologias dos sistemas especialistas e a inteligência artificial (tecnológico) e a classificação decimal universal e a linguística (normativo), a autora destaca que uma das grandes problemáticas reveladas pela pesquisa foi a construção de conceitos e a representação da informação.

Muitas pesquisas com enfoques diferentes foram desenvolvidas desta época até os dias atuais, iremos destacar algumas como o trabalho de Campos (2001) que enfatiza a necessidade de metodologias apropriadas para elaboração de modelos conceituais que possam representar unidades do conhecimento na produção de hiperdocumentos ou hipertextos. A autora aborda a ontologia como instrumento de representação conceitual e analisa os elementos que a compõem e seus tipos.

Lara (2004) traz o estudo da relação entre os elementos da terminologia, o termo, a definição e suas teorias com os elementos da linguagem documentária e os instrumentos de organização da informação. Tálamo e Lenzi (2006) analisa a terminologia à luz da organização do conhecimento e do mapeamento conceitual necessário. Para as autoras, "para formar conceitos a partir das ocorrências dos termos é necessário comparar, refletir e abstrair"(TÁLAMO; LENZI, 2006, p. 05).

Kobashi (2007) descreve sobre elementos pragmáticos e semânticos na construção de instrumentos de representação de informação. Enfatiza a necessidade da analise da relação entre termos de um sintagma nominal que para o autor designam nomeando fenômenos e objetos de campos especializados.

O recente trabalho de Francelin (2010) também merece destaque, pois ao analisar o conceito e sua relação com a Organização da Informação e do Conhecimento trouxe grande contribuição para esta pesquisa.

5.1.2 Ciência da Computação na visão dos autores da CI

Para Saracevic (1996b) um marco histórico na origem da Ciência da Informação é o artigo de Vannevar Bush, respeitado cientista do Masschusetts Institute of Technology, MIT, e chefe do esforço científico americano durante a Segunda Guerra Mundial.

Nesse artigo, em meio a era da 'Avalanche' do conhecimento para (SOUSA, 2007), Bush fez duas coisas consideradas fundamentais para Saracevic (1996b) . Identificou o problema da explosão informacional, o crescimento exponencial da informação e de seus registros, particularmente em ciência e tecnologia e a dificuldade de acesso rápido à informação relevante e propôs usar as tecnologias de informação para combater o problema.

Segundo Saracevic (2009) Bush não foi o primeiro nem o único a falar sobre o assunto, mas motivou, por sua posição e *status*, muitos cientistas e profissionais de diversas áreas, alem dos governos, a tentar solucionar o problema. No final dos anos cinquenta a Ciência da Informação já estava com rumo certo, com equipes se formando e financiamentos acontecendo.

As origens da Ciência da Informação, entretanto, remontam ao ano de 1948, com o nascimento da primeira grande sociedade científica dos Estados Unidos, a American Society for Information Science (ASIS). Saracevic (2009) informa que Bush, além de ter escrito o intrigante artigo em 1945, também participou da criação de importante instituição para a Ciência da Informação, a National Science Foundation (NSF) (Fundação Nacional da Ciência)[1] nos Estados Unidos, em 1950, que a princípio tinha a finalidade de "Promover o intercâmbio de informações científicas entre os cientistas dos EUA e de outros países estrangeiros". Porém, em 1958, uma Lei Nacional de Defesa da Educação, ampliou o mandato da NSF, passando a ter como meta empreender programas para desenvolver métodos novos ou melhorados, incluindo sistemas mecanizados, para tornar a ciência da Informação disponível. Segundo Saracevic (2009), a evolução da Ciência da Informação, pelo menos nos Estados Unidos, foi enormemente influenciada pelo apoio do governo norte americano.

Na década de sessenta, surgem os primeiros conceitos e definições; ocorrem os debates sobre origens e fundamentos teóricos, a identificação dos marcos, o estabelecimento das relações interdisciplinares com outros campos do conhecimento e se vislumbra a atuação dos profissionais desta nova era. Também no início dessa mesma década, constata-se o registro oficial da Ciência da Informação, durante evento promovido pelo Georgia Institute of Technology (Estados Unidos), onde foi discutida a criação de novas tecnologias de informação, consequência natural do crescimento da produção científica. Apesar da ênfase na educação e em treinamentos

[1] NSF - tradução do autor

profissionalizantes, a realização de debates teóricos permitiu que se chegasse a uma primeira definição do que seria a Ciência da Informação.

Sousa (2007) cita uma definição, de 1962, do Geórgia Institute of Technology , para Ciência da Informação:

> Ciência que estuda as propriedades e o comportamento da informação, as forças que regem seu fluxo e os meios de processamento para acessibilidade e utilização ótimas. O processo inclui a origem, disseminação, coleta, organização, armazenamento, recuperação, interpretação e uso da informação (SOUSA, 2007, p. 02).

Há consenso entre os autores Saracevic (1996a) e Sousa (2007) ao admitir em que as origens da Ciência da Informação encontram-se na Biblioteconomia, em especial nas áreas de documentação e recuperação da informação, e que seu surgimento está intimamente ligado à revolução científica e técnica que se seguiu à II Grande Guerra, com destaque ao desenvolvimento das Tecnologias de Informação e Comunicação (TICs).

Atualmente estudos sobre a representação semântica, as tecnologias para recuperação da informação, o processamento de linguagem natural e a inteligência artificial são algumas das linhas de interseção entre a computação e a Ciência da Informação.

5.1.3 Ciência da Informação e seu objeto de estudo

Jaime Robredo, em sua obra *Da Ciência da Informação revisitada aos sistemas humanos de informação* , define Ciência da Informação de forma bem objetiva como "o estudo, com critérios, princípios e métodos científicos, da informação" (ROBREDO, 2003, p.105).

O objeto de estudo da Ciência da Informação, portanto, é a informação, porém o termo "informação" tem sido utilizado para as mais diversas situações.

Zeleny (1987) propôs a hierarquia dado, informação, conhecimento e sabedoria, a cadeia do conhecimento, também denominada DIKW Chain por Hey (2004), para tentar distinguir os contextos informacionais.

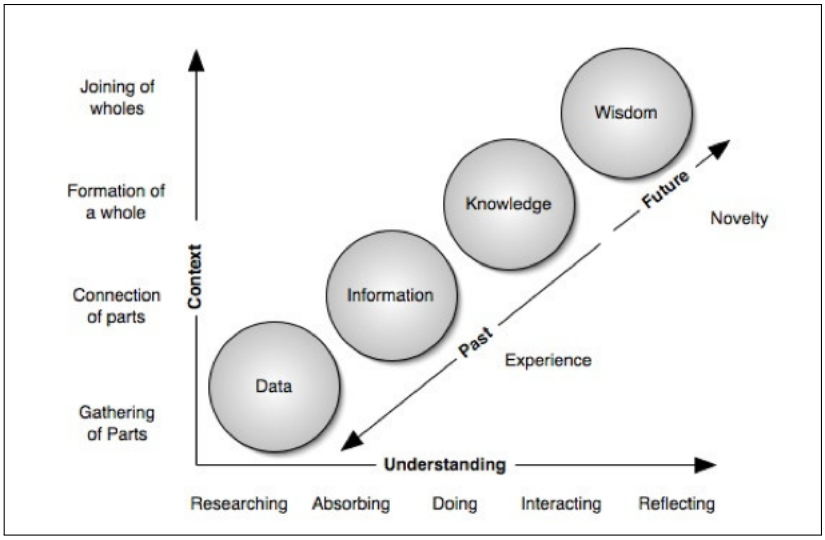

Figura 4: Cadeia Dado, Informação, Conhecimento e Sabedoria - DIKW
Fonte: (CLARK,)

Esta cadeia é bastante utilizada na Ciência da Informação e (ZELENY, 1987) define os seus elementos da seguinte forma:

1. dado – signo(s) sem significado contextual, informação não processada;
2. informação – dado(s) com significado contextual;
3. conhecimento – informação coordenada e aplicada por um sujeito;
4. sabedoria – reflexões sobre o conhecimento.

Hey (2004) ao comentar sobre a cadeia faz uma relação do dado com um líquido puro que precisa ser destilado para virar informação, após isso precisa de um novo processo de destilação para virar conhecimento, e quando o elemento está bem pastoso, ele vira sabedoria. Propõe a imagem de uma pirâmide, figura 5, para demonstrar a cadeia, por acreditar que é preciso ter muitos dados para obter informação, como no processo de mineração de dados. É necessário, também, bastante informação sobre um domínio para se ter conhecimento e finalmente, a sabedoria que se obtém após a reflexão sobre muitos conhecimentos adquiridos.

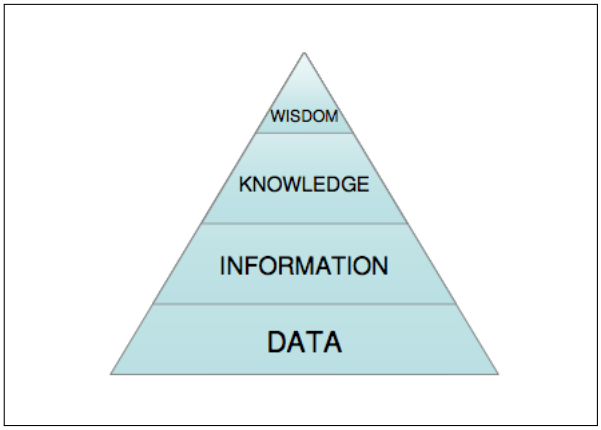

Figura 5: Hierarquia DIKW
Fonte: (HEY, 2004)

Siqueira (2008) ao analisar a tecnologia e a natureza da informação refuta essa hierarquia e as definições de Zeleny afirmando que o dado é que seria o fundamento principal da Ciência da Informação neste modelo e não a informação. Propõe então uma nova hierarquia:

1. a informação sintática – aquela que não possui significado contextual. É um signo sintático cuja forma é objeto de observação;
2. a informação semântica – aquela que possui significado contextual para um sujeito;
3. a informação pragmática – aquela que está codificada e preparada para uso. É uma informação manipulada por um sujeito com fins de utilidade planejada.

O autor considera que esta hierarquia toma a informação como fundamento de organização das coisas e descreve sobre a informação:

> A sua expressão sintática é o dado, a sua expressão semântica é a relação significativa para um sujeito e sua expressão pragmática é a codificação em um suporte para uso, é o caso do conhecimento, codificado na mente pelas relações neurais, ou em um livro, pela codificação de letras; ou em um software,

> pela estruturação de máquinas de estados regidas pela lógica booleana dos bits (SIQUEIRA, 2008, p. 94).

Este trabalho adota a visão de Siqueira (2008) sobre a informação e sua relação com o dado e o conhecimento.

5.2 Organização da Informação e do Conhecimento

A organização parece ser uma tendência fundamental nos seres humanos. Todos nós utilizamos elementos de organização no processo de aprendizagem desde crianças. Taylor A. G.; Joudrey (2009), afirmam que os humanos desenvolveram as mais sofisticadas habilidades cognitivas para categorizar, reconhecer padrões, ordenar, relacionar e criar grupos de pensamentos e ideias. Os autores identificam como ponto fundamental para o desenvolvimento da Organanização da Informação o sonho de Paul Otlet e Henri La Fontaine em organizar toda a produção científica mundial até 1892, através do Universal Bibliographic Control (UBC), dando origem, segundo eles, a todos as outras ferramentas. Otlet e La Fontaine fundaram em 1895, o instituto Internacional de Bibliografia (IBB) com sede na Bélgica.

Sousa (2007) cita que este era o momento do 'Caos' Documentário, do volume crescente de documentos e da diversidade de tipos de documentos, afirmando que a própria mudança de nome desta entidade no tempo, de IBB para Federação Internacional de Documentação (FID) em 1937 e, posteriormente, em 1988, para Federação Internacional de Informação e Documentação (mantida a sigla FID), reflete a mudança dos conceitos do campo de atuação da entidade, sendo a passagem da era da Bibliografia para a era da Documentação e posteriormente para a era da Informação.

Sousa (2007) relata ainda que nesta fase como na fase de explosão da informação, as atenções se voltaram para os esquemas de classificação bibliográfica visando encontrar uma melhor ordenação lógica para as coleções e para a organização temática de itens de biblioteca. Nomeia, por data de criação a Classificação Decimal de Dewey (CDD), e a Classificação Decimal Universal (CDU), a Classificação da Biblioteca do Congresso Americano (LC Library of Congress); a Classificação de Assunto de Brown (Subject Classification), a Classificação Bibliográfica de Bliss (Bibliographic Classification), e a Classificação Facetada de Ranganathan (Colon Classification), com destaque para essa última que segundo a autora a obra Ranganathan em Prolegomena to Library Classification (1967) é um marco teórico e clássico de referência sobre classificação em biblioteca. A teoria de Classificação Facetada será detalhada no item 5.3.1.

5.2.1 Recuperação da Informação (RI)

Taylor A. G.; Joudrey (2009, p.427) afirmam que “nós organizamos porque precisamos recuperar”[2]. Saracevic (2009), porém, ao descrever sobre o surgimento e a evolução da recuperação da informação, enfatiza a diferença entre a RI e os métodos e sistemas relacionados que por muito tempo a precederam, tais como classificações, título por assuntos, métodos de indexação, ou descrições bibliográficas. Considera, também, que a noção fundamental utilizada na descrição bibliográfica e em todos os tipos de classificação ou categorização é a “tematicidade” (*aboutness*), concentram-se em descrever e categorizar os objetos de informação e, em contrapartida, a noção fundamental usada em RI é a relevância (*relevance*).

A recuperação não é sobre qualquer tipo de informação, até porque existem muitas, e sim sobre a informação que é relevante ou que satisfaça o usuário, para Sousa (2007).

Para Saracevic (2009), ao escolher a relevância como uma noção básica e fundamental na RI, os sistemas de informação, serviços e atividades relacionados com ela, além de todo o campo da Ciência da Informação, foi em uma direção que difere das abordagens adotadas na biblioteconomia, documentação e serviços de informação relacionados.

Segundo Figueiredo (1977, p.75), foi com a publicação de *Sources of information on specific subjects* por S.C. Bradfordd em 1934, na qual o autor afirmou que "há periódicos de âmbito, obviamente e a priori, relevantes a assuntos investigados", e, pela primeira vez, o conceito de relevância apareceu dentro do contexto de medida de uma fonte de informação.

Desde então, o conceito de relevância vem sendo analisado, mas foi de Cuadra et al. (apud SARACEVIC, 1975) uma das primeiras definições discutidas na literatura. Segundo os autores, a relevância é uma relação entre uma declaração de informação requisitada por uma consulta e algum documento contido na base. Nesse caso, relevância seria uma propriedade do sistema e, portanto, depende apenas de como este adquire, representa, organiza e associa as informações (SARACEVIC, 1996b).

Para Shamber L.; Eisenberg (1990), a relevância é um julgamento de qualidade entre a informação recuperada e a necessidade de informação do usuário. Nessa definição, o usuário é que determina se o documento recuperado atende as suas necessidades. Relevância aqui se refere aos contextos subjetivos que são empregados pelo usuário para julgar os objetos informacionais. Aspectos cognitivos, situacionais e psicológicos dos usuários são fundamentais para a compreensão do julgamento.

[2]Tradução nossa

Os dois conceitos acima definem duas categorias de relevância citadas na literatura e Shamber (1994) usa uma nomenclatura para referenciar esses dois tipos. Chama a primeira de orientada ao sistema e a segunda de orientada ao usuário.

Saracevic (2007), ao falar da relevância, faz a seguinte afirmativa:

> Relevância é como uma árvore de conhecimento. A estrutura básica do sistema de relevância na ciência da informação é uma dualidade: a relevância de tópicos e a relevância do usuário. Cada uma tem seus galhos devem ser bem exploradas, mas fazem parte da mesma árvore (SARACEVIC, 2007, p. 1931).

Capurro R.; Hjorland (2007, p.31), falam sobre o conceito de Informação para a Recuperação da Informação (RI) e observam que RI, normalmente, significa recuperação de documentos, considerando que os sistemas efetivamente recuperam documentos e não a informação contida neles. Afirma, também, que os termos "recuperação de documentos" e "recuperação de textos" são frequentemente usados como sinônimos na Ciência da Informação.

Frei (1996) expressa visão similar, porém com relação à recuperação de referências:

> Pesquisadores acadêmicos têm estudado, por anos, como indexar, armazenar e recuperar referências bibliográficas, denominando esta disciplina de 'recuperação de informação' e não de 'recuperação de referências'. Assim, desde longo tempo, RI tem se ocupado em localizar um tipo de informação bastante restrito e o termo recuperação de informação é, na verdade, equivocado. Recuperar referências bibliográficas relevantes é, certamente, um problema válido e útil para algumas pessoas. Mas ele claramente não reflete a maioria dos problemas que devem ser enfrentados com a explosão informacional contemporânea (FREI, 1996, p. 03).

Saracevic (2009, p.03) cita Calvin N. Mooers (1919-1994), físico e matemático, criador da RI, que definiu recuperação da informação como "O processo de procura ou descoberta com relação às informações armazenadas ... útil a [um usuário]." Podemos destacar o termo descoberta contido na definição de Mooers e a utilidade da informação para o usuário como elemento primordial.

Yates e Neto (1999, p.01), caracterizam Sistemas de Recuperação da Informação (SRI) como "sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação". Essa definição mais recente já incorpora os processos de representação e organização da informação.

Campos (2001) enfatiza que a questão primordial, posta quanto ao tratamento e à recuperação de informação, diz respeito à qualidade no tratamento das informações e à adequação a uma solicitação de busca dada por um usuário através de um controle terminológico que venha garantir precisão nas informações recuperadas em meio eletrônico.

Araújo Jr. (2007) em sua obra, *Precisão no processo de busca e recuperação da informação*, descreve a necessidade de se observar a qualidade do processo de medição da precisão para garantir um resultado satisfatório em mecanismos computacionais. Relembra que as medidas mais utilizadas para avaliar os sistemas de recuperação da informação são a Precisão e a Revocação, que segundo o autor, foram usadas pela primeira vez por Cleverdon em seu estudo em 1962 e definidas a partir da seguinte fórmula:

$$R = \frac{a}{a+b}$$

$$P = \frac{a}{a+c}$$

Onde:

R = revocação;

P = precisão;

a = refências úteis e recuperadas;

b = refêrencias úteis não recuperadas; e

c = refêrencias inúteis e recuperadas;

A Revocação seria, portanto, um índice para medir a cobertura da aplicação, o quanto o sistema automaticamente chega próximo de uma varredura manual e a Precisão é uma medida que avalia a quantidade de erros necessários para alcançar os objetos requisitados a um sistema. Em testes de aplicações se assinala os itens relevantes ou úteis anteriormente e executa a aplicação para medir seus resultados. Estas medidas foram utilizadas na análise de resultados gerados a partir da ferramenta de extração automática de Contextos Definitórios realizada neste estudo.

Gomes e Campos (2004), afirma que para garantir esta precisão verifica-se a necessidade de ferramentas taxonômicas e terminológicas para o tratamento semântico de informações contidas em bases de dados. Citam a ontologia como um ferramenta com essa finalidade e definem ontologia da seguinte maneira:

> Ontolgia é um conjunto de conceitos padronizados onde termos e definições devem ser aceitos por uma comunidade no âmbito de um domínio e tem por finalidade permitir que múltiplos agentes compartilhem conhecimento. Uma ontologia consiste em termos, definições, e axiomas relativos a eles (GOMES; CAMPOS, 2004, p. 02).

Recentemente o conceito de ontologia como instrumento de organização conceitual surgiu

e os estudos sobre o tema cresceram, podemos citar o trabalho de Guarino (1998). Porém alguns autores identificaram a necessidade de compreender as diferentes formas de se organizar para obter um melhor resultado na recuperação.

5.2.2 Tipos de organização

O termo "Organização do Conhecimento" (OC) tem sido utilizado na área de Ciência da Informação por alguns autores e é completamente renegado por outros que utilizam "Organização da Informação" (OI).

Taylor A. G.; Joudrey (2009) utilizam apenas OI, pois consideram que o conhecimento existe na mente do indivíduo que estudou um assunto. Sempre que ele registra esse conhecimento ele se torna informação. Ou seja, um livro não contém conhecimento, contém uma representação do conhecimento do autor, que pode ser tão imperfeita quanto a sua dificuldade em explicar os conceitos como os compreende.

Sousa (2007), usa o termo "organização do conhecimento" e coloca o conhecimento como uma informação contextualizada; descreve a CI na fase da necessidade de conhecimento e alerta que são necessários métodos de processamento da informação diferenciados para obter esse objetivo.

Bräscher e Café; (2010, p.91) aprofundam a discussão e propõem o uso dos dois termos, porém para processos diferentes. Para as autoras, OI, "é um processo que envolve a descrição física e de conteúdo dos objetos informacionais", estando, portanto, no mundo dos objetos físicos. A OC, por sua vez, "visa à construção de modelos de mundo" e se constitui numa estrutura conceitual. Complementam acerca da Organização do Conhecimento:

> Organização do Conhecimento tem por base a análise do conceito e de suas características para o estabelecimento da posição que cada conceito ocupa num determinado domínio, bem como das suas relações com os demais conceitos que compõem esse sistema nocional (BRÄSCHER; CAFÉ;, 2010, p. 93).

Alvarenga (2003) ao falar sobre a descrição de conteúdos (classificação, indexação e elaboração de resumo), afirma que não são os documentos que são classificados, mas os conceitos contidos nos documentos. Bräscher e Café; (2010) concordam, entretanto não identificam esses processos de representação conceitual como parte da OC. As autoras consideram que se refere a um objeto informacional em particular e a visão de apenas um autor, enquanto a representação do conhecimento é fruto de uma análise de domínio e procura refletir uma visão consensual sobre um modelo de abstração do mundo real, construído para determinada realidade.

Uma pergunta que se faz necessária é: o que é um conceito? Francelin (2010) afirma que os elementos, as características e as linhas de força teóricas que estão nas bases das respostas para esta pergunta refletem escolhas teórica-epistemológicas e permitem identificar os princípios adotados por autores da área da Organização da informação e do Conhecimento.

O conceito era definido pela norma ISO 704 (1987) como 'unidade do pensamento' (*units of thought*), porém durante a reunião da ISO-TC 37, em que se aprovaria a nova versão da Norma, foi possível aprovar a definição de conceito como 'unidade do conhecimento' (*unit of knowledge*) (GOMES; CAMPOS, 2004).

Essa mudança ocorreu pelo trabalho de Dahlberg (1978b) com a Teoria do Conceito que contrariando a ideia de Wuster, criador da terminologia e que teve grande influência na criação das normas ISO, segundo Campos (2001, p.101), afirma que o termo "pensamento"pode ser subjetivo e impreciso propondo definir o conceito como "unidade de conhecimento". Francelin (2010) considera que essa questão é discutível, porém para a autora o conceito definido como unidade de conhecimento caminha para algo já externalizado, não restrito à mente daquele que pensa, pois, para se ter uma "totalidade de proposições verdadeira sobre o mundo", como ocorre na ciência (DAHLBERG, 1978b, p. 6) é necessário que tais proposições sejam expressas e comunicadas pela linguagem.

5.3 Representação da Informação e do Conhecimento

A seguir detalharemos três teorias importantes para a representação da informação e do conhecimento e fundamentais em processos automatizados para auxilio a organização da informação e do conhecimento. A Teoria da Classificação Facetada de Ranganathan, a Teoria da Terminologia de Wüster e Cabré e a Teoria do Conceito de Dahlberg.

5.3.1 Teoria da Classificação Facetada

Desenvolvida pelo indiano Shiyali Ramamrita Ranganathan na década de 1930 , a Teoria da Classificação Facetada tem sido amplamente discutida e apontada como uma solução para a organização do conhecimento, pois seus princípios têm como finalidade acompanhar as mudanças e a evolução do conhecimento. (CAMPOS, 2001; ALARCON, 2004; LARA, 2004)

Para Ranganathan e Gopinath (1967, p. 66), o conhecimento é "a totalidade das ideias conservadas pelo ser humano", através da observação dos fatos, coisas e processos e , segundo a autora, os esquemas de classificação bibliográfica teriam como função, além de permitir a

organização dos documentos nas estantes, a representação do conhecimento registrado numa dada área de assunto.

Prescott (2003) ao falar sobre a teoria de Ranganathan observa que:

> A expressão análise em facetas foi adotada por Ranganathan para indicar a técnica de fragmentar um assunto complexo em seus mais diversos aspectos ou partes constituintes, que são as facetas, utilizando, para estabelecer a relação entre as "categorias fundamentais", de noções abstratas, denominadas Personalidade, Matéria, Energia, Espaço, Tempo, conhecidas pela sigla PMEST. Personalidade é a característica que distingue o assunto; Matéria é o material físico do qual um assunto pode ser composto; Energia é uma ação que ocorre com respeito ao assunto; Espaço é o componente geográfico da localização de um assunto; Tempo é o período associado com um assunto (PRESCOTT, 2003, p. 01).

A respeito do total de cinco categorias denominadas fundamentais para representar o universo de assuntos em classes bastante abrangentes, Ranganathan apresenta o seguinte argumento:

> Alguém pode perguntar: Por que as idéias fundamentais postuladas são em número de cinco? Por que não três? Por que não seis? Isto é possível. Há liberdade absoluta para todos tentarem. Uma pessoa pode talvez gostar de seis. Ela deve classificar nessa base alguns milhares de artigos variados. Se elas produzirem resultados satisfatórios arranjando os assuntos dos artigos ao longo de uma linha, aquele postulado pode ser aceito. Isto não é uma matéria a ser discutida ex cathedra sem um teste completo e prolongado. Trabalhar com base em cinco idéias fundamentais produziu resultados satisfatórios nos vinte últimos anos' (RANGANATHAN; GOPINATH, 1967, p.70)[3].

Tálamo e Lenzi (2006), afirmam que esta estruturação em categorias na organização de conceitos e, em consequência, na elaboração de uma classificação, permite o entendimento da natureza do conceito, além de ser um recurso para a formação das estruturas conceituais. Ou seja, as categorias permitem a sistematização do conhecimento.

Campos (2001) cita os elementos que constituem a teoria de classificação facetada e acredita que para melhor compreensão das idéias de Ranganathan essa sequencia é a mais didática:

1. Unidades Classificatórias - Essas unidades representam os conceitos e suas relações e na Teoria da Classificação Facetada elas são o assunto básico, áreas mais abrangentes do conhecimento e ideia isolada que sozinha não é um assunto, mas combinadas podem gerar um assunto. Por exemplo, Milho denota uma ideia isolada, mas se combinada com o assunto básico Agricultura forma o assunto Cultivo de Milho. A autora considera que

[3]Tradução nossa

a ideia isolada pode ser considerada um conceito, porém em alguns casos funciona como unidade combinatória que tem por função facilitar a formação da notação, sendo, neste tipo de tabela, a notação o representante do conceito. Afirma que com isto é possível representar conceitos que não estão nomeados na língua, como por exemplo, Psicologia + Pré-adolescente.

2. Características - São usadas para comparar os elementos que estão sendo classificados, objetiva formar classes e, dentro destas, os renques e cadeias.

3. Renques e Cadeias - servem para diferenciar na formação das classes, as séries verticais e horizontais de conceitos. Renques formam séries horizontais, pois são formadas a partir de uma única característica. Por exemplo: Macieira e Parreira são elementos da Classe Árvore Frutífera, formada pela característica da divisão - tipo de árvores frutíferas. Cadeia são séries verticais de conceitos. Árvore - Árvore Frutífera - Macieira. Os renques e cadeias revelam a organização da estrutura hierárquica desta classificação, evidenciando as relações de gênero-espécie e de todo/parte.

4. Facetas - é "um termo genérico usado para denotar algum componente - pode ser uma assunto básico ou um isolado - de um assunto composto, tendo, ainda, a função de formar renques, termos e números."(RANGANATHAN; GOPINATH, 1967, p.88)[4]

5. Categorias fundamentais - Postulado das cinco categorias que fazem o primeiro corte classificatório do domínio e garantem a visão de conjunto dos agrupamentos que ocorrem na estrutura. Personalidade, Matéria, Energia, Espaço, Tempo, conhecidas pela sigla PMEST.

6. Universo do Conhecimento - "é a soma total, num dado momento, do conhecimento acumulado. Ele está sempre em desenvolvimento contínuo. Diferentes domínios do Universo do Conhecimento são desenvolvidos por diferentes métodos. O método Científico é um dos métodos reconhecidos de desenvolvimento. O método Científico é caracterizado pelo movimento sem fim em espiral (RANGANATHAN; GOPINATH, 1967, p.94).

A autora destaca, ainda, dois pontos importantes na Teoria da Classificação Facetada. O primeiro é o enfoque no documento como registro de conhecimento, sendo as unidades que o constituem não mais os assuntos, mas os conceitos, que Ranganathan, segundo a autora, denomina de isolados ou ideia isolada. O segundo ponto é a série de princípios que visam permitir que os conceitos possam ser estruturados de forma sistêmica, isto é, os conceitos se

[4]Todas as citações de Ranganathan são traduções nossa

organizam em renques e cadeias, estas estruturas em facetas e estas em uma dada categoria fundamental (CAMPOS, 2001).

5.3.2 Teoria da Terminologia

Terminologia é um termo que aparece com vários significados na literatura e para Cabré (1995) nos remete a pelo menos três aspectos: a disciplina, a prática e o produto gerado desta prática. Como disciplina é a matéria que se ocupa dos termos especializados; como prática é um conjunto de princípios para a organização dos termos; como produto é o conjunto de termos de uma determinada especialidade.

O engenheiro austríaco E. Wüster (1898-1977) é considerado o pai da terminologia, quando em sua tese de doutorado, intitulada de A normalização internacional da terminologia, expôs pela primeira vez de forma sistematizada uma teoria terminológica. Sua teoria, mas tarde consolidada em suas aulas na Universidade de Viena, foi denominada Teoria Geral da Terminologia (TGT) e foi o marco para a consolidação da terminologia como disciplina.

Para Cabré (2003), Wüster buscou uma série de objetivos com a TGT:

- Eliminar a ambiquidade na linguagem especializadas através da padronização terminológicas, a fim de torná-las ferramentas eficientes de comunicacao.
- Convencer os usuários de linguagens técnicas dos benefícios de padronização terminológicas.
- Estabelecer a terminologica como uma disciplina para todos os efeitos práticos e para dar-lhe o *status* de uma ciência.

Segundo Kamikawachi (2009) a TGT tem como primazia o conceito e apresenta como proposta a compilação de conceitos e termos para normalização sem considerar a polissemia e as ambiguidades. Desta forma, é necessário garantir a unificação de conceitos e termos através da correspondência exata para facilitar a comunicação nos vários domínios da Ciência e da Tecnologia. (CAMPOS, 2001). Essa característica normativa fica clara quando a teoria de Wuester se ajusta aos objetivos da normalização técnica e esta na base do Comitê 37 da ISO - Fundamentos da Terminologia.

Maciel (2001) recorda que a TGT foi concebida a princípio para as chamadas áreas aplicadas das Ciências duras, como Engenharia, Eletrotécnica e Mecânica, onde se pressupõem sistemas de conceitos delineados com precisão e práticas bem determinadas O autor considera

que a teoria de Wuster propõe uma metodologia que segue a direção onomasiológica, isto é, começa pela identificação dos conceitos básicos da área em foco a fim de chegar ao sistema conceitual desta mesma área.

A TGT recomenda ainda que os conceitos devem ser identificados, nomeados através de um termo e definidos por autoridades competentes, reunidas em comitês oficiais de normalização linguística, visto que a comunicação profissional não pode ficar sujeita a variações e flutuações que a língua natural sofre. Esta atribuição do significado do conceito em termos e definições não pode ficar preso à memória dos especialistas, mas deve de forma concreta ser armazenada nos produtos terminológicos. (MACIEL, 2001)

Campos (2001), considera que para a TGT o conceito é uma unidade de pensamento, constituído de características que refletem as propriedades significativas atribuídas a um objeto, ou a uma classe de objetos. Para a autora, a característica que constitui um conceito é também um conceito e através dela é possível comparar conceitos, classifica-los em um sistema de conceitos, sintetizá-los através da definição e denominá-los através do termo.

Maciel (2001) afirma porém, que a abordagem clássica não se aprofunda em investigações sobre a gênese do conceito, se apoia na filosofia do positivismo lógico e apresenta o conceito como um construto mental, elaborado a partir da síntese das características de fenômenos do mundo real ou imaginário. O conceito é, então, identificado por um símbolo, o signo linguístico, e para o autor sua descrição por meio da lingua é através da definição.

> O principal papel da definição, portanto, é fixar a referência do termo ao conceito e estipular os traços que o caracterizam. Tais traços servirão como elos de seu relacionamento com os outros conceitos dentro da estrutura hierarquizada de conhecimento de uma área temática (MACIEL, 2001, p. 42).

As últimas décadas a TGT passou a ser bastante questionada por conceituados autores da terminologia, entre eles, Cabré (1995). Apesar de reconhecer os méritos da visão terminológica divulgada pela Escola de Viena, a autora considera, a teoria clássica insuficiente para atender as necessidades atuais da comunicação da ciência.

Nesse contexto outras abordagens teóricas surgiram com destaque para a Teoria Comunicativa da Terminologia(TCT), apresentada por Maria Teresa Cabré, lider do grupo IULATERM, do Instituto de Linguística Aplicada da Universidade Pompeu Fabra, de Barcelona.

Cabré (2003) propõe a TCT como uma teoria que possa ser aplicada a todas as áreas do conhecimento, que considera as unidades terminológicas como unidades de conhecimento, significação, denominação e comunicação no quadro do discurso especializado real.

Para a TCT a unidade terminológica é o centro do objeto de conhecimento da terminologia e

deve ser visto como um poliedro com três pontos de vista: o cognitivo (O conceito), a linguística (o termo) e comunicativo (a situação). Cada uma das três dimensões, sendo inseparáveis, na unidade terminológica, são portas de acesso direto ao objeto (CABRÉ, 2003).

Oliveira (2009) considera que a principal ruptura da TCT com a TGT é que a primeira reconhece a polissemia dos termos no espaço das comunicações especializadas. Entretanto, adverte, que esta visão polissêmica não induz o desinteresse pela relação termo-conceito, pelo contrário, o componente conceitual é importante na medida em que representa uma determinada indentificação dos termos, tendo em vista sua intima ligação com a definição terminológica.

A Teoria Comunicativa da Terminologia (TCT) tem sido adotada em vários grupos de pesquisa no Brasil nos trabalhos terminográficos em áreas do conhecimento. Almeida, Aluísio e Oliveira (2007) propõem uma sequência de etapas que, segundo a autora, deve fazer parte de qualquer trabalho terminográfico que segue a TCT. São elas:

1. Coleta (ou extração) de termos - obtenção do conjunto terminológico que comporá as unidades léxicas que serão inseridas na ontologia e dicionários ;

2. Elaboração do mapa conceitual ou criação de uma ontologia - Semelhante a uma árvore de domínio, só que os conceitos/termos estão ali armazenados;

3. Inserção dos termos na ontologia e sua validação por especialistas - a partir de campos nocionais, pede-se para os especialistas assinalarem os termos semanticamente relevantes ;

4. Elaboração e preenchimento das fichas terminológicas - a ficha é um dossiê do termo, fundamental em uma pesquisa terminológica. Não tem um modelo ideal, cada projeto têm suas necessidades;

5. Elaboração e incremento da base definicional - "Tem como função armazenar todos os excertos definitórios aos termos, de forma a facilitar a redação da definição"(ALMEIDA; ALUÍSIO; OLIVEIRA, 2007, p.04);

6. Elaboração das definições e informações enciclopédicas - Considerada a etapa mais complexa e importante numa pesquisa, visto que um bom dicionário especializado se avalia pela qualidade das suas definições;

7. Edição dos verbetes - É uma seleção de alguns campos da ficha para constarem do modelo de verbete final.

Destacamos as considerações de (ALMEIDA; ALUÍSIO; OLIVEIRA, 2007) quanto a armazenar os excertos definitórios na base definicional:

> É imprescindível armazenar essas informações, uma vez que:1) somente com o preenchimento de um número suficiente de excertos definitórios é que a redação de uma definição pode ser iniciada; 2) a quantidade e qualidade de excertos devem ser suficientes para elucidar o redator das definições, uma vez que este não é um especialista da área-projeto; 3) as definições, depois de elaboradas, são submetidas à apreciação dos especialistas, caso eles encontrem algum problema conceitual, questionem as fontes bibliográficas ou peçam que o trabalho seja refeito, é possível um retorno a essas informações constantes na base definicional, não sendo necessária uma volta aos textos originais (ALMEIDA; ALUÍSIO; OLIVEIRA, 2007, p. 03-04) .

Oliveira (2009) desenvolveu, em sua tese de doutorado, uma ferramenta automatizada que se propõe auxilar o terminólogo em todas estas etapas do trabalho terminológico (e-termos)[5]. Considera, porém, que antes da fase de coleta ou extração dos termos é necessário a compilação de córpus de especialidade. Seu trabalho faz um mapeamento de várias ferramentas computacionais que auxiliam o trabalho terminológico e o tratamento automático de textos como extração de termos e criação automática de ontologias. Adverte, entretanto, que existem muitos trabalhos isolados e nem sempre com um resultado adequado.

5.3.3 Teoria do Conceito

Ingetraut Dahlberg proferiu duas palestras na Conferência Brasileira de Classificação Bibliográfica em 1972, "Teoria da classificação, ontem e hoje"(DAHLBERG, 1972b) e "O futuro das linguagens de indexação"(DAHLBERG, 1972a), sendo estes trabalhos publicados apenas em 1979, em Brasília por, meio do Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT. Nestas palestras já se perguntava como resolver os problemas do processamento informacional em bases de dados (FRANCELIN, 2010).

A solução proposta por Dahlberg (1978b) foi sistematizar a compreensão da natureza dos conceitos através de uma teoria, chamada Teoria do Conceito. Nesta teoria “Cada enunciado verdadeiro representa um elemento do conceito” (DAHLBERG, 1978b, p.02) e a soma total dos enunciados verdadeiros de um objeto fornece o conceito do mesmo. Estes enunciados verdadeiros podem conter conceitos individuais, como aqueles que identificam um objeto especifico no tempo e no espaço e em conceitos gerais, que identificam categorias ou grupos de objetos. Exemplo:

[5] e-termos: para mais informações visite o site http://www.etermos.cnptia.embrapa.br/

Tabela 1: Exemplo de Conceitos Individuais e Gerais

Conceitos Individuais	Conceitos Gerais
A UnB	As universidades
A vitória magnifica do Flamengo na partida de futebol contra o Fluminense no dia 15 de janeiro de 1976	As partidas de futebol
O descobrimento do Brasil no ano de 1500	As descobertas marítimas

Fonte: Dahlberg (1978b) adptado

Tanto os conceitos individuais quanto os conceitos gerais podem ser enunciados em linguagem natural. Exemplo:

Tabela 2: Exemplo de Enunciados Individuais e Gerais

Conceito	Enunciado
Individual: IBICT (Instituto Brasileiro de informação em Ciência e Tecnologia	- é uma instituição - situada no Rio de Janeiro - relacionada com a coordenação dos sistemas de informação no Brasil - possui cerca de 60 funcionários, etc.
Geral: Instituição	- é constituída por um grupo de pessoas - que trabalham com determinada finalidade - possuindo adminstração comum - localizada em determinado lugar - durante determinado tempo, etc.

Fonte: Dahlberg (1978b)

Para Francelin (2010), a formulação de enunciados verdadeiros com os atributos dos conceitos permite que se identifiquem características que tanto serão específicas de um único conceito, como também serão compartilhadas por outros conceitos. Esta identificação ocorre por meio da análise de conceitos. Sendo possível, mesmo sem um nome que designe um conceito, saber o que ele é e formulá-lo pelo conjunto de suas características.

Dahlberg (1978b) define conceito como "unidade do conhecimento"o que difere da primeira versão da norma ISO 704 que utilizava "unidade do pensamento"como proposto por Wuster. Unidade do conhecimento, para Campos (2001), é mais apropriado, pois pressupõe um entendimento mais objetivo de algo observável e apresenta o que chama de "Modelo para Construção de Conceitos".

Dahlberg (1978a) considera três passos envolvidos na formação do conceito: 1) o passo re-

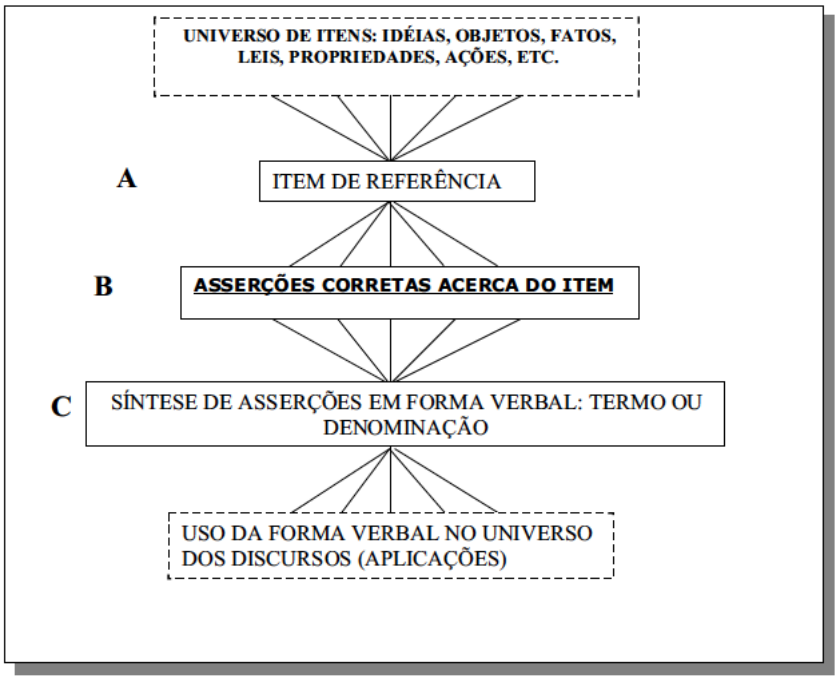

Figura 6: Modelo para construção de conceitos de Dalhberg
Fonte: (CAMPOS, 2001)

ferencial, 2) o passo predicacional e c) o passo representacional. Estes podem ser representados graficamente na forma de um triângulo.

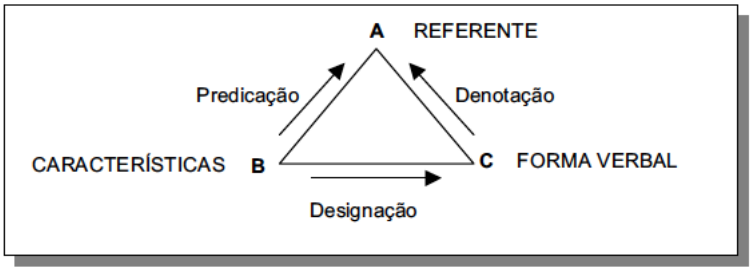

Figura 7: Triângulo de Dalhberg
Fonte: (CAMPOS, 2001)

Cada afirmação correta sobre o referente é um elemento do conhecimento sobre ele e o total de afirmações sobre o referente forma a unidade de conhecimento, ou seja, o conceito.

Campos (2001) considera que a Teoria do Conceito possibilitou um método para a fixação do conteúdo do conceito e para seu posicionamento em um Sistema de Conceitos, sendo o conceito não apenas um elemento de significação do termo e sim o termo como um elemento do próprio conceito.

"A equivalência entre o termo (*definiendum*) e as características necessárias de um referente de um conceito (o *definies*) com o propósito de delimitar o uso do termo em um discurso"(DAHLBERG, 1978b, p.178), resulta na definição deste conceito dentro de um sistema. Para Campos (2001), assim, a definição não é mais colocada em segundo plano, como um recurso auxiliar para minimizar dúvidas sobre o uso do termo e serve, por sua vez, como um recurso para estabelecer as fronteiras da intensão do conceito, da fixação do conceito e seu posicionamento no próprio Sistema de Conceitos.

A definição, para Dalhberg, trata de determinar ou fixar os limites de um conceito ou idéia. Propõe então uma definição da definição :

> Podemos então definir a definição da seguinte maneira: Definição — df delimitação ou fixação do conteúdo de um conceito (intensão, ou conjunto de características ou atributos) (DAHLBERG, 1978b, p.02).

Nesse capítulo foram apresentados o conceito de informação e conhecimento para a pesquisa, tipos de organização da informação e do conhecimento, teorias de análise de conteúdo para representação da informação e do conhecimento e o papel das definições nessas teorias. Demostra a importância das definições ou dos enunciados definitórios para o mapeamento semântico de uma área de domínio. No próximo capítulo trataremos em detalhe das definições contidas em textos.

6 Contextos Definitórios

Este capítulo traz o conceito do objeto de estudo desta pesquisa o Contexto Definitório. Para isso ele se estrutura assim:

- Em primeiro lugar, se identifica os diversos conceitos e tipos de definições, caracterizando a definição como elemento de ligação entre os objetos e os conceitos. Descreve também sobre as relações semânticas dos conceitos apresentados na expressão das definições, citando o trabalho de autores que pesquisam sobre estes temas.
- Na sequência, a identificação de definições em textos são abordadas, citando estudos feitos em inglês, espanhol e francês. Várias nomenclaturas são utilizadas para identificar parte do texto que caracterizam uma definição ou parte dela, tais como, enunciado definitório, contextos ricos em conhecimento e a adotada neste estudo, os Contextos Definitórios, que são conceituados e têm seus elementos abordados nesta seção.
- Ao final do capítulo, é comentado sobre a função dos córpus para análise de estruturas linguísticas, além das fórmulas estatísticas utilizadas para a composição de amostras para análise.

6.1 Definições

No dicionário Aurélio da Língua Portuguesa (edição eletrônica 2005) o significado da palavra "definição"tem o seguinte enunciado:

Definir

verbo transitivo direto

1. Enunciar os atributos, as características específicas de uma coisa (objeto, idéia, ser) de tal modo que ela não se confunda com outra;

2. Dizer exatamente, explicar a significação de;

3. Demarcar, fixar;

4. Tomar resolução, decidir-se por;

Como pode ser visto, no dicionário, o ato de definir é caracterizado pela delimitação de fronteiras do sentido de um objeto, ideia ou ser, enfatizando as particularidades de maneira que transmitam o seu sentido real e o caracterizem unicamente.

A norma ISO 704 afirma que a definição pode ser de dois tipos: definição intensional e definição extensional. Para Lara (2004) as definições intensionais ou de gênero e espécie são aquelas onde se faz menção ao conceito genérico mais próximo, já definido ou supostamente conhecido e as características distintivas que delimitam o conceito a ser definido.

Para Juan Carlos Sager, importante pesquisador Argentino radicado na Inglaterra, a definição é "una descrición linguistica de un concepto, basada en el listado de un número de características que transmiten el significado del concepto"(SAGER, 1993, p.68)[1].

O autor coloca que as definições fazem um vínculo com os conceitos e os termos através de uma equação na qual o termo é a incógnita e mediante o ato de definir se cria a exata referencia de um termo a um conceito (SAGER, 1993).

De Bessé (apud SEPPÄLÄ, 2004) coloca que a definição permite fazer a ligação entre a realidade (referente) e o conceito, visto que o conceito é algo abstrato. Sendo então o elemento que está no centro do triângulo semiótico, figura 8.

O triângulo semiótico , figura 8, foi proposto pelos linguístas Ogden e Richards em 1923 para demonstrar o signo e a relação entre o objeto, o conceito e o termo. Sendo o referente o objeto real ,"qualquer parte de mundo concebido ou percebido"(ISO 1087), o conceito (referência) "uma unidade do pensamento constituída mediante uma abstração a partir das propriedades comuns de um conjunto de objetos", ISO 5963, e o termo (significante) como "uma designação de um conceito definido em uma língua especifica por meio de uma expressão linguística", ISO 1087.

Rey (apud SEPPÄLÄ, 2004) propõe um esquema do triângulo semiótico adaptado para a definição que ilustra sua visão sobre a função de interface entre o conteúdo semântico, a unidade linguística e a realidade externada pela linguagem. (Figura 9)

[1]Tradução nossa - Definição é uma descrição linguística de um conceito, baseado na lista de inúmeras características que transmitem o significado do conceito.

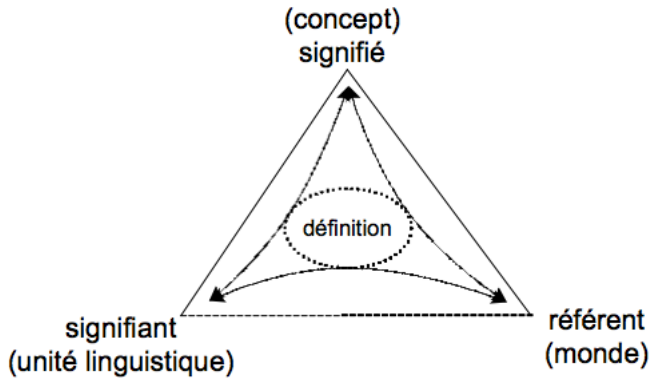

Figure 1 : Triangle sémiotique[22]

Figura 8: Triângulo semiótico
Fonte: (SEPPÄLÄ, 2004)

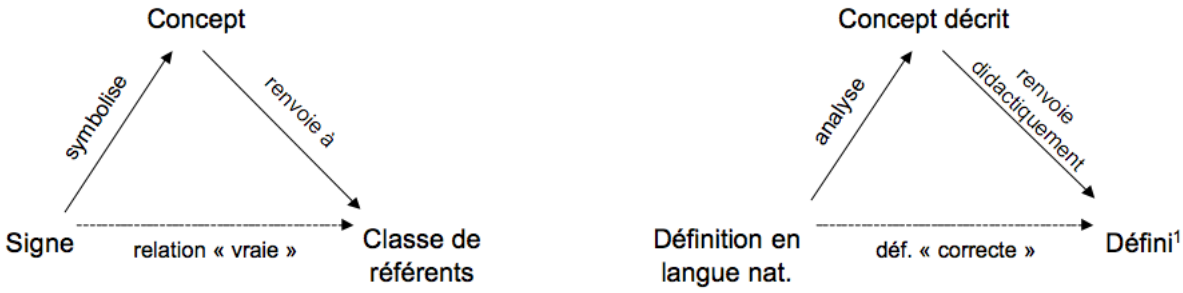

Figura 9: Triângulo semiótipo adptado para definição por Rey
Fonte: (SEPPÄLÄ, 2004)

Lara (2004) enfatiza a importância da definição em mecanismos de organização da informação via linguagem documentária, onde existe a necessidade de estruturar campos lógico-semânticos na elaboração de tesauros . Ela diz:

> Um dos instrumentos fundamentais para tal estruturação é a definição, que permite, dentre inúmeras possibilidades de organização, a determinação do campo de interpretação do termo e sua inserção em um campo temático. Para que isso fique mais claro, recorremos a um exemplo. Diante dos termos "casado", "solteiro", "viúvo", "separado", "divorciado, a constituição dos grupos casados e não-casados altera-se substancialmente conforme varia a definição de casamento tomada como ponto de partida. Vejamos: Casamento: evento relativo à união legal entre pessoas de sexo diferente. No Brasil, a legalidade da união é estabelecida no casamento civil, ou religioso com efeito civil, sendo que o indivíduo só poderá casar legalmente se o seu estado civil for solteiro,

> viúvo ou divorciado. Os casamentos têm como fonte principal as informações dos cartórios de registro civil. (Fundação Seade) Segundo a definição acima, poderemos agrupar em casados, os termos "casado", "separado"e, em "não-casados", "viúvo", "solteiro", "divorciado", supondo que a agregação é sustentada pela definição legal de casamento. Se, entretanto, a definição tomada for baseada nos costumes, e não na lei, o termo "separado"seria agrupado junto com os "não-casados", reorganizando-se a hierarquia (LARA, 2004, p. 01).

6.1.1 Tipos de definições

Larivière (1996, p.409), propõe a divisão das definições, conforme sua finalidade, em três tipos:

1. Definição lexicográfica (DL): utilizada nos dicionários de língua e enciclopédicos que se propõem a explicitar os significados distinguindo os sentidos e o emprego dos signos (ou palavras) de uma língua;

2. Definição enciclopédica (DE): utilizada nas enciclopédias e nos dicionários enciclopédicos, propõe-se a fornecer um conjunto de conhecimento sobre uma coisa;

3. Definição terminológica (DT): utilizada nos vocabulários especializados, propõe-se caracterizar (delimitar e distinguir de outras noções) as noções denominadas por um termo e que representam uma coisa no interior de um sistema organizado.

As definições Terminológicas são as mais encontradas em textos de especialistas, portanto, o tipo de definição estudada nesta pesquisa.

Flowerdew (apud AGUILAR, 2009) classifica as definições de acordo com suas estruturas:

1. Definições formais: são aquelas que apresentam uma estrutura do tipo gênero próximo + diferença específica: X = Y + características.

2. Definições semi-formais: são muito recorrentes em textos técnicos e se diferenciam da primeira pois especifica apenas a diferença específica.

3. Definições não formais : Não têm uma estrutura formal específica e podem ser representadas de forma linguística (uso de predicativos verbais, frases adverbiais, etc.) ou de forma não linguística (marcadores tipográficos, símbolos, fórmulas matemáticas, etc.)

Sierra e Alarcón (2003), também propõem uma tipologia de contextos definitórios que serão apresentados no item 5.2.3, de acordo com a estrutura de sua composição. Esta classificação é a seguinte:

1. Definição analítica ou aristotélica: quando é informado de forma explícita o gênero próximo e a diferença específica;
2. Definição sinonímica: quando apenas o gênero próximo é explicitado, estabelecendo uma equivalência conceitual com o termo que está sendo definido;
3. Definição funcional: quando apenas a diferença específica é explicitada, oferecendo uma definição de um conceito a partir do seu uso ou aplicação de uma situação dada;
4. Definição extensional: quando apenas a diferença específica é explicitada, apresentando uma definição que enumera os componentes que compõem um objeto representado pelo termo a definir. Esta enumeração de componentes seguem uma ordem baseadas em relações de todo e suas partes ou das partes e seu todo;

Os autores indicam que para cada tipo deste de definição existem verbos associados, mas detalharemos esses padrões linguísticos na seção 7.2.

6.1.2 Relações semânticas

As definições podem conter relações semânticas entre os conceitos apresentados em sua expressão. Sager (1993) sintetiza as duas principais, baseadas na estrutura gênero próximo + caracteristica específica:

Relação genérica (espécie-gênero): que pode ser descrita como "tipo de"mediante as fórmulas:

- X é um tipo de A
- X, Y , Z são tipos de A
- A contém X, Y, Z
- A contém o subtipo X

Relação partitiva (parte-todo): que indica ligação entre conceitos, os quais consistem em mais de uma parte e suas partes constituidas. Ocorre mediante as fórmulas:

- X é um componente de A
- X, Y, Z são componentes de A
- A consiste em X
- A consiste em X, Y, Z

Estudos na área porém identificaram que estes dois tipos são insuficientes para mapear o relacionamento entre os conceitos. Marshman (2003), por exemplo, estudou a extração de enunciados definitórios e as relações mais produtivas foram:

- relação hiperonímica/hiponímica: vincula um item específico ao seu correspondente genérico ou um genérico ao seu específico.
- relação meronímica: expressa vínculo entre um todo e suas partes;
- relação causal: vincula uma causa a seu efeito. As seguintes fórmulas (genéricas) podem ser empregadas para validar essa relação: X causa Y, em que X é o agente causal e Y é o efeito; ou X foi causado por Y, em que X é o efeito e Y é o agente causal;
- relação de finalidade: expressa a utilidade de uma entidade. A fórmula (genérica) X serve para Y, em que X é a entidade e Y é a função.

Seppälä (2004) propôs uma tipologia para anotação das diferenças específicas identificadas em enunciados definitórios de um corpus estudado em sua pesquisa. A autora afirma que os tipos foram sendo atualizados a medida que o trabalho foi sendo executado, até chegar neste quadro de tipos de relações, figura 10:

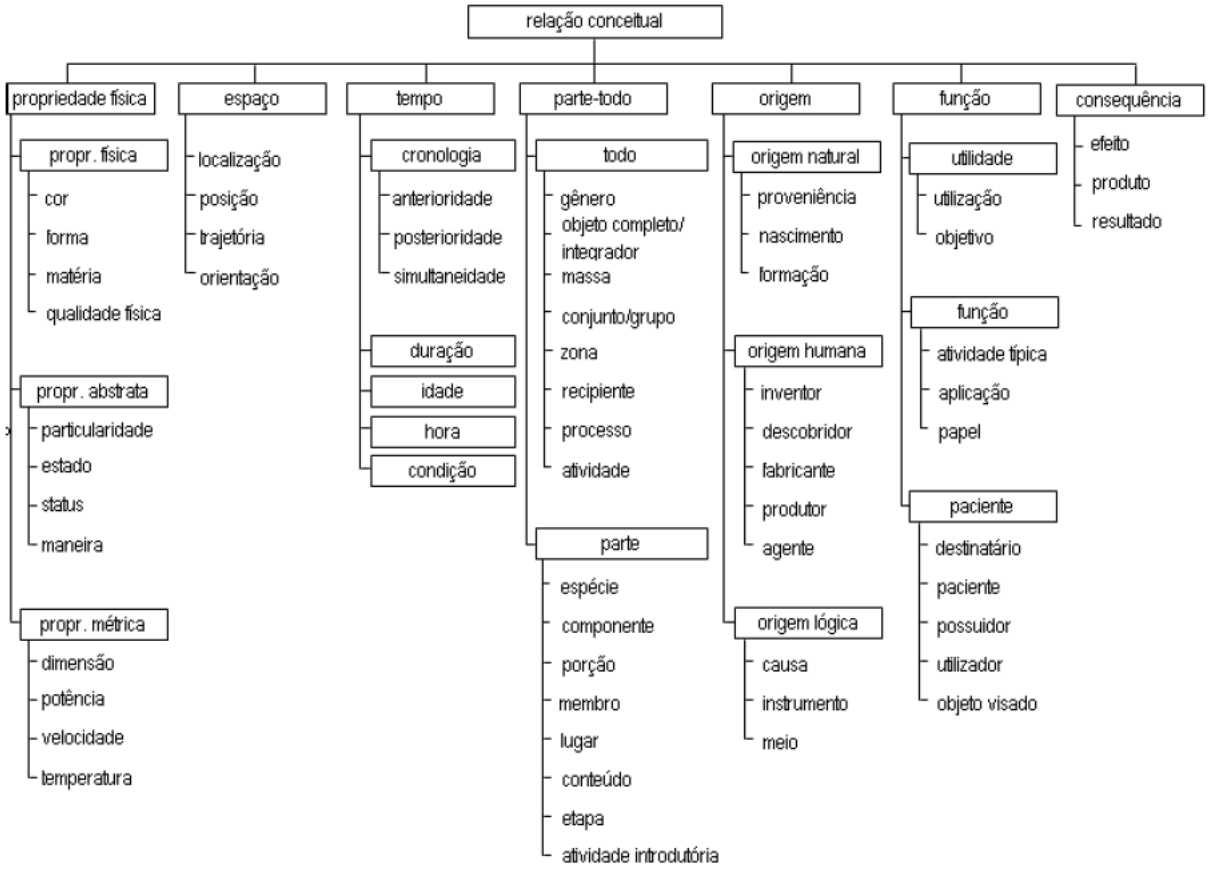

Figura 10: Tipologia conceitual de Sepalla
Fonte: (KAMIKAWACHI, 2009)

A estrutua Qualia foi proposta por Pustejovsky (1991), em sua obra intitulada *The Generative Lexicon*. O autor propõe quatro níveis de estruturação de uma unidade léxica: estrutura argumento, estrutura do evento, estrutura Qualia e estrutura de herança. A estrutura Qualia diz , segundo o autor, os atributos essenciais de um objeto. Função quale:

- Formal: distingue um objeto dos demais que pertencem a um domínio maior;
- Agentivo: apresenta elementos envolvidos na origem de um objeto;
- Télico: caracteriza a finalidade e função de um objeto;
- Constitutivo: contem a relação entre um objeto e suas partes constituintes.

Para Kamikawachi (2009) o construto da estrutura Qualia constitue questionamentos básicos que se fazem a respeito de unidades léxicas e por isso devem fazer parte de um definição terminológica.

Papéis *Qualia*
Formal: o que é *x*?
Constitutivo: *x* é feito de quê?
Télico: qual a função de *x*?
Agentivo: qual foi a causa de *x*?

Figura 11: Papéis Qualia de Pustejovsky(1991)
Fonte: (KAMIKAWACHI, 2009)

Auger (1997) , pesquisador Françês, em seu trabalho de extração semi-automática de termos e definições na língua francesa, propõe a seguinte tipologia:

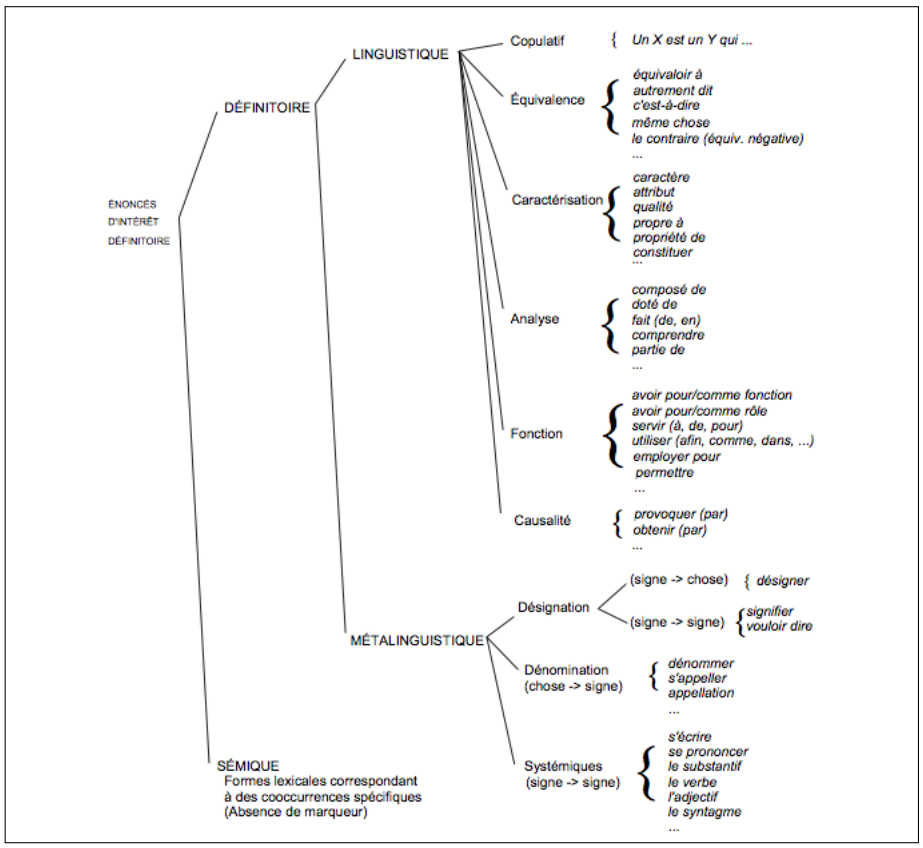

Figura 12: Classificação de Enunciados definitórios de Auger (1997)

Fonte: (AUGER, 1997)

Sua estrutura está dividida em dois tipos de enunciados definitórios os linguísticos e os metalinguístos, detalharemos esses tipos quando apresentarmos os enunciados definitórios no item 6.2.1

6.2 Definições em textos

Pearson (1998) em seu livro *Terms in Context* toma como ponto de partida os tipos de relação comunicativa possíveis em textos de especialistas e relaciona três tipos:

- Comunicação de especialista - especialista;
- Comunicação de especialista - profissionais da área em questão;
- Comunicação de especialista - principiante;

De acordo com Pearson estes três tipos não mantêm uma regularidade na identificação de definições, assinala que os últimos níveis são os que contêm o maior número de ocorrências devido a necessidade que o autor têm em esclarecer qualquer dúvida sobre o sentido dos conceitos tratados no texto.

6.2.1 Enunciado definitório

Para Auger (1997), um Enunciado Definitório (ED) é uma predicação que introduz e determina a informação conceitual associada a uma definição e identifica três elementos que compõem um ED.

- a) Um termo a definir;
- b) Uma expressão definitória;
- c) Uma partícula que associe o elemento a com o b, de modo que ambas as partes constituam uma estrutura predicativa, onde o termo funcione como sujeito e a expressão definitória como um predicado;

O autor estabelece dois tipos de EDs de acordo com o verbo que opera como núcleo:

- EDs com verbos linguísticos: aquelas predicações cujos verbos são de uso geral na língua e indicam sinonímia, funcionalidade, causalidade, etc. (ser/estar);

– EDs com verbos metalinguísticos: aquelas predicações cujo núcleo é um verbo que estabelece alguma ligação semântica significativa com a mesma linguagem;

A partir destes tipos de EDs, Auger propõe uma tipologia já apresentada na figura 12 no item 6.1.2.

6.2.2 Contextos ricos em conhecimento

Meyer (2001) define contextos ricos em conhecimento como contextos que indicam ao menos uma característica conceitual do termo, são um atributo ou relação e na prática terminográfica são úteis para:

– Prover definições;

– Prover pontos de partida pra formular definições;

– Incrementar o conhecimento do terminógrafo sobre a área em que trabalha.

Meyer (2001) ao estudar os Contextos ricos em conhecimento (CRC) propõe uma divisão em dois tipos: CRC definitórios e CRC explicativos. Os definitórios são os mais completos e seguem uma definição aristotélica: definição = gênero próximo + diferença específica, que na fórmula de Meyer é dada como:

X = Y + características distintas.

Essa fórmula contém: o X que representa o termo que se define; o Y é a classe geral que pertence X; as características distintas representam a informação que distingui X dos demais membros da sua classe; e o "=" que indica que tanto o X quanto suas características distintas devem poder mudar de posição sem alterar o sentido da oração.

Os explicativos são aqueles que só proporcionam informações sobre as características do termo, sem incluir a classe geral que este pertence.

6.2.3 Contextos Definitórios

Sierra e Alarcón (2003) a partir dos estudos de Meyer (2001) e Rodriguez (1999) e no âmbito do projeto coordenador pelo professor Gerardo Sierra, do Grupo de Engenharia Linguística, da Universidade Nacional do México – UNAM, propõem uma estrutura linguística para identificação de uma definição, o que eles chamam de Contexto Definitório (CD).

Sierra (2009) explica que partiu da definição de contexto proposta por De Bessé (1995) para definir um CD.

> Contexto é o entorno linguístico de um termo conformado por um enunciado, ou seja, as palavras ou frases ao redor deste termo e que condiciona sua existência, sua forma , seu funcionamento, seu significado, seu valor e seu emprego. Possui duas funções básicas: clarear o significado de um termo e ilustrar seu funcionamento (De Bessé, 1995, p. 03) .

Para Alarcón (2009), entende-se como Contextos Definitórios aqueles contextos de textos de especialistas onde contêm informação relevante sobre os atributos, características e relações conceituais dos termos. Estas informações para o autor, permitem entender o significado e a forma que aparecem com outros termos, além de conhecer as relações que estabelecem com outros termos para poder situá-los no contexto global do domínio de conhecimento que pertencem.

Aguilar (2009) entende Contexto Definitório como qualquer fragmento textual onde se introduza e associe um termo a uma definição. Os CDs são compostos de um termo (T), uma definição (D) que se encontram conectados mediantes a um padrão definitório (PD). Esses CDs podem apresentar outros tipos de informações metalinguísticas e pragmáticas referentes à forma, condições de uso ou alcance operativo que foi denominado por padrão pragmático (PPR) Sierra (2009)

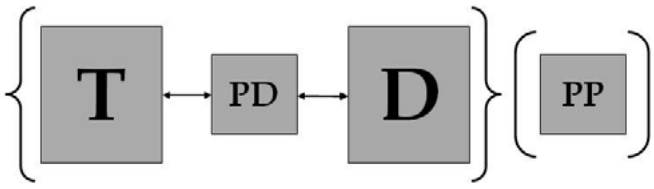

Figura 13: Estrutura de um Contexto Definitório
Fonte: (SIERRA, 2009)

Exemplo: <PPR> Tradicionalmente </PPR>, <T>la logística </T> <PD> se define como </PD><D> el arte militar que estudia el movimiento, transporte y estacionamiento de las tropas fuera del campo de batalla</D>. (SIERRA, 2009, p.17)

Os Contextos Definitórios são o objeto de estudo desta pesquisa e suas características serão melhor relatadas no capítulo 6, onde seus elementos serão apresentados com o enfoque da extração automática.

6.3 Córpus de análise

Tanto a linguística computacional quanto a terminologia utiliza córpus textuais para analises e estudos em linguagem natural para reconhecer estruturas linguísticas, termos e definições.

Uma das definições mais usadas e completas de córpus é a de Sanchez (apud SARDINHA, 2000):

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito na língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SANCHEZ, 1995 apud SARDINHA, 2000, p. 8) .

Sardinha (2000) identifica quatro pré-requisitos para a formação de um corpus computadorizado:

- Composição - Deve ser composto de textos autênticos, em linguagem natural.
- Autenticidade - Textos autênticos são escritos por nativos da língua.
- Seleção - O conteúdo do córpus deve ser escolhido criteriosamente, para que possa ter as características necessárias para sua análise. O objetivo do estudo é que determina como deve ser composto o córpus.
- Representatividade - O córpus deve ser uma porção representativa de uma variedade linguística, porém, para o autor, esta é o requisito mais complicado pois depende do objetivo do estudo. Cita vários critérios que podem ser utilizados para dar a representatividade de acordo com o estudo proposto: modo (falado ou escrito), tempo (um período, vários períodos), seleção (amostragem, estático, dinâmico), conteúdo (especializado, mais de um idioma), autoria (de aprendiz, nativo da língua), finalidade (de estudo, de referência, treinamento ou teste).

6.3.1 Repositório Institucional da Universidade de Brasília - RIUnb

O Repositório Institucional é um conjunto de serviços oferecidos pela Biblioteca Central da Universidade de Brasília para a gestão e disseminação da produção científica e acadêmica da comunidade universitária. Todo o seu conteúdo está disponível publicamente e, por estar

amplamente acessível, proporciona maior visibilidade da produção científica da instituição. Sobre os repositórios institucionais, Café et al. (2003, p.04) colocam que a função principal do repositório institucional é "[...] preservar e disponibilizar a produção intelectual da instituição representando-a, documentando-a e compartilhando-a em formato digital".

Para garantir a qualidade do material que será disponibilizado é essencial que, antes de se publicar no Repositório, ele tenha sido avaliado pelos pares, no caso de artigos de periódicos e trabalhos apresentados em eventos, afinal "o processo de avaliação de originais pelos pares é, até o momento, o que confere credibilidade ao conhecimento científico divulgado" (STUMPF, 2008, p.19). Para os demais materiais disponibilizados, a qualidade é garantida com a avaliação crítica daqueles com os mesmos interesses em comum.

O Repositório Institucional da UnB é constituído de material produzido pelos membros de sua comunidade acadêmica (professores, alunos, entre outros) em termos de artigos de periódicos, livros ou capítulos de livros, trabalhos apresentados em eventos e outros materiais que forem considerados pertinentes e relevantes como produção intelectual da comunidade da Universidade.

A partir deste repositório, foi feita a composição do córpus de estudo desta pesquisa. O tamanho da amostra e sua composição estão descritos no capítulo da metodologia, item 4.3.

7 Extração de Contextos Definitórios

Figura 14: Nuvem de tags do estudo

Fonte: Produzida pelo autor

Neste capítulo abordam-se os métodos para processamento automático de textos, além dos instrumentos necessários para extração de contextos definitórios. Divide-se da seguinte forma:

– Primeiramente, métodos e técnicas de processamento automático de textos são descritos. O Processamento de Linguagem Natural, a Descoberta de Conhecimento em Textos, a Extração da Informação são brevemente descritos, pois são ferramentas para a pesquisa.

– Em seguida, métodos de extração de contextos definitórios utilizados em língua espanhola são relatados, com detalhamento dos padrões identificados no trabalho de Sierra e Alarcón (2003).

– Ao final, o conceito de gramática definitória é descrito e caracterizado como um dos elementos fundamentais para possibilitar a extração automática de contextos definitórios. Algumas gramáticas em língua inglesa, francesa e espanhola são citadas, além de expressões linguísticas identificadas no trabalho de Kamikawachi (2009) que investigou uma base de definições em língua portuguesa.

7.1 Métodos para processamento de textos

7.1.1 Processamento de Linguagem Natural - PLN

O conteúdo armazenado em texto é considerado não estruturado e não pode ser manipulado por ferramentas de mineração de dados convencionais. Porém, na linguística, um texto é a unidade maior na estrutura de uma língua natural. Todo texto possui um padrão implícito, uma estrutura, que pode ser reconhecida, analisada e processada.

No início do século XX, mas precisamente em 1916, com a publicação do livro *Curso de Linguística Geral*, Ferdinand de Saussure (1857-1913) foi considerado o fundador da linguística moderna. Silveira (2003, p.23) enfatiza que "a partir dessa publicação, os estudos da linguagem conquistaram uma autonomia, centrada no reconhecimento de que a língua tinha uma ordem própria. Foi do reconhecimento dessa ordem enquanto estrutura que surge por volta de 1960 o estruturalismo."

O estruturalismo de Saussure se propunha a abordar qualquer língua como um sistema no qual cada um dos elementos só pode ser definido pelas relações de equivalência ou de oposição que mantém com os demais elementos. Esse conjunto de relações é que forma a estrutura. É uma abordagem que veio a se tornar um dos métodos mais extensamente utilizados para analisar a língua, a cultura, a filosofia da matemática e a sociedade na segunda metade do século XX. O método estruturalista consiste em se desmontar o objeto e remontá-lo visando-se entender suas relações internas, as leis ou regras que regem sua constituição e o seu funcionamento.

No livro Estruturas Sintáticas (1955), Chomsky apresenta sua idéia da gramática gerativa e sugere que a capacidade para produzir e estruturar frases é inata ao ser humano (isto é, é parte do patrimônio genético dos seres humanos). Apresentou, também, sua teoria de que os "enunciados"ou "frases"das línguas naturais devem ser interpretados em dois tipos de representação distintas: as "estruturas superficiais", correspondendo à estrutura patente das frases, e as "estruturas profundas", uma representação abstrata das relações lógico-semânticas das mesmas.

O método estruturalista é base para o Processamento de Linguagem Natural (PLN) aplicado em textos. Esta aplicação do PLN busca exatamente descobrir a estrutura ou os padrões lingüísticos de um texto e auxiliar na transformação de dados em informação de forma automática.

Podemos identificar quatro níveis de processamento de um texto, segundo Dias et al. (2007):

- Morfológico: quando as unidades mínimas dotadas de significado, os morfemas, são isolados para identificação dos traços de gênero, número e conjugação verbal

(pessoa, número, tempo);

– Sintático: quando a distribuição das palavras resulta em determinadas funções que elas desempenham na sentença. Para formar um enunciado dotado de um sentido completo, as palavras são combinadas seguindo uma regra estrutural bastante definida. Na manipulação dessas regras, faz-se uso de um conjunto de categorias definido em termos da sua função sintática (sujeito, objeto direto, complemento nominal e assim por diante) e classes gramaticais (substantivo, verbo, adjetivo, pronome, numeral, etc.);

– Semântico: quando o conteúdo significativo da palavra implica relações de natureza ontológica e referencial para a identificação dos objetos. O significado é inerente ao signo lingüístico e esta presente não só na palavra como unidade completa, mas nas suas unidades constitutivas;

– Pragmático-discursivo: quando a força expressiva das palavras remete à identificação dos objetos do mundo em termos do seu contexto de enunciação e condições de produção discursiva.

O Processamento de Linguagem Natural (PLN), de forma mais ampla, pode ser considerado a área que estuda e desenvolve mecanismos para o tratamento computacional da linguagem. Os primeiros estudos aconteceram no início da década de 50, quando o governo americano estimulou as instituições e pesquisadores a trabalharem com tradução automática. De lá para cá, o PLN mostrou evolução significativa, principalmente para a língua inglesa, entretanto ainda não proporciona a infra-estrutura exigida para oferecer o desejado suporte à Sociedade da Informação.

Podemos destacar os trabalhos do Núcleo Interistitucional de Linguística Computacional (NILC)[1] que, desde 1993, vem pesquisando sobre o PLN, contendo pesquisadores da Universidade de São Paulo (USP), Universidade Federal de São Carlos (UFScar) e Universidade Estadual Paulista (UNESP) de Araraquara. (BONFANTE, 2003; GREGHI, 2002; CASELI, 2007)

O projeto chamado Linguateca[2] também é uma referência na área. Possui pesquisadores do português de portugual e além disso, publica trabalhos de estudos em processamento da linguagem de todo o mundo.

[1] NILC : mais informações em http://www.nilc.icmc.usp.br/nilc/
[2] Linguateca : mais informações em http://www.linguateca.pt/

7.1.2 Descoberta de Conhecimento em Textos - DCT

A Descoberta de Conhecimento em Textos (DCT) é uma área para manipulação de textos e descoberta de conhecimento de forma automática.

Schiessl (2007) observa que documentos textuais que fazem sentido aos seres humanos, uma vez que esses reconhecem capítulos, parágrafos e sentenças, necessitam de pré-processamento antes de sua manipulação ou mineração por computadores. O autor informa também que a DCT é oriunda da Descoberta de Conhecimento em Dados (DCD), que compreende a seleção, o pré-processamento e adequação dos dados aos algoritmos, a utilização de técnicas de mineração e, finalmente, a análise e interpretação dos resultados para aquisição do conhecimento.

A DCT difere da DCD por lidar com dados não estruturados ou preparados para manipulação por computador. Para utilização de técnicas da DCD, ou *Data Mining* que, apesar de ainda não estarem consolidadas já se encontram em um patamar de amadurecimento aceitável, é necessário um processo de identificação de estruturas e padrões implícitos contidos nos textos. (WEISS et al, 2005).

Wives (2004), adverte entretanto que:

> A DCT não inclui somente aplicação de técnicas tradicionais de DCD, mas também qualquer técnica que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto. Com isso, muitos métodos foram adaptados ou criados para suportar esse tipo de informação semi-estruturada ou sem estrutura, que é o texto (WIVES, 2004, p. 24).

É na etapa de pré-processamento dos textos onde os processos da PLN trabalham efetivamente ao trazer um ganho na identificação e qualificação dos dados e consequentemente no processo de descoberta de conhecimento como um todo. Nessa etapa, o texto passa pelo processo chamado ‘tokenização’, que consiste em recortar o texto em unidades menores, as palavras, para tratamento. As palavras podem ser caracterizadas de diversas maneiras de acordo com sua natureza individual, sua função na sentença ou no texto. O uso de técnicas de PLN nessa etapa da DCT é comum, porém o nível de processamento ainda é bem superficial.

Geralmente, o morfológico é o único nível de processamento utilizado nos processos de DCT. A utilização de lista de palavras não significativas "stopwords"e a lematização, processo para retirar os afixos e sufixos dos termos, estão nesse nível de processamento, no qual a função ou o significado da palavra não é levado em consideração e se usam apenas métodos quantitativos para seleção e agrupamento de textos.

Bräscher (2002) destaca o problema em utilizar métodos apenas quantitativos na recuperação da informação em textos em língua portuguesa por existirem palavras que mudam completamente de significado de acordo com sua função linguística na sentença. Essa observação pode ser aplicada também na DCT. O simples fato de identificarmos que a palavra "cobre" é um verbo em uma sentença já propiciará uma classificação diferente do texto que possui a palavra "cobre" como nome de um elemento químico.

Para isso, é fundamental a utilização do segundo processamento, o processamento sintático, ou *parsing*. O *parsing* diz respeito à interpretação automática (ou semi-automática) de sentenças de linguagem natural por meio de programas de computador conhecidos como parsers ou analisadores sintáticos.

Unidade Lexical	Classificação Sintática
A	ART
Tecnologia	SUB
Da	PRE
Informação	SUB
Constrói	VER

Figura 15: Exemplo de etiquetagem sintática
Fonte: Produzido pelo autor

Para textos em português existem alguns analisadores sintáticos, com destaque para o trabalho de Maia (2008) que desenvolveu uma ferramenta chamada Ogma[3]. Ogma tem a finalidade de extração de Sintagmas e ,para isso, faz uma etiquetação sintática em língua portuguesa.

Existem várias abordagens ou técnicas de Descoberta de Conhecimento em Texto. Loh (1999) identifica 17 tipos e subtipos diferentes, englobando todas as manipulações a partir de texto ou partes do texto para identificação de informações. As principais citadas por Wives (2000) são:

- Agrupamento ou *clustering* - A técnica de agrupamento ou generalização é a abordagem tradicional de DCT. O processo consiste na seleção, tratamento e estruturação dos dados do texto para utilização de técnicas de mineração de dados. Tem o propósito de identificar documentos que fazem parte de um mesmo padrão e classifica-los em grupos. A descoberta está na criação sistêmica de classificações.
- Classificação ou categorização – a abordagem de classificação utiliza o mesmo processo da anterior, porém o usuário (analista da informação) informa os grupos (tema

[3]Ogma - ferramenta disponível em: http://www.luizmaia.com.br/ogma/ .

ou assunto) e o sistema seleciona os documentos que fazem parte daquela classe, descobrindo as características principais de cada uma, que possam identificá-la para o usuário e distingui-la das demais classes. Pode ser usada no SRI no auxílio ao processo de indexação automática.

– Sumarização – essa técnica tem por objetivo selecionar as frases mais significativas do documento ou de uma série de textos e produzir um resumo ou sumário. Pode utilizar a técnica de Extração de Informação (EI) mencionada a seguir.

– Extração de informação (EI) – Essa técnica consiste em retirar do texto informações específicas, segundo um padrão informado, e representar essa informação de forma estruturada. Esta técnica está presente no fluxo da DCT, na fase de pré-processamento e transformação. Por se tratar da técnica utilizada na pesquisa, detalharemos suas características na próxima sessão.

Estas técnicas podem ser combinadas e várias abordagens podem ser utilizadas de acordo com o objetivo da extração. Wives L.; Loh (2000) classificando, ainda, os tipos possíveis de mineração de textos, comentam sobre o método de Descoberta por análise linguística, que consiste na descoberta de generalizações escondidas através da análise de padrões sintáticos nos textos. Outro tipo caracterizado pelos autores é a Descoberta por análise de conteúdo, que propõe a investigação linguística dos textos para apresentar informações sobre o tema, assunto ou até mesmo um índice ou resumo. O presente estudo pode ser enquadrado nesses dois tipos de Descoberta propostos por Wives L.; Loh (2000).

7.1.3 Extração da Informação - EI

A Extração de Informação (EI), cujo objetivo é encontrar informações específicas dentro dos textos, pode ser feita isolando-se partes relevantes do texto, extraindo-se informações destas partes e transformado-as em informações mais digeridas e melhor analisadas.

Os primeiros estudos na EI foram na Ciência da Computação, dentro da área de Processamento de Linguagem Natural (PLN). Para Scarinci (1997), a Extração de Informação diverge da PLN por não se ater a todo processamento do texto, apenas às partes específicas, as que se deseja extrair, determinadas pelo usuário do sistema através de padrões a serem analisados.

A Extração de Informação ignora partes do texto que não casem com um domínio pré-definido de normas, que devem ser tão claras e exatas quanto for possível especificar. Esses padrões podem ser morfológicos, sintáticos ou com auxilio de elementos de representação do conhecimento, semânticos e pragmáticos.

Segundo Scarinci (1997), os processos de extração podem ser:

- Estatísticos: com base em freqüência de ocorrência de padrões;
- Léxicos: com base na ocorrência de termos simples ou compostos ou por formatos de termos;
- Sintáticos: com base nas relações entre termos, ou baseados em conhecimento, com uso de regras pré-definidas.

Ao se ater a partes do texto que interessam ou são relevantes para o usuário que o manipula, seu processamento é mais rápido que o PLN. Entretanto, possuem os mesmos desafios para manipulação de documentos, sendo necessário um grande domínio do contexto e da estrutura dos objetos informacionais manipulados quando os níveis de processamento saem do morfológico e sintático.

Segundo Wives (2000):

> As técnicas de Extração de Informações (EI) não possuem uma classificação muito bem definida. Elas podem ser enquadradas na área de RI, pois são compreendidas algumas vezes como técnicas especiais de indexação ou por extraírem de um texto ou conjunto de textos somente as informações mais relevantes para o usuário. Por outro lado, se não fossem extraídas, talvez essas informações não fossem facilmente identificadas pelo usuário (poderiam estar implícitas ou passar despercebidas). Vistas dessa forma, elas são enquadradas na área de descoberta de conhecimento (WIVES, 2000, p. 91) .

Contudo, Constantino (1997) afirma existir uma diferença grande entre a RI e a EI. Para o autor, a Recuperação da Informação tem como foco identificar documentos relevantes em uma coleção e a EI visa identificar informações relevantes em um documento e produzir uma representação dessa informação.

A Extração da Informação é um meio que pode auxiliar a indexação automática, a Recuperação da Informação, a Descoberta de Conhecimento em Textos e a Organização do Conhecimento. Ao extrair informações conforme um padrão definido pelo usuário e armazenar de forma estruturada o conteúdo contido em textos que fação parte de um mesmo domínio, contribui para diversos processos de organização da informação e do conhecimento.

7.2 Identificação de Contextos Definitórios

Os pesquisadores mexicanos Gerardo Sierra e Rodrigo Alarcon, segundo Aguilar (2009), delinearam um projeto de pesquisa orientado para o reconhecimento e extração de termos e

definições em textos técnicos e científicos, particularmente situados em contextos definitórios.

Este projeto, desenvolvido no Grupo de Ingeniería Linguística (GIL) da Universidad Automata do México, têm gerado vários resultados para essa linha de pesquisa, como:

- Uma descrição linguística sobre o comportamento dos CDs, junto com suas unidades constitutivas, em textos de especialistas em espanhol (ALARCON, 2003).
- Identificação de um grupo de verbos associados a predicações verbais cuja função é servir como nexo entre termos e definições (SIERRA; ALARCÓN, 2003).
- Uma delimitação de uma tipologia de definições baseada em relações que estabelecem com o tipo de predicação verbal que se vincula (AGUILAR et al., 2004).

A partir desses trabalhos, além das teses de doutorado de Aguilar (2009) e Alarcón (2009), alguns elementos importantes para a extração automática de Contextos Definitórios foram relatados, neste capítulo, como formas de se identificar um CD em um texto de especialista.

Para Sierra (2009), um elemento chave no processo de reconhecimento de Contextos Definitórios de forma automática é a identificação de padrões que servem para conectar o termo com sua definição ou para ressaltar visualmente sua presença dentro do texto.

Sierra e Alarcón (2003) identifica entre os elementos de um CD esses padrões, os chamam de padrões definitórios e os classificam inicialmente em 4 tipos: padrões tipográficos, padrões sintáticos, padrões mistos e padrões compostos. Em trabalho mais recente (ALARCÓN, 2009), os dois primeiros foram os únicos adotados, visto que os padrões mistos possuem os tipográficos e os sintáticos e os compostos definem mais de um termo, sendo assim divididos em sua classificação recente:

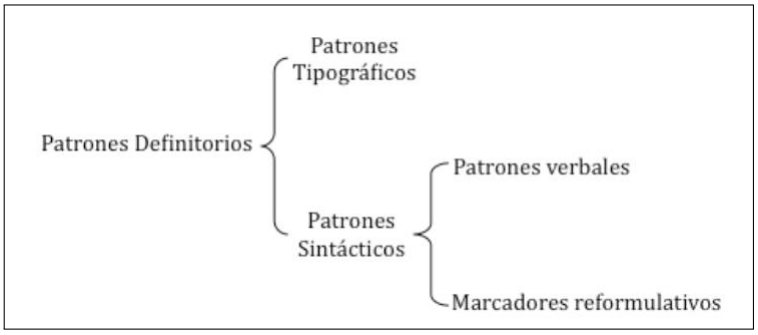

Figura 16: Tipologia de padrões definitórios
Fonte: (ALARCÓN, 2009)

Esses padrões descritos na figura 16 serão detalhados na sessão 7.2.1 e 7.2.2. Pearson (1998) também considera que existe uma série de padrões gramaticais que permitem associar

um termo com uma definição, porém o autor classifica os dois tipos de Sierra e Alarcón (2003) em um único padrão denominado padrão metalinguístico e cria outro, realizadores definitórios. Os dois tipos de Pearson (1998) são detalhados como:

- Padrões metaliguísticos: elementos sintáticos ou tipográficos que servem para realçar um termo ou outra unidade de informação conceitual, podendo ser frases que explicam a respeito do sentido de um termo, como, por exemplo: neste sentido, para este trabalho, de acordo com, etc, ou também elementos tipográficos, como negrito, parênteses, etc;
- Realizadores definitórios: que se dividem em dois tipos. Sendo o primeiro aquele que introduz pela primeira vez a definição de um termo e o segundo aquele que mostra explicações sobre um termo que já tenha sido definido anteriormente. Em ambos os casos, para o autor, operam predicações verbais do tipo "is a", "is defined as", "consists of"e similares.

Detalharemos a seguir os tipos de definição propostos por Sierra e Alarcón (2003).

7.2.1 Padrões tipográficos

Sierra e Alarcón (2003) afirmam que a tipologia de um texto serve como ajuda visual para o leitor identificar facilmente algum elemento importante e separá-lo do resto do texto comum, como os termos e suas definições.

Considera ainda que em alguns casos se define um termo sem a necessidade de ter um verbo como conector, sendo substituído sintaticamente, por signos de pontuação como dois pontos, ponto e vírgula ou vírgula. Como exemplo:

- Diseñõ: Desarrollho de configuraciones para la resolución de algún problema en base y sujetándose a sus restricciones (SIERRA; ALARCÓN, 2003).
- Desastre. Pertubación de la actividad normal que ocasiona pérdidas o daños extensos o graves (ALARCÓN, 2009).

Alarcón (2009) relata que os padrões tipográficos mais frequentes encontrados para resaltar os elementos constituintes dos CDs foram o itálico, negrito, sublinhado, letras maiúsculas, cabeçalhos e, entre as pontuações, os dois pontos e o ponto e traço.

7.2.2 Padrões sintáticos

Sierra (2009) enfatiza que um caminho para extrair de forma automática CDs em textos de especialistas é identificar as estruturas sintáticas recorrentes dos conectores que unem os elementos que compõem o CD.

Alarcón (2009) afirma que estes conectores sintáticos que ligam os elementos de um CD podem ter como núcleo um verbo e, nesse caso denomina-os de padrão verbal. No caso da estrutura possuir outro tipo de forma sintática o autor os classifica de marcadores reformulativos.

Exemplo: "El índice secundário es a menudo um índice denso, es decir, contiene todos os valores posibles de la clave primaria"(SIERRA, 2009, p.20).

Alarcón (2009) e (SIERRA, 2009) caracterizam também outro tipo de padrão, o pragmático, que, na visão dos autores, identifica explicitamente as condições de uso ou o alcance do termo definido, como a localização geográfica, as instituições que utilizam, o nível de especialização, a frequencia de uso, etc. Esses padrões são muito úteis junto com os verbais para identificar um CD no texto quando não existem padrões tipográficos (SIERRA, 2009).

Aguilar (2009), ao estudar a ligação das predicações na estrutura dos Contextos Definitórios, em especial nos padrões sintáticos, considera que a estrutura predicativa estabelece uma sequência de organização sintática entre termos, verbos e definições de tal modo que o termo pode ocupar a posição de sujeito ou de objeto. Estabelece ainda que os verbos, quando operam com núcleos do predicativo, têm uma ligação estreita com a definição de tal maneira que pode determinar o tipo de ligação que a definição terá.

O autor considera que existem dois tipos possíveis de estrutura sintática:

- Predicação primária: Uma sequência do tipo termo + verbo + definição, no qual o termo é o sujeito, o verbo, o núcleo e a definição é o predicativo que se associa ao sujeito. Exemplo: um eror de programación es un falho en la semántica de un programa (AGUILAR, 2009, p.81).
- Predicação secundária: Uma sequencia do tipo autor + termo + verbo + definição, quando o sujeito indica o autor da definição, o termo equivale a um objeto da predicação, o verbo opera como núcleo e a definição é introduzida através do predicado associado ao objeto. Exemplo: Turing definió la inteligencia artificial como aquella inteligencia exhibida por artefactos creados por humanos (AGUILAR, 2009, p.82).

7.3 Gramática de padrões definitórios

No âmbito dos trabalhos do grupo de engenharia linguística da Universidade do México, a partir dos trabalhos de Sierra e Alarcón (2003) e, posteriormente, Aguilar (2009), o grupo identificou uma série de verbos que podem identificar um contexto definitório.

Definición	Verbo	Adverbio o preposición	Unidades nominales	Tipo de Predicación
Analítica (género próximo/ diferencia específica)	Referir Representar Ser Significar	*A*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
	Caracterizar Comprender Concebir Conocer Considerar Definir Describir Entender Identificar Visualizar	*Como* *Por*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Secundaria
Sinonímica (género próximo)	Denominar Equivaler Llamar Nombrar Ser	*También* *A* *Igual a* *Similar a*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
Funcional (diferencia específica)	Emplear (se) Encargar Funcionar Ocupar Permitir Servir Usar Utilizar	*De* *Para*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
Extensional (diferencia específica)	Componer Comprender Consistir Constar Contar Constituir Contener Incluir Integrar Es/son parte Es / son + **:** (dos puntos)	*De* *Por* *Con*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria

Figura 17: Gramática de padrão definitório em Espanhol proposta por Sierra e Alarcón (2003), Aguilar (2009)
Fonte: (AGUILAR, 2009)

A figura 17 identifica os verbos que compõem a gramática de padrões definitórios proposta pelo grupo GIL em língua espanhola. Esses verbos identificam os tipos de definição propostos por Alarcón (2009) citados no item 6.1.1, o tipo de advérbio, preposição e as unidades nominais necessárias para que o verbo seja definitório, além do tipo de predicação como classificou (AGUILAR, 2009), item 7.2.2.

O pesquisador Auger (1997) também utilizou uma série de verbos definitórios na identificação de estruturas definitórias na língua francesa. Citamos, a seguir (Figura 18), parte de sua gramática.

Catégorie	Marqueur
analyse	comporter
analyse	composer de
analyse	comprendre
analyse	constituer de
analyse	décrire
analyse	doter de
analyse	fait (de, en)
analyse	formé de
analyse	munir de
analyse	partie de
analyse	pièce de
caract.	apparenter à
caract.	attribut de
caract.	caractère de
caract.	caractériser
caract.	caractéristique
caract.	communément
caract.	dans le domaine
caract.	dans le jargon
caract.	dans le langage (la langue)
caract.	dans la parlure
caract.	dans le parler
caract.	définir
caract.	définition
caract.	dénotation, dénoter

Figura 18: Parte da gramática de padrão definitório em Francês proposta por Auger
Fonte: (AUGER, 1997)

Esta gramática de Auger (1997) também relaciona os verbos com os tipos de definições propostas por ele e citados neste trabalho no item 6.1.2.

Em inglês, temos o trabalho de Rodriguez (2004), que identifica uma série de verbos definitórios em suas investigações para extração automática de estrutura definitória em textos.

Verb or VP	Ocurrences
call	36
Is, use	19
refer	17
known as	8
use, refer	5
Imply, mean	4
Apply, define, restrict	3
Designe, use, mean	2
amounts to, coins, conceptualize, consider, correspond, denote, dub, evoke, extend, favor, has, has become, includes, indicates, insist, name, note, reserve, speak, stands for, substitute, strech the meaning, taken to embrace, termed, terms	1

Figura 19: Verbos mais identificados em análise de córpus de sociologia por (RODRIGUEZ, 2004)
Fonte: (RODRIGUEZ, 2004)

Na figura 19 acima, Rodriguez identifica os principais verbos encontrados em estruturas definitórias em córpus de estudo da área de sociologia.

Marshman (2003), pesquisadora canadense, faz um trabalho no qual analisa dois corpos em línguas diferentes, o francês e o inglês. Em português, entretanto poucos trabalhos foram identificados nesta revisão. Podemos citar o trabalho das portuguesas Pinto e Oliveira (2004), que analisaram um córpus em português de portugual, mas a estrutura de verbos utilizada no trabalho não foi possível ser acessada.

Já no Brasil, Kamikawachi (2009) ao analisar uma base de definições já previamente anotadas em um córpus do grupo Geterm, da Universidade de São Carlos, identifica uma série de expressões linguísticas que compõem as definições. A autora separa essas expressões nos tipos semânticos de definições analisados por ela, baseada nas classificações propostas por Seppälä (2004) e Pustejovsky (1991).

Na figura 20 , citamos os tipos de expressões identificas no trabalho de Kamiquawachi para o tipo de relação semântica denominada agentivo. As demais expressões em português, separadas por sua classificação semântica, encontradas no estudo de Kamikawachi (2009) e utilizadas neste estudo, fazem parte do anexo desta pesquisa.

Expressões linguística – AGENTIVO	Expressões linguística – AGENTIVO
As causas mais freqüentes são	Obtida por
As principais causas ou fatores de risco são	Obtido a partir
Causada pela	Ocorre devido ao
Causada por	Pode ocorrer devido à
Causado pela	pode ocorrer pela
Causado pelo	Pode ocorrer por
causado pelos	pode ser causada por
como proveniente	Pode ser causado por
de origem	Pode ser de origem
Decorre de	pode ser desencadeado por
decorrente de	Pode ser obtido por
desenvolvida pelo	Pode ser provocada por
Desenvolvida por	Pode ser resultado de
Devido a	Pode ter origem
Devido à	podem ser provocadas por
devido ao	provocada geralmente por
É causado pela	provocada pela
É ocasionada por	Provocada por
É originada pela	que ocorre devido à
É proveniente de	que podem ser originados
É provocada por	que provém do
É provocada por	realizada pelo
É provocado pela	realizado pelo
e são geralmente provocadas	resultante de
Em consequência	São causadas por
em decorrência	São fatores que propiciam
Em decorrência de	Suas causas mais freqüentes são
Obtém-se por	Suas causas possíveis são
obtida pelo	trazida da

Figura 20: Expressões identificadas no trabalho de kamiquawachi para o tipo de relação semântica agentivo
Fonte: (KAMIKAWACHI, 2009)

Nesse capítulo foram apresentadas as técnicas e métodos de manipulação automática da informação armazenada em textos como o Processamento de Linguagem Natural, com foco na manipulação de textos, e a Descoberta de Conhecimento em Textos, com ênfase no método de Extração da Informação, além dos elementos utilizados em trabalhos anteriores de identificação e extração de estruturas definitórias como padrões definitórios e gramáticas definitórias em língua espanhola Sierra e Alarcón (2003), Sierra (2009), em língua francesa Auger (1997), em inglês com Rodriguez (2004) e as expressões linguísticas em português encontradas no trabalho de Kamikawachi (2009) junto a uma base de definições do grupo Geterm, da Universidade de São Carlos.

Parte I

Resultados

8 Criação da gramática de padrões definitórios

8.1 Breve análise da revisão de literatura

Após a revisão de literatura podemos identificar 3 grandes grupos de organização:

- Organização da Informação que utiliza a análise de conteúdo para as ocorrências individuais de objetos informacionais.
- Organização do Conhecimento que tem por base a análise do conceito. Organiza as unidades do conhecimento de vários documentos pertencentes ao mesmo domínio.
- Propomos enquadrar, ainda, os processos de catalogação, gestão de arquivos e bibliografia, que têm um olhar para a análise física dos objetos, como Organização de Documentos.

A Ciência da Informação têm como elemento base a informação registrada e o fluxo principal é organizar para recuperar, contudo nos repositórios já constituídos de documentos textuais e que adotaram a Organização de Documentos para se recuperar informação, o processo se inverte e a extração de informação, nos ítens em liguagem natural, possibilita uma reorganização destes elementos melhorando a eficiência dos sistemas de recuperação da informação.

A análise das teorias descritas no capitulo 03 demonstram técnicas tanto da linguística quanto da Ciência da Informação para adentrar nos textos e representar de forma fiel os conceitos ali tratados, possibilitando uma compreensão sobre cada documento e sobre o repositório e sua área do saber.

Os termos e as definições contidas nos textos são elementos fundamentais para o mapeamento semântico dos contéudos neles contidos, porém é necessário separa-los e reorganiza-los de forma a auxiliar em sua recuperação e efetiva comunicação aos usuários de sistemas de informação.

Os processos automáticos citados nesta pesquisa são uma tentativa de auxiliar neste processo de reorganização e re-representação da informação e do conhecimento contido nos textos.

Um mapa mental contendo as áreas de estudo sobre o conhecimento representado em texto, baseado na revisão de literatura, e suas relações com as Ciências foi desenvolvido com a finalidade de mapear os elementos que compõe os estudos desta área. Não tem a finalidade de dividir os elementos em cada Ciência, visto que isso não é possível nos dias atuais, mas apenas visualizar as interações de seus elementos.

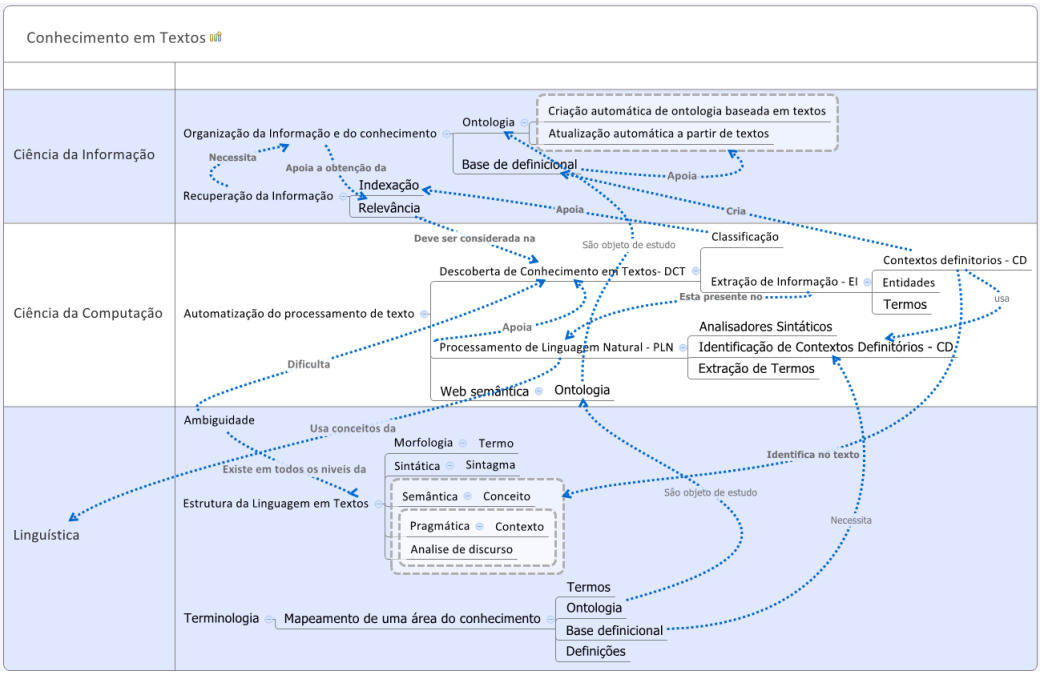

Figura 21: Mapa mental de estudos em conhecimento em textos

Fonte: Produzido pelo autor.

8.2 Análise manual dos documentos da amostra

Conforme dito na metodologia, a amostra é composta de 179 documentos, sendo 53 teses e 126 dissertações da Faculdade da Ciência da Informação da Universidade de Brasília, do período de 2006 a 2011.

Este grupo de documentos disponibilizados no Repositório da Universidade contêm algumas teses e dissertações anteriores a este período, mas que só foram disponibilizados em meio digital em 2006. O grupo com todas as teses e dissertações com acesso via repositório compôs o grupo total.

Segundo a fórmula apresentada no item 4.1 da metodologia foi possível calcular a amostra ideal para analise do grupo de documentos selecionados para esta pesquisa. Com 95% de grau de confiança, erro amostral considerado de 13, um máximo de 150 e o mínimo de 65 ocorrências em cada documento e uma população de 179 documentos, identificou-se que seria necessário a analise de 10 documentos de forma manual.

O site de estatística, encontrado em www.random.com, foi utilizado para selecionar de forma aleatória os 10 documentos a serem investigados. Conforme o detalhamento da metodologia , item 4.5, os 10 documentos foram separados em dois grupos de 5 para analise. O primeiro grupo foi composto, de acordo com a tabela 3, abaixo. O segundo grupo será descriminado no item 8.4.3.

Tabela 3: Relação dos documentos do primeiro grupo analisados.

ANO	AUTOR	TíTULO
2006	João Pereira Marciano	Segurança da Informação: uma abordagem social.
2007	Tiago Miranda Marques	Abordagens de recomendação para recuperação de perfis: uma proposta de modelo
2008	Grazielle Noronha Campos	Características e perfil dos bibliotecários das bibliotecas de instituições de ensino superior privadas do Distrito Federal e as expectativas dos empregadores

Continua na próxima página...

Tabela 3 – Continuação...

ANO	AUTOR	TíTULO
2009	Wagner Junqueira de Áraujo	A sergurança do conhecimento nas práticas da gestão da segurança da informação e da gestão do conhecimento
2011	Fernando Silva	Critérios de seleção de obras raras adotados em bibliotecas do Distrito Federal

Fonte: Produzida pelo autor.

Optou-se por adotar a estrutra Qualia de Pustejovsky (1991) para identificação e classificação de contextos definitórios. A estrutura Qualia responde as principais indagações sobre um objeto, qual seja: o que é, do que é feito, qual sua função e sua causa.

Ao analisar os 5 primeiros documentos foram encontrados 537 estruturas definitórias divididas, segunda a classificação de Sierra e Alarcón (2003) quanto ao seu padrão de identificação, conforme tabela abaixo:

Tabela 4: Total de Contextos Definitórios do primeiro grupo por padrão.

Código do Documento	Padrão tipográfico	Padrão sintático
Doc01	03	70
Doc02	03	69
Doc03	16	120
Doc04	10	143
Doc05	13	90
TOTAL =>	45	492

Fonte: Produzida pelo autor.

Os padrões tipográficos não são o foco deste pesquisa, a busca automática se realizará em estruturas sintáticas, a partir de uma gramática criada no âmbito do estudo. Por isso, neste momento buscamos identificar as estruturas linguísticas que apareceram na análise manual.

Identificamos na pesquisa, algumas estruturas e as separamos de acordo com a presença e a ordem dos elementos, tais como, Autor, Termo , Definição e expressões linguísticas, objetos que podemos definir como:

- Autor - Elemento a quem podemos atribuir a definição.
- Termo - Objeto a ser definido ou caracterizado.
- Definição - Qualquer enxerto definitório, não necessariamente contêm todas as características do objeto.
- Expressão linguística - qualquer elemento da língua que possa identificar a presença de um contexto definitório, de forma única ou em uma composição. Estes elementos são os primeiros a comporem a gramática de padrões definitórios proposta, um dos resultados desta pesquisa.

A seguir, segue um detalhamento das estruturas identificadas, com a relação das expressões linguísticas apresentadas conforme apareceram nos textos analisados.

Estrutura 01: Esta estrutura é uma das mais completas, pois podemos identificar duas expressões linguísticas que mapeiam o contexto definitório, quando combinadas. Alem disso, ela possui uma primeira Expressão linguística, o Autor da definição, o Termo, uma segunda Expressão e a Definição bem delimitados. Esta estrutura corresponde a 12% dos contextos definitórios identificados, com padrão sintático, nos documentos analisados manualmente e iremos chamá-la de EATED.

Expressão linguística 01	AUTOR		TERMO	Expressão linguística 02	DEFINIÇÃO
De acordo com Segundo De acordo também com Na definição do Para A definição apresentada por Do ponto de vista Conforme apresentado por	AUTOR	,	TERMO	é um é " é a é antes de tudo esta relacionado com esta relacionado á quer dizer é prepara : observa busca está a qual se caracteriza é visto como é aquela que é definido como podem ser definidos como: é definida como: é definido como: refere-se em seu sentido mais atual, pode ser considerado designa apresenta são caracteriza-se pelo é caracterizada por	DEFINIÇÃO

Figura 22: Estruturas linguísticas, EATED, encontradas nos documentos analisados
Fonte: Produzido pelo autor.

Exemplos de estrutura EATED:

No trabalho de Araújo (2009):

> "Segundo Von Krogh et al. (2001): A criação de conhecimento é um processo frágil, que não se sujeita às técnicas de gestão tradicionais. " (ARAÚJO, 2009, p. 81)

No estudo de Campos (2008):

> "Para Flory(2005), empregabilidade é a qualidade de manter-se no mercado, ser desejado pelos alvos e coerentes com a missão. " (CAMPOS, 2008, p. 20)

Estrutura 02: Esta estrutura difere da primeira apenas na ordem dos elementos, tendo o Autor como primeiro elemento. Contudo, também é completa, contendo alem do Autor, uma Expressão, o Termo, outra Expressão e a Definição. Correspondeu a 14% dos contextos definitórios e a chamaremos de AETED.

AUTOR	Expressão linguística 01	TERMO	Expressão linguística 02	DEFINIÇÃO
	afirma		é a	
	afirma que		é	
	afirmava que		é	
	sugere que		deve ser definida	
	entende que		é:	
	entendem que		é um	
	alertam que		se dedica à	
	afirmam que		deve servir como	
	que afirma que		consiste em	
	quando afirma que		diz respeito a	
	para quem		é	
	consideram		como	
	considera		como	
	conceituam		como	
	conceitua		como	
AUTOR	define	TERMO	da seguinte forma:	DEFINIÇÃO
	define		como sendo	
	determinou os principais compo		que são	
	entende		como	
	apresenta		como	
	em cujo prefácio ele define		como	
	situa		como	
	situam		como	
	lembram		como	
	entende-se por			
	entendem		como	
	preconizam		como parte da	
	considera-se		como	
	definem		como	
	define		em	
	define		como	

Figura 23: Estruturas linguísticas, AETED, encontradas nos documentos analisados
Fonte: Produzido pelo autor.

Exemplos de estrutura AETED:

Em Marciano (2006):

> "Husserl define a verdade como sendo a concordância perfeita entre o significado (formulado pelo observador) e o que é dado (objeto). contextualizando o conhecimento como mais um dos fenômenos de estudo vistos por meio do epoché. (STEGMÜLLER, 1977, p. 58-91)." (MARCIANO, 2006, p. 33)

No texto de Campos (2008):

> "Wilson (2006) afirma que a gestão do conhecimento é uma extensão dos conceitos da gestão da informação." (CAMPOS, 2008, p. 77)

Na pesquisa de Silva (2011):

> "Cunha (2008), p. 234, define o livro raro como o livro que, pelas características da edição, existência de autógrafo do autor ou alguma razão especial, é considerado valioso" (SILVA, 2011, p. 34)

Estrutura 03: A terceira estrutura é mais simples, pois o autor não esta bem caracterizado, porém é possível mapear o Termo, uma Expressão linguística e a Definição. Esta estrutura corresponde a maior percentagem encontrada na amostra inicial com quase 62% dos contextos definitórios e a chamaremos de TED.

TERMO	Expressão linguística 01	DEFINIÇÃO
TERMO	designa a deve ser vista como devem ser entendidos como devem ser vistas como é é " é a é a variável que é apresentado constantemente como é aquela em que é aquele é aquilo que é caracterizada é composto de é composto pela é composto por é conhecida como é considerado é definido como é definido pela é encarado como é entendida como é o é o conceito de é também o resultado é um é um termo utilizado para designar é um tipo de é uma é vista como é, por consequência, entendidas como enxerga o esta atrelado às está sendo utilizado para fazem parte de uma funciona como identifica	DEFINIÇÃO

Figura 24: Estruturas linguísticas, TED, encontradas nos documentos analisados
Fonte: Produzido pelo autor.

Exemplos de estrutura TED:

Em Araújo (2009):

> "O conhecimento é também o resultado dos relacionamentos que a organização manteve ao longo do tempo com seus clientes, fornecedores e parceiros.(CHOO, 2003, p.179)" (ARAÚJO, 2009, p. 57)

Em Marques (2007):

> "Um sistema de reomendação é um sistema de informação que auxilia o usuário a recuperar informação através da previsão de seus interesses, informando-lhe conteúdo, fontes de consulta ou outras informações. " (MARQUES, 2007, p. 23)

Estrutura 04: Estrutura mais simples, pois não é possível mapear o autor com facilidade, pórem representa 4% dos contextos identificados e contêm uma Expressão linguística, o Termo, uma segunda Expressão linguística e a Definição. Denominaremos esta estrutura de ETED.

Expressão Linguística 01	TERMO	Expressão Linguística 02	DEFINIÇÃO
utiliza um conceito de		que...:	
É usual a visão de que		constitui	
a concepção de		compreende	
uma das definições apresentada para		é	
conceitos como		:	
trata do		como	
todos entendem que		é um	
entende-se		como	
define-se	TERMO	como	DEFINIÇÃO
É considerado		"	
considera-se		a	
depreende-se por		uma	
o objetivo do		é	
a finalidade principal		é	
o objetivo fundamental da		é	
o objetivo próprio de		é	
o objetivo da		é	
o objetivo de		é	

Figura 25: Estruturas linguísticas, ETED, encontradas nos documentos analisados
Fonte: Produzido pelo autor.

Exemplo de estrutura ETED:

Na pesquisa de Marques (2007):

> "O objetivo da recuperação, dada uma pergunta formalizada por descritores que a definem corretamente, é que o sistema de informação providencie a comparação desses com aqueles que descrevem o documento e obtenha as referências bibliográficas que atendem à pergunta em questão (ROBREDO, 2005). " (MARQUES, 2007, p. 24)

Estruturas mais complexas também foram identificas, porém seu mapeamento é mais complicado.

Em Marciano (2006):

> "A interconexão entre a Fenomenologia e a Ciência da Informação mostra-se ainda mais evidente quando se observa que a primeira conceitua a linguagem como origem e expressão do conhecimento, ao passo que a última situa o documento, sua principal fonte de estudo, como veículo do conhecimento codificado e formalizado por meio da linguagem." (MARCIANO, 2006, p. 36)

A tabela 5, a seguir, resume a percentagem identificada das estruturas definitórias no grupo amostral analisado manualmente, por tipo proposto.

Tabela 5: Percentagem das Estruturas identificadas na Amostra

Estrutura definitória	Presença na amostra (%)
EATED	12%
AETED	14%
TED	62%
ETED	04%
Estruturas Complexas	08%

8.3 Primeira versão da gramática

A primeira versão da gramática de padrões definitórios em língua portuguesa para os textos da Faculdade foi elaborada a partir das expressões linguísticas identificadas na amostra e da tradução para o português da gramática proposta por Sierra e Alarcón (2003) em língua espanhola, citada no item 7.3 da revisão de literatura.

A gramática proposta por Sierra e Alarcon é composta de verbos e estes estão vinculados a sua classificação de tipos de definições: analítica, sinonímica, funcional e extensional.

Porém, com a finalidade de aumentar a cobertura da gramática, optou-se por utilizar também as expressões linguísticas encontradas no trabalho de Kamikawachi (2009) em língua portuguesa. Kamiquawachi identificou inúmeras expressões nas definições analisadas em seu estudo, expressões que estão relatadas em seu trabalho conforme apareceram nas definições Kamikawachi (2009, p.89) e divididas em todas as classificações semânticas relatadas em sua pesquisa. Seu trabalho, entretanto, analisou um córpus de definições já previamente selecionadas, não encontradas, por tanto, em textos de especialistas. O foco do trabalho desta pesquisadora

foi separar as definições e classificar as relações semânticas existentes, difere desta pesquisa que pretende identificar os contextos definitórios de forma automática, sem detalhar todos os tipos possíveis de relação. Entretanto, as expressões linguísticas podem dar sinais importantes para identificar a presença do contextos definitório.

Como nesta pesquisa foi adotada a estrutura Qualia, proposta por Pustejovsky (1991), apenas as expressões linguísticas que compunham os tipos constitutivo (apresenta-se como, caracterizado pela, constituído de, é formado por, etc), télico (atua como, é empregado como, muito utilizado como, provoca a, etc) e agentivo (causado por, é originada pela,obtido por,etc) foram aproveitadas. O tipo Formal não foi estudado por Kamikawachi, sendo usado apenas os verbos da gramática de Sierra e Alarcón (2003) e as expressões identificadas na análise manual como base para este tipo.

Para se ter uma convergência entre as duas propostas, foi criada a tabela 6, que identifica a relação dos dois tipos de classificação.

Tabela 6: Relação da estrutura Qualia com as categorias de CDs

Qualia de Pustejovsky (1991)	Descrição da categoria	Categoria de Sierra e Alarcón (2003)
Constitutivo	Objeto e seus componentes	Extensional
Télico	Função do objeto	Funcional
Télico	Finalidade do objeto	—-
Formal	Hiperonímia	—-
Formal	Sinonímia	Sinonímia
Agentivo	origem do objeto	—-
—-	Gênero próximo e diferença específica	Analítica

O enquadramento dos verbos ou expressões pode ser feito através desta tabela, ou seja, todos os elementos linguísticos encontrados podem ser associados a esses dois tipos de classificação, e consequentemente, os contextos definitórios serão também enquadrados nas duas classificações.

Vale destacar que o tipo de Contexto definitório analítico indica que existe a presença do gênero próximo e da diferença específica, mas não detalha o tipo de diferença, podendo, por este motivo, ser enquadrada em qualquer classificação da estrutura Qualia. Contudo, como o objetivo da pesquisa é identificar os contextos definitórios, expressões linguísticas com indicativo claro de definição (define como, é definido como, etc) fazem parte da gramática, mas o contexto definitório não tem um detalhamento do tipo de diferença específica e classificação Qualia.

A primeira versão da gramática foi composta, então, pelos três grupos de elementos linguísticos: a gramática de Sierra e Alarcón (2003), incorporada das expressões de Kamikawachi (2009) para os grupos da estrutura Qualia, conforme tabela 4, e as expressões encontradas na analise manual, que não faziam parte das relacionadas por Kamikawachi.

8.4 Análise da extração automática com a manual

8.4.1 Execução da ferramenta e análise do primeiro grupo

A gramática proposta foi então associada à uma ferramenta, para identificação de contextos definitórios, construída no âmbito desta pesquisa. Esta ferramenta propicia a inclusão de padrões sintáticos que servem de expressões de busca no texto analisado e recuperam ou sinalizam os contextos que possuem os elementos solicitados.

A ferramenta foi construída em módulos, onde cada estrutura identificada na análise manual e citada na tabela 5 foi tratada separadamente. A estrutura EATED, por exemplo, foi mapeada na ferramenta seguindo sua estrutura: Expressão linguística, complemento textual que consideramos como o autor, podendo ter uma vírgula ou não, um segundo complemento textual, que é o termo, seguido de outra expressão linguística e por fim, a definição.

Após executar a ferramenta, configurada para a estrutura EATED, nos documentos analisados manualmente, identificamos automaticamente contextos definitórios, contudo o número de CDs identificados foi diferente. Enquanto a forma manual selecionou 71 CDs nesta estrutura, a forma automática identificou 97, sendo 42 válidos ou iguais aos encontrados de forma manual. Este processo ocorreu para todas as estruturas, exeto as definições com estruturas complexas que não foram investigadas neste estudo.

O cálculo da Precisão e da Revocação foram feitos baseado nas fórmulas citadas na seção 4.2 da revisão de literatura e permitem uma boa avaliação do método. Como assinalamos os contextos definitórios de maneira manual, podemos considerar que eles são os ítens relevantes e os erros são aqueles identificados pela ferramenta, mas que não são válidos ou não foram identificados manualmente. O resultado para todas as estruturas segue na tabela 7.

Tabela 7: Comparação número de Contextos definitórios identificados de forma automática x forma manual

Estrutura	CDs identificados manualmente	CDs identificados automaticamente	CDs válidos automaticamente	%Revocação	%Precisão
EATED	78	97	42	54%	43%
AETED	73	101	61	83%	60%
ETED	21	30	18	86%	60%
TED	320	400	298	93%	74%
TOTAL	492	573	459	85%	67%

A revocação, que mede dos CDs válidos, quantos foram recuperados, teve um índice muito bom, com 85% no total, tendo o pior desempenho para a estrutura EATED com 54%. Já a precisão teve um resultado apenas regular com 67% no total. Vale destacar a estrutura TED, que têm uma revocação muito alta 93%, mas também possui uma precisão bastante elevada 74%.

A figura 26 é um exemplo de comparação entre a marcação manual e a identificação automática da ferramenta, que assinala cada estrutura com uma cor especificada anteriormente.

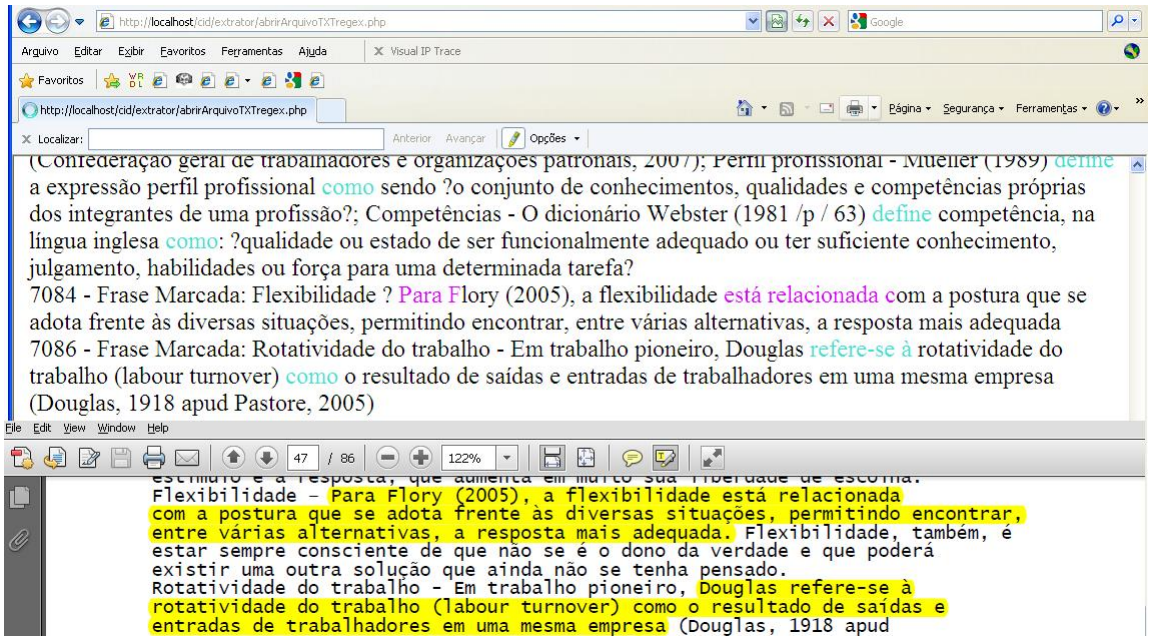

Figura 26: Telas dos documentos marcados de forma manual e automática
Fonte: Produzido pelo autor

8.4.2 Adequação da gramática

A partir destes números se realizou uma investigação nos CDs inválidos recuperados e nos CDs não encontrados para analisar o motivo para a divergência. Percebeu-se que ajustes para algumas expressões linguísticas poderiam ser feitos e o resultado melhorou significativamente. Um dos ajustes efetuados ocorreu na primeira expressão linguística da estrutura EATED que ao colocar "para"ou "segundo"em minúsculo recuperava mais erros do que acertos, optou-se por colocar a primeira letra em maiúscula, ou seja, "Para"e "Segundo", que identifica o inicio de uma frase. Com os ajustes o desempenho aumentou ainda mais, como demonstra a tabela 9.

Tabela 8: Comparação Número de Contextos definitórios identificados de forma automática x forma manual

Estrutura	CDs identificados manualmente	CDs identificados automaticamente	CDs válidos automaticamente	%Revocação	%Precisão
EATED	78	90	75	96%	83%
AETED	73	101	61	83%	60%
ETED	21	30	18	86%	60%
TED	320	400	298	93%	73%
TOTAL	492	621	452	92%	73%

Na segunda rodada do grupo 01 o melhor desempenho na revocação passou a ser da estrutura EATED, que passou a recuperar 96% dos contextos definitórios válidos. Essa melhora, elevou a revocação geral para 92% , índice espetacular, contudo é preciso observar o desempenho da gramática em outros documentos, o que foi feito na próxima seção.

8.4.3 Execução da ferramenta e análise do segundo grupo

Após a primeira rodada de comparação entre os contextos definitórios identificados manualmente com os assinalados pela ferramenta de forma automática no primeiro grupo de documentos, partiu-se para o caminho inverso no segundo grupo da amostra.

Os documentos do segundo grupo, tabela 9, foram analisados primeiro pela ferramenta e depois foi feita a análise manual dos mesmos documentos. Os resultados são descritos na tabela 10.

Tabela 9: Relação dos documentos do segundo grupo analisados.

ANO	AUTOR	TíTULO
2004	João Batista Simao	Universalização de serviços públicos na internet para o exercício da cidadania: Análise crítica das ações do Governo Federal
2006	Márcia Loureiro Paulo	Monitoramento Informacional nos curtumes do Estado de Mato Grosso do Sul
2008	Renilda Goncalves Amaral	A função da biblioteca pública escolar no contexto da formação integral do educando: estudo de caso
2010	Cleone Silvestre Neto	Estudo de necessidades de informação dos produtores de hortaliças orgânicas não certificados do Distrito Federal
2010	Katiucia Goncalves Amaral	Modelos de negócios para periódicos científicos eletrônicos de acesso aberto

Fonte: Produzida pelo autor.

Tabela 10: Comparação Número de Contextos definitórios identificados de forma automática x forma manual- Segundo grupo de documentos

Estrutura	CDs identificados manualmente	CDs identificados automaticamente	CDs válidos automaticamente	%Revocação	%Precisão
EATED	65	57	42	65%	73%
AETED	39	34	23	59%	68%
ETED	11	10	6	55%	60%
TED	397	403	351	88%	87%
TOTAL	512	504	422	82%	84%

O resultado ficou muito próximo da primeira rodada do grupo 01 para a revocação geral,

83%. Porém a precisão geral foi maior, alcançando 84%. Após analisar este resultado, novamente se realizou uma investigação nos erros e nos CDs não identificados. Podemos destacar ainda um decréscimo das 3 primeiras estruturas com relação a percentagem alcançada na primeira rodada do grupo 01. Ao analisar os CDs não identificados de forma automática, verificou-se que novas expressões apareceram e que não tinham sido mapeadas. Essas novas expressões linguísticas foram acrescidas a ferramenta o que proporcionou um acréscimo de 10% para a estrutura EATED e 15% para a estrutura AETED.

Com esse resultado, considerou-se que a gramática de padrões definitórios para os documentos da Faculdade da Ciência da Informação estava criada e pronta para ser aplicada nos documentos armazenados em seu repositório instituicional.

8.5 Extração automática de Contextos definitório na Base da Faculdade da Ciência da Informação

Para a extração dos Contextos Definitórios nos documentos da faculdade, a gramática gerada foi incorporada á ferramenta, que também foi acrescida de um novo módulo. Este módulo, ao percorrer o documento, extrai os contextos definitórios e os armazena em um arquivo texto separado, salvo em um outro diretório, porém com o mesmo nome do documento analisado, acrescido de "CDS"à frente. Contabiliza ainda, em uma tabela de banco de dados, o total de CDs extraídos por estrutura mapeada, separando também, as expressões linguísticas identificadas.

Não foi possível a verificação de 12 documentos que foram salvos em um formato diferente do padrão normal, o que impossibilitou o seu mapeamento através da ferramenta. Uma investigação mais detalhada será necessária para resolver estes casos, contudo para fins estatísticos deste estudo, esses documentos não foram contabilizados.

Para a contabilização dos resultados os documentos foram separados por nível de titulação, ou seja, dissertações e teses foram separados para geração das estatísticas. Os documentos analisados contabilizaram os seguintes resultados:

A estrutura TED aparece como a que identifica o maior número de contextos 78,5%. Por conter apenas uma expressão linguística, esta estrutura é mais livre porém como verificado na analise manual da seção 8.4, possui também uma precisão muito alta. Os resultados para o grupo de Teses, segue na tabela 11.

O grupo das Teses identificou uma quantidade grande de CDs com a estrutura TED, em

Tabela 11: Média de Contextos definitórios identificados na base - Dissertações

Estrutura	média por documento	% do total
EATED	13,4	11%
AETED	10	8%
ETED	3,2	2,5%
TED	96,4	78,5%
Média de CDs por Documento	123	100%

Tabela 12: Média de Contextos definitórios identificados na base - Teses

Estrutura	média por documento	% do total
EATED	12,8	7%
AETED	17,8	10%
ETED	5,5	3%
TED	145	80%
Média de CDs por Documento	180	100%

média 145, porém, alguns documentos saíram completamente do padrão. A ferramenta chegou a identificar 270 CDs com essa estrutura em apenas um documento.

Com relação às expressões linguísticas identificadas, criamos uma lista com as 5 expressões que mais apareceram para cada estrutura.

Tabela 13: TOP 05 - Estrutura EATED - Expressões identificadas

Expressão 01	Expressão 02
Para	é/são
Segundo	é/são
Para	apresenta
Para	representa
Segundo	apresenta

Na estrutura EATED, figura 13, a combinação que mais apareceu foi a expressão linguística "Para"+ autor + termo e a outra expressão "é/são", finalizando com a definição.

Já na estrutura AETED, tabela 14, a combinação Autor + afirma/sugere que + Termo + é foi a que mais se destacou. Porém, o Autor(es) + define/consideram/refere-se à + Termo + como, também apareceu bastante.

Segundo a tabela 15, na estrutura ETED, a composição objetivo de + Termo + é, foi a

Tabela 14: TOP 5 - Estrutura AETED - Expressões identificadas

Expressão 01	Expressão 02
afirma	é
define	como
sugere que	é
consideram	como
refere-se à	como

campeã de ocorrências.

Tabela 15: TOP 5 - Estrutura ETED - Expressões identificadas

Expressão 01	Expressão 02
objetivo de	é
entende-se	como
considera-se	um
define-se	como
a concepção de	como

Porém, o grande destaque, em termos de ocorrências, foi a expressão é, seguido de um, a ou o. Esta expressão teve uma média de quase 60 ocorrências por documento analisado.

Tabela 16: TOP 10 - Estrutura TED - Expressões identificadas

Expressão 01	Média por documento
é um/a/o	59,3625
, ou seja,	12,45
, que é,	5,025
, isto é,	4,4625
tais como:	1,725
identifica	1,4625
se refere a	1,2375
é considerado	0,9
para designar	0,5625
funciona como	0,4125

O grande número de CDs identificados nos documentos, tabela 11 e 12, comprovam que nos textos acadêmicos, direcionados para principiantes ou profissionais da mesma área, são encontrados um grande número de definições, o que permite um mapeamento semântico da área em estudo. A seguir faremos as considerações finais do estudo.

9 Considerações finais

Os objetivos propostos para este trabalho, conforme Seção 1, são analisados a seguir.

– Objetivo 1: Construir uma gramática de padrões definitórios para textos da Ciência da Informação em língua portuguesa a partir dos trabalhos de Sierra e Alarcón (2003) e Kamikawachi (2009). O percurso metodológico utilizado permitiu uma investigação apurada de documentos escritos em língua portuguesa, com as características da formalidade de um texto científico, de uma única área do conhecimento e com um número significativo de contextos definitórios descritos. Percebeu-se que existe um padrão de escrita nesses textos, que pode ser mapeado. Os estudos de Sierra e Kamikawachi auxiliaram na composição deste mapeamento. A gramática em língua portuguesa criada é o principal resultado alcançado com o estudo e poderá servir de base para extração de termos e definições como demostrado neste estudo. Essa gramática poderá, também, auxiliar na construção automática de tesauros, ontologias e bases terminológicas, uma vez que possibilita a identificação automática de conceitos e termos.

– Objetivo 2: Validar a gramática proposta através da comparação dos contextos definitórios extraídos de forma automática com grupo de contextos identificados de forma manual. Este objetivo permitiu a validação da gramática ao apontar os índices considerados significativos de precisão e revocação, conforme Seção 7.4. Outro resultado relevante foi a criação uma base de CDs não encontrados automaticamente e de estruturas identificadas automaticamente, mas que não são consideradas um contexto definitório. Esta base serviu de instrumento de compreensão dos problemas de um mapeamento sintático e pode servir de base para investigações futuras.

– Objetivo 3: Identificar de forma automática os contextos definitórios (CDs) nas teses e dissertações da Faculdade de Ciência da Informação da Universidade de Brasília - UNB, contidas no RIUnB. Neste aspecto, a pesquisa possibilitou a criação de uma base de Contextos Definitórios, que apesar de não possuir todos os ítens extraí-

dos validados, os índices de precisão da gramática apresentados na Seção 7.4, nos levam a acreditar que grande parte dos CDs extraídos representam um mapeamento semântico significativo dos textos de dissertações e teses da Faculdade de Ciência da Informação da UNB.

Além dos objetivos atingidos, podemos destacar também, a ferramenta construída no âmbito da pesquisa, que poderá ser utilizada em novas investigações em Descoberta de Conhecimento em Textos e Extração da Informação, em especial em estudos relacionados a Contextos Definitórios. O método criado necessita primordialmente de uma estrutura computacional consistente para uma efetiva extração, o que foi validado a partir dos resultados identificados.

Contudo, não podemos deixar de comentar as limitações de uma gramática de padrões definitórios. A riqueza da língua portuguesa, impossibilita um mapeamento completo das possibilidades de se definir ou caracterizar um termo. A gramática precisa ser dinâmica e novas expressões devem ser incluídas para obter um aumento nos índices de precisão do método de extração, principalmente quando aplicada em outros contextos ou em áreas diferentes.

Concluindo, a pesquisa possui um conhecimento na análise de textos em língua portuguesa de forma automática, em especial com relação ás estruturas definitórias, que pode auxiliar em diversas pesquisas futuras nas área de Processamento de Linguagem Natural e Descoberta de Conhecimento em Textos que detalharemos na seção seguinte.

9.1 Possibilidades futuras de pesquisa

Vários trabalhos futuros podem ser indicados, dentre os quais podem ser sugeridos:

- A aplicação da gramática de padrões definitórios em um outro contexto ou em outro tipo de documento, como os artigos, por exemplo, ou em uma outra área do conhecimento, a avaliação de seu desempenho, o acréscimo de novas expressões linguísticas, seriam ações interessantes para o aprimoramento da gramática. Outro ponto importante é a investigação das estruturas consideradas complexas e não mapeadas no âmbito desta pesquisa. É possível que outras estruturas possam ser identificadas e, apesar da ocorrência nos documentos analisados ter sido baixa, é mais um ponto de identificação e possível extração de CDs de forma automática. Ademais, os padrões tipográficos também não foram estudados e, não obstate a complexidade atual de utilização destes elementos em processamento textual, também é um aspecto para investigações futuras.

– Com relação à ferramenta, precisam ser desenvolvidos módulos gráficos de inclusão de arquivos, de novas estruturas e expressões, além de uma página para navegação nos Contextos Definitórios identificados, com operações de edição e exclusão. É possível, ainda, uma otimização do código para acelerar o processamento dos textos. Esses módulos e ajustes serão desenvolvidos e a ferramenta será patenteada.

– A base de Contextos Definitórios extraídos da Faculdade de Ciência da Informação da Universidade de Brasília - UNB, também é um campo fértil de estudos. A validação dos CDs identificados, o estudo aprofundado dos erros e o mapeamento de exeções pode melhorar o índice de Revocação da ferramenta; Estudar os CDs válidos, separando os termos, a função semântica das definições e o estudo dos elementos pragmáticos também são possíveis pesquisas futuras.

Índice Remissivo

Referências Bibliográficas

AGUILAR, C. *Análisis linguistico de definiciones em contextos definitorios*. Tese (Doutorado em Linguística) — Universidad Nacional Autónoma de México, 2009.

AGUILAR, C. et al. Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados. In: CABRÉ M. T., E. R. y. T. C. (Ed.). *La terminologia en el siglo XXI*. Barcelona, 2004. p. 259–269.

ALARCON, A. M. D. T. B. F. E. Sistema de classificação facetada e tesauros: instrumentos para organização do conhecimento. *Ciência da Informação*, v. 33, n. 2, p. 161–171, maio/agosto 2004.

ALARCON, R. *Anális linguístico de contextos definitorios em textos de especialidad*. 2003.

ALARCÓN, R. *Descripción y evaluación de um sistema basado em reglas para La extracción automática de contextos definitorios*. Tese (Doutorado em Linguística) — Universidad Pompeu Fabra - Barcelona, 2009.

ALMEIDA, G. M. B.; ALUÍSIO, S. M.; OLIVEIRA, L. O método em terminologia: revendo alguns procedimentos. In: ISQUERDO APARECIDA NEGRI; ALVES, I. M. O. (Ed.). *Ciências do léxico: lexicologia, lexicografia, terminologia.* [S.l.]: Editora da UFMS/Humanitas, 2007. III.

ALVARENGA, L. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. *Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.*, n. 15, 1 sem. 2003.

ARAÚJO JR., R. H. d. *Precisão no processo de busca e recuperação da informação.* Brasília: Thesaurus, 2007.

ARAÚJO, W. J. *A sergurança do conhecimento nas práticas da gestão da segurança da informação e da gestão do conhecimento*. Tese (Doutorado em Ciência da Informação) — Faculdade de Ciência da Informação da Universidade de Brasília - UNB, 2009.

AUGER, A. *Repérage des énoncés d'intérêt définitoire dans les bases de données textualles*. Dissertação (Doutorado em Letras) — Faculté des Lettes, Université de Neuchâtel, 1997.

BONFANTE, A. G. *Parsing Probabilístico para o Português do Brasil*. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo, 2003.

BRÄSCHER, M. A ambiguidade na recuperação da informação. *DataGramaZero - Revista da Ciência da Informação*, v. 3, n. 1, Fev 2002.

BRÄSCHER, M.; CAFÉ;, L. Organização da informação ou organização do conhecimento? In: LARA M.L.G. DE; SMIT, J. O. (Ed.). *Temas de pesquisa em Ciência da Informação no Brasil*. São Paulo: Escola de Comunicação e Artes da USP, 2010. p. 85–103.

BUSH, V. As we may think. the atlantic on-line. *The Atlantic Monthly*, v. 12, n. 1, p. 101–108, julho 1945. Disponível em: <http://www.theatlantic.com/unbound/flashbks/computer/bushf-.htm>. Acesso em: 06.06.2010.

CABRÉ, M. T. La terminología hoy: concepciones, tendencias y aplicaciones. *Ciência da Informação*, v. 24, n. 3, 1995.

CABRÉ, M. T. Teorias da terminologia. *John Benjamins Publishing Company*, 2003.

CAFÉ, L. et al. Repositórios institucionais: nova estratégia para publicação científica na rede. In: *Congresso Brasileiro de Ciências da Comunicação*. Belo Horizonte: Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, 2003.

CÂMARA JR., A. T. *Indexação Automática de Acórdãos por meio de Processamento de Linguagem Natural*. Dissertação (Mestrado em Ciência da Informação) — Departamento de Ciência da Informação/Universidade Federal de Santa Catarina, 2007.

CAMPOS, G. N. *Características e perfil dos bibliotecários das bibliotecas de instituições de ensino superior privadas do Distrito Federal e as expectativas dos empregadores*. Dissertação (Mestrado em Ciência da Informação) — Faculdade de Ciência da Informação da Universidade de Brasília - UNB, 2008.

CAMPOS, M. L. de A. *A Organização de Unidades do Conhecimento em Hiperdocumentos: o modelo conceitual como um espaço comunicacional*. Tese (Doutorado em Ciência da Informação) — Programa de Pós-Graduação em Ciência da Informação, do convênico CNPq/IBICT - UFRJ/ECO, 2001.

CAPUANO, E. A. *Mineração e modelagem de conceitos como praxis de gestão do conhecimento para inteligência competitiva*. Tese (Doutorado em Ciência da Informação) — Faculdade de Ciência da Informação da Universidade de Brasília - UNB, 2010.

CAPURRO R.; HJORLAND, B. O conceito de informação. *Ciência da Informação*, v. 12, n. 1, p. 148–207, jan./abr. 2007.

CASELI, H. *Indução de léxicos bilíngues e regras para tradução automática*. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo, 2007.

CLARK, D. *http://www.nwlink.com/ donclark/performance/understanding.html*.

COCHRAN, W. G. The estimation of sample size. In: *Sampling techniques*. 3. ed. New York: John Wiley, 1977. p. 72–90.

CONSTANTINO, M. *Financial Information Extraction using pre-dened and user-denable*. Tese (Doutorado em Ciência da Computação) — University of Durham, 1997.

CUADRA, C. A. et al. Experimental studies of relevance judgments: Final report. *System Development Corporation*, v. 1-3, 1967.

DAHLBERG, I. O futuro das linguagens de indexação. In: *Conferência Brasileira de Classificação Bibliográfica*. [S.l.: s.n.], 1972a.

DAHLBERG, I. Teoria da classificação ontem e hoje. In: *Conferência Brasileira de Classificação Bibliográfica*. [S.l.: s.n.], 1972b.

DAHLBERG, I. A referent-oriented analytical concept theory of interconcept. *International Classification*, v. 5, n. 3, p. 142–150, 1978.

DAHLBERG, I. Teoria do conceito. *Revista da Ciência da Informação*, v. 7, n. 2, p. 101–107, 1978.

De Bessé, B. Le contexte terminographique. *Meta: Journal des traducteurs/ Meta: Translators' Journal*, v. 36, n. 1, p. 111–120, mars 1995.

De Bessé, B. Notes de cours. In: ____. Genève: École de traduction et d'interprétation, 1996. cap. Chapitre 2: Aspects cognitifs, p. 41–67.

DIAS, G. M. et al. Introdução ao processamento das linguas naturais e algumas aplicações. *NILC-TR-07-10*, Agosto 2007.

FIGUEIREDO, L. M. O conceito de relevância e suas implicações. *Ciência da Informação*, v. 6, n. 2, p. 75–78, 1977.

FLOWERDEW, J. *Definitions in Science Lectures*. [S.l.]: Applied Linguistics, 1992.

FRANCELIN, M. M. *Ordem dos conceitos na Organização da Informação e do Conhecimento*. Tese (Doutorado em Ciência da Informação) — Programa de Pós-Graduação em Ciência da Informação da Escola de Comunicações e Artes (ECA) da Universidade de São Paulo (USP), 2010.

FREI, H. P. Information retrieval. *Academic research to pratical applications*, 1996.

GIL, A. C. *Métodos e técnicas de pesquisa social*. São Paulo: Editora Atlas, 1999.

GOMES, H. E.; CAMPOS, M. L. de A. Tesauro e normalização terminológica: o termo como base para intercâmbio de informações. *DataGramaZero - Revista da Ciência da Informação*, 2004.

GRACIO, M. C. C.; OLIVEIRA, E. F. T. Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de ciência da informação. *DataGramaZero - Revista da Ciência da Informação*, v. 6, n. 3, jun 2005.

GREGHI, J. G. *Projeto e desenvolvimento de uma base de dados lexicais do português*. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulos, 2002.

GUARINO, N. Formal ontology and informaton systems. *Proceedings of the 1st International Conference*, IOS Press, Trento, Italy, p. 3–15, June 1998.

HEY, J. The data, information, knowledge, wisdom chain. *The Metaphorical Link*, 2004.

KAMIKAWACHI, D. S. L. *Aspectos semântico da definição terminológica (DT): Descrição linguística e proposta de sistematização*. Dissertação (Mestrado em Linguística) — Universidade Federal de São Carlos - UFSCar, 2009.

KOBASHI, N. Y. Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. *DataGramaZero - Revista da Ciência da Informação*, v. 8, n. 6, 2007.

LARA, M. L. G. de. Diferenças conceituais sobre termos e definições e implicações na organização da linguagem documentária. *Revista da Ciência da Informação*, v. 33, n. 2, p. 91–96, maio/ago 2004.

LARIVIÈRE, L. Comment formuler une définition terminologique. *journal des traducteurs*, v. 41, n. 3, p. 405–418, 1996.

LOH, S. *Descoberta de Conhecimento em Texto*. Dissertação (Mestrado em Ciência da Computação) — Instituto de Informática, Universidade Federal do Rio Grande do Sul - UFRGS, 1999.

MACIEL, A. M. B. *Para o reconhecimento da especificidade do termo jurídico*. Tese (Doutorado) — Programa de Pós-Graduação em Letras - Universidade Federal do Rio Grande do Sul, 2001.

MAIA, L. C. G. *Uso de Sintagmas nominais na classificação automática de documentos eletrônicos*. Tese (Doutorado em Ciência da Informação) — Universidade Federal de Minas Gerais, 2008.

MARCIANO, J. L. P. *Segurança da informação: uma abordagem social*. Dissertação (Doutorado em Ciência da Informação) — Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2006.

MARQUES, T. M. *Abordagens de recomendação para a recuperação de perfis: uma proposta de modelo.* Dissertação (Mestrado em Ciência da Informação) — Faculdade de Ciência da Informação da Universidade de Brasília - UNB, 2007.

MARSHMAN, E. *The cause relation in biopharmaceutical corpora: English and French patterns for knowledge extraction*. Tese (Doutorado) — School of Translation and Interpretation - University of Ottawa, 2003.

MENDONÇA, E. S. A linguistica e a ciência da informação: estudos de uma interseção. *Revista da Ciência da Informação*, v. 29, n. 3, p. 50–70, set/dez 2000.

MEYER, I. Extracting knowledge-rich contexts for terminography. In: BOURIGAULT, D. (Ed.). *Recent advances in computational terminology*. Amsterdan: John Benjamins Publishing Company, 2001. cap. 14, p. 279–302.

NUNES, T. A. S. P. H. de Medeiros Caseli; Maria das G. V. Mapeamento da comunidade brasileira de processamento de línguas naturais. Relatório de pesquisa. 2009.

OLIVEIRA, L. H. M. *e-Termos: Um ambiente colaborativo web de gestão terminológica*. Tese (Doutorado em Ciência da Computação) — Instituto de Ciências Matemáticas e de Computação da Universidade de São Carlos - USP, Agosto 2009.

PEARSON, J. *Terms in Context*. [S.l.]: John Benjamins Publishing Company, 1998.

PINTO, A. S.; OLIVEIRA, D. *Extracção de Definições no Corpógrafo*. [S.l.], outubro 2004.

POPPER, k. R. *Objective Knowldege. An evolutionary approach*. 2ª. ed. New York: Oxford University Press, 1972.

PRESCOTT, L. *Ranganathan and Facet Analysis*. 2003.

PUSTEJOVSKY, J. The generative lexicon. *Computacional Linguistics*, v. 17, n. 4, 1991.

RANGANATHAN, S.; GOPINATH, M. A. *Prolegomena to library classification*. 3. ed. [S.l.]: London: Asia Publishing House, 1967.

REY, A. L'impossible définition. In: ____. Paris: Le lexique images et modèles: du dictionnaire à la lexicologie, 1977. p. 98–113.

ROBREDO, J. *Da Ciência da Informação Revisitada aos Sistemas Humanos de Informação.* [S.l.]: Thesaurus, 2003.

RODRIGUEZ, C. *Operaciones Metalinguísticas e Explícitas en Textos de especialidad*. Barcelona, 1999.

RODRIGUEZ, C. *Metalinguistic Information Extraction from Specialized Texts to Enrich Computational Lexicons*. Tese (Doutorado) — Departament de traducció i filologia - Universita Pompeu Fabra, Barcelona, 2004.

SAGER, J. C. *Curso práctico sobre ele procesamiento de la terminología*. Madrid: Fundación Germán Sánchez Ruipérez, 1993.

SANCHEZ, A. Definicion e historia de los corpus. In: ____. Madrid: Corpus Linguisticos de Espanol Contemporaneo, 1995.

SANTOS, D. *Introdução ao Processamento de Linguagem Natural através das Aplicações.* [S.l.]: Tratamento das Línguas por Computador – Uma Introdução à Lingüística Computacional e suas Aplicações., 2001.

SARACEVIC, T. Relevance a review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, v. 39, n. 4, p. 235–351, October 1975.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*, v. 1, n. 1, p. 41–62, jan./jul. 1996.

SARACEVIC, T. *Relevance Reconsidered*. [S.l.]: Information science: Integration in Perspectives, 1996b. (COLIS 2, 201-208).

SARACEVIC, T. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *Journal of the American Society for Information and Technology*, v. 3, n. 58, p. 1915–1933, 2007.

SARACEVIC, T. Information science. In: *Marcia J. Bates and Mary Niles Maack*. [S.l.]: Encyclopedia of Library and Information Science. New York & Francis, 2009. p. 2570–2586.

SARDINHA, T. B. Linguística de corpus: Histórico e problemática. *Delta - Documentação de Estudos em Linguística Teórica e Aplicada*, v. 16, n. 2, p. 323–367, 2000.

SCARINCI, R. G. *SES - Sistema de Extração Semântica de Informações*. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, 1997.

SCHIESSL, J. M. *Descoberta de Conhecimento em Texto aplicada a um sistema de atendimento ao consumidor.* Dissertação (Mestrado) — Departamento de Ciência da Informação/Universidade Federal de Santa Catarina, 2007.

SEPPÄLÄ, S. *Composition et formalisation conceptuelles de la définition terminographique*. Dissertação (Doutorado em Tratamento Informático Multilingue) — École de traduction et dínterprétation, Université de Genève, Genebra, 2004.

SHAMBER, L. Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST)*, v. 29, p. 3–48, 1994.

SHAMBER L.; EISENBERG, M. N. M. A re-examination of relevance toward a dynamic, situational definition. *Information Processing and Management*, v. 26, n. 6, p. 755–776, 1990.

SIERRA, G. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones linguísticos. *Linguamática*, v. 1, n. 2, p. 13–37, Dezembro 2009.

SIERRA, G.; ALARCÓN, R. El rol de las predicaciones verbales en la extracción automática de conceptos. *Estudios de Linguística Aplicada*, v. 38, p. 129–144, 2003.

SILVA, F. *Critérios de seleção de obras raras adotados em bilbiotecas do Distrito Federal*. Dissertação (Mestrado em Ciência da Informação) — Faculdade de Ciência da Informação da Universidade de Brasília - UNB, 2011.

SILVEIRA, E. *As marcas do movimento de Saussure na fundação da linguistica*. Tese (Doutorado em Linguística) — Programa de Pós-Graduação em Linguística. Universidade Estadual de Campinas, 2003.

SIQUEIRA, A. H. Sobre a natureza da tecnologia da informação. *Ciência da Informação*, v. 37, n. 1, p. 85–94, jan./abr. 2008.

SOUSA, R. F. D. Para entender a ciência da informação. In: ____. Salvador: Toutain, Lídia Maria Batista, EDUFBA, 2007. cap. Organização do Conhecimento, p. 103–124.

STUMPF, I. Avaliação pelos pares nas revistas de comunicação: visão dos editores, autores e avaliadores. *Perspectivas em Ciência da Informação*, v. 13, n. 1, p. 18–32, jan./abr. 2008.

TÁLAMO, M. F. G. M.; LENZI, L. A. F. Terminologia e documentação: a relação solidária das organizações do conhecimento e da informação da inovação tecnológica. *DataGramaZero - Revista da Ciência da Informação*, 2006.

TAYLOR A. G.; JOUDREY, D. N. *The organization of information*. 3rd. ed. [S.l.]: Westport, Conn.: Libraries Unlimited, 2009.

WANDERLEY, M. A. Linguagem documentária: Acesso à informação. *Revista da Ciência da Informação*, v. 2, n. 2, p. 175–217, 1973.

WEISS et al. *Text minig: Predictive Methods for Analyzing Unstructed Information*. New York: Springer Science and Business Media, LLC, 2005.

WIVES, L. *Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva*. Dissertação (Exame de qualificação) — Instituto de Informática - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

WIVES, L. *Utilizando conceitos como descritores de textos para o processo de conglomerados (clustering) de documentos*. Tese (Doutorado em Ciência da Computação) — Instituto de Informática - Universidade Federal do Rio Grande do Sul, 2004.

WIVES L.; LOH, S. Tecnologia de descoberta de conhecimento em informações textuais: ênfase em agrupamento de informações. *PPGC/Universidade Federal do Rio Grande do Sul*, 2000.

YATES, R. B.; NETO, B. R. *Modern Information Retrieval*. [S.l.]: ACM Press, 1999.

ZELENY, M. Management support systems: Toward integrated knowledge management. *Human Systems Management*, v. 7, p. 59–70, 1987.

ZINS, C. Classification schemes of information science: 28 scholars map the field. *Journal of The American Society for Information Science and Technology*, v. 58, n. 5, p. 645–672, february 2007.

ZINS, C. Conceptions of informationn science. *Journal of The American Society for Information Science and Technology*, v. 58, n. 3, p. 335–350, february 2007. Knowledge Mapping Research, 26 Hahaganah Street, Jerusalem 97852, Israel.

ZINS, C. Conceptual approaches for defining data, information and knowledge. *Journal of The American Society for Information Science and Technology*, v. 58, n. 4, p. 479–493, february 2007. Knowledge Mapping Research, 26 Hahaganah Stree, Jerusalem 97852, Israel.

Anexo 01

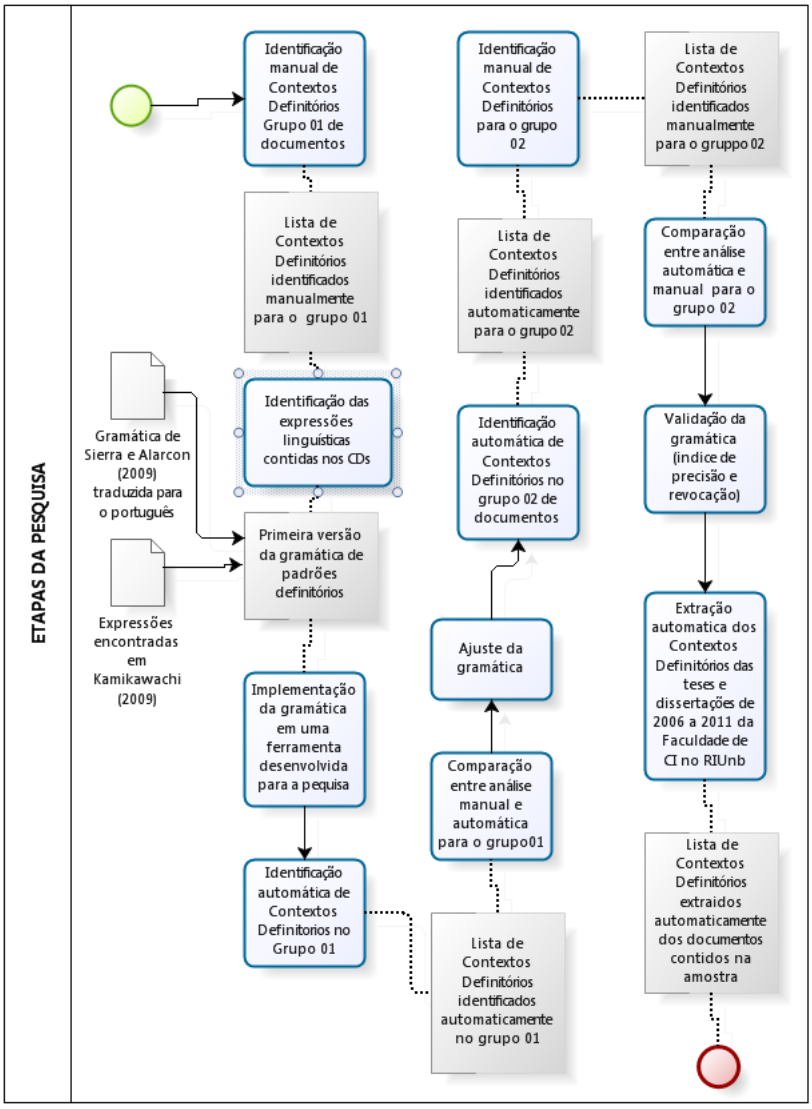

Figura 27: Etapas do estudo

Fonte: Produzida pelo autor

Anexo 02

Expressões linguísticas identificadas para o tipo télico.

Expressões linguística – TÉLICO
a fim de
a fim de que
a qual possibilita dar
Além da utilização como
Atua como
atuam como
com a finalidade de
com o fim de
com o objetivo de
com o propósito de
com objetivo de
conferindo à
cuja função é
cuja principal característica é propiciar
Cumpre função
Dentre os objetivos, encontram-se
destinado ao
É empregada
é empregada com o objetivo de
É empregada como
É empregado em
É empregado na
É geralmente empregada como
possibilita
É preferencialmente usado
É responsável por
é utilizada nas
é utilizada no
É utilizada para
é utilizada para
É utilizado em
É utilizado para
Empregado em
empregada na
Empregado na

Expressões linguística – TÉLICO
Empregado nas
Empregado nos
geralmente utlizada para
muito utilizada como
Muito utilizado como
Muito utilizado na
no qual se controla
Normalmente se utiliza
O objetivo é
Os objetivos são
Os principais efeitos
para evitar
Para fins
para obter
pode ser empregado como
pode ser utilizada na
pode ser utilizada para
Pode ser utilizado
possibilitando a
promove a
Propicia
propiciando
próprio para
provoca a
que avalia
que avalia a
que controla
que determina
que fornece
que identifica
que permite
que produz
Que quantifica
que registra
que se dedica a

Figura 28: Expressões identificadas no trabalho de kamiquawachi para o tipo de relação semântica télico

Fonte: (KAMIKAWACHI, 2009)

Expressões linguísticas identificadas para o tipo télico (continuação).

Expressões linguistica – TÉLICO
que serve para
que tem como finalidade
que tem como funcionalidade
que tem por finalidade
que têm por objetivo
que tem por objetivo
que tem por objetivo atingir
Que tem por objetivo aumentar
Que tem por objetivo avaliar
Que tem por objetivo avaliar
que tem por objetivo detectar
que tem por objetivo determinar
Que tem por objetivo identificar
Que tem por objetivo mensurar
que tem por objetivo mostrar
que tem por objetivo realizar
que tem por objetivo verificar
que visa
realizado para
Responsável por
São utilizadas
serve como
Sua aplicação favorece
Sua finalidade é
Tal recurso objetiva
Tem como finalidade
Tem como objetivo
Têm diversas aplicações
tem por finalidade
Tem por objetivo
Utilizada em
utilizada na
utilizada no
utilizado com a finalidade de
Utilizado em
utilizado junto à
Utilizado na
utilizado na
utilizado no
utilizado para
Utilizado para
Visa

Figura 29: Expressões identificadas no trabalho de kamiquawachi para o tipo de relação semântica télico - continuação

Fonte: (KAMIKAWACHI, 2009)

Expressões linguísticas identificadas para o tipo constitutivo

Expressões linguística – CONSTITUTIVO
Apresenta
Apresenta propriedades como
Apresenta-se como
Apresenta-se na
baseada na
Baseia-se fundamentalmente
caracterizada pela
caracterizada pelo
Caracterizada por
caracterizado pela
caracterizado pelo
Caracterizado por
composto de
composto essencialmente por
Consiste em
Consiste na
Consiste num
consiste numa
Constituída basicamente de
Constituída especialmente de

Expressões linguística – CONSTITUTIVO
Constituída essencialmente de
constituída essencialmente pelo
Constituída essencialmente por
Constituídas por
Constituído de
Constituído por
cuja aparência é
cujo constituinte principal é
É baseada na
é composta por
é composto, basicamente
é constituída de
é constituído por
É formado por
em que há
em que ocorre
em que são efetuados
em que se
em que se realiza
Engloba

Figura 30: Expressões identificadas no trabalho de kamiquawachi para o tipo de relação semântica constitutivo

Fonte: (KAMIKAWACHI, 2009)

Expressões linguísticas identificadas para o tipo constitutivo (continuação)

Expressões linguística – CONSTITUTIVO
essencialmente constituído por
feito de
Formada por
Formado por
na qual
na qual as
na qual um
no qual
no qual a
no qual o
no qual ocorre o
no qual os
por meio da qual se observa
Possui forma
que apresenta
Que apresentam características
que consiste
que consiste em
que consiste na
que consiste num
que consiste numa
que descreve
que envolve
que inclui
que indica
que possui
que se apresenta como
que se caracteriza pela
que se caracteriza pelo
que se caracteriza por
que se constitui em
relacionada à
relacionada ao
se baseia
Seu material pode ser
tem formato
Tem o formato de

Figura 31: Expressões identificadas no trabalho de kamiquawachi para o tipo de relação semântica Constitutivo - continuação

Fonte: (KAMIKAWACHI, 2009)