Universidade Federal de Santa Catarina
Centro Tecnológico
Programa de Pós-Graduação em Ciência da Computação

Cláudia Maksud Mechereffe

ESTRUTURA SINTR+: UM MODELO DE SUPORTE AO USUÁRIO NA RECUPERAÇÃO DE INFORMAÇÕES

Dissertação de Mestrado

Florianópolis
2005

Cláudia Maksud Mechereffe

ESTRUTURA SINTR+: UM MODELO DE SUPORTE AO USUÁRIO NA RECUPERAÇÃO DE INFORMAÇÕES

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação.
Profª Edla Maria Faust Ramos, Drª

Florianópolis
2005

Cláudia Maksud Mechereffe

ESTRUTURA SINTR+: UM MODELO DE SUPORTE AO USUÁRIO NA RECUPERAÇÃO DE INFORMAÇÕES

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____
Prof. Raul Sidnei Wazlawick, Dr.
Coordenador do PGCC

Banca Examinadora

_____
Profª Edla Maria Faust Ramos, Drª
Orientadora PGCC

_____
Prof. Heronides Maurílio de Melo Moura, Ph. D.

_____
Profª Maria Marta Leite, Drª.

_____
Prof. Raul Sidnei Wazlawick, Dr.

A alma é uma borboleta. Há na vida, um momento em que uma voz nos diz que chegou o momento de uma grande metamorfose: é preciso abandonar o que sempre fomos para nos tornarmos uma outra coisa.

Rubem Alves

AGRADECIMENTOS

O momento de agradecimento permite lembrar com gratidão de todas as pessoas que conviveram comigo e que foram importantes nesta etapa de construção de conhecimento.

Agradeço à Universidade Federal de Santa Catarina, ao Programa de Pós-Graduação em Ciência da Computação e a todos os professores que oportunizaram o aprendizado alcançado.

Em nome destes professores e pelo seu profissionalismo, à Edla Faust Ramos, pelas tão valiosas orientações, pela confiança, pela crença no meu trabalho e na minha pessoa.

Em especial, agradeço a minha mãe e ao meu pai, Heloisa e Antonio, que me ensinaram a lutar e persistir e pelo carinho, apoio e estímulo que sempre me deram.

Ao David, meu companheiro, pela compreensão, paciência, carinho e apoio.

Ao Paulo Bueno e à Leila Di Pietro pelo estímulo, paciência, pelas grandes contribuições e ajudas prestadas.

À Renata Brizzi, à Josiele Azevedo, à Danielle Hennings e à Adriana Santos pelo apoio e por suas contribuições.

Ao Carlos Eduardo Nascimento pelo apoio e incentivo prestado.

E, aos meus irmãos, Beatriz e Régis, por sempre acreditarem em mim.

E a todos os meus amigos por serem especiais em minha vida.

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

SIGLAS

RI: Recuperação de Informação

SRI: Sistemas de Recuperação de Informação

SN: Sintagma Nominal

LG: Léxico Gerativo

EQ: Estrutura de Qualia

SMART: System for the Manipulation and Retrieval of Text

SV: Sintagma Verbal

SEL: Léxico de Enumeração de Sentidos

PLC: Paradigma Léxico-Conceitual

XML: Extensible Markup Language

UML: Linguagem de Modelagem Unificada

UP: Processo Unificado

O.O: Orientado a Objetos

NG: N-Grama

TT: Termo-Termo

TR: Termo-Relacionamento

RT: Relacionamento-Termo

TR+: Termo-Relacionamento/Relacionamento-Termo

SINTR+ Sintagma Nominal com TR+

BD: Banco de Dados

RESUMO

Este trabalho tem como objetivo apresentar um novo modelo de sistema informatizado de suporte ao usuário no processo de recuperação de informações. A proposta consiste em apoio durante a definição da *query* de busca e baseia-se na identificação das possibilidades de sistematização e junção do modelo de Kuramoto com a estrutura de Gonzalez. Para a sua construção foi necessário analisar e sintetizar o modelo de suporte ao usuário de Kuramoto (baseado na determinação dos Sintagmas Nominais), a estrutura de *Qualia* do Léxico Gerativo de Pustejovsky e, termos e RLBs (relações lexicais binárias) do modelo TR+ de Gonzalez. O resultado que se espera alcançar é possibilitar a realização de uma interação que venha a proporcionar uma negociação adequada dos significados entre o usuário e a máquina, negociação essa que deve resultar em fator fundamental na melhoria da eficiência dos processos de busca. O modelo de Kuramoto, baseado em uma hierarquia de Sintagmas Nominais, suporta inicialmente essa interação. Com a definição da *query* de busca e da Estrutura de *Qualia* de Pustejovsky, implícita no modelo TR+ de Gonzalez, foi possível obter uma maior relevância dos documentos recuperados através de um cálculo de peso de descritores (termos e relacionamentos) evidentes nos documentos. As etapas gerais do modelo proposto são: a extração de Sintagmas Nominais e a sua hierarquização automática em níveis, o pré-processamento (*toquenização* e etiquetagem), o processo de nominalização e a captura de RLBs. Delineado preliminarmente o modelo partiu-se para as etapas de levantamento e análise de requisitos, representada pelos diagramas e pelas descrições dos casos de uso, chegando-se ao desenvolvimento do seu modelo conceitual que culminou a construção dos diagramas de classes e de seqüência para a aplicação proposta. Ao final conclui-se que a alternativa indicada neste trabalho, além de ser exeqüível, apresenta ganhos qualitativos nos resultados de uma busca em recuperação de informações e, também, quantitativos, no que se refere a um menor tempo na fase de indexação (rapidez) e um tamanho menor de arquivos de índice gerados (memória).

Palavras-chave: Recuperação de Informação; Sintagmas Nominais; Estrutura de *Qualia;* Termos e RLBs.

ABSTRACT

This work has the presentation of a new model of a support information system to the user in the process of information retrieval. The proposal consists in the support during the definition of a search query based on the identification of the possibilities of informatization and junction of a Kuramoto model along with the Gonzalez structure. For its construction, it was necessary to analyze and synthesize the support model to the Kuramoto user (base don the determination of Nominal Syntagm), the *Qualia* structure of the Lexical Semantics of Pustejovsky, and having the LBRs (lexical binary relations) of the Gonzalez TR+ model. The result we expect to reach is the possibility of actually performing an interaction that may result in an adequate negotiation of meanings between the user and the machine, knowing that this negotiation should result in a fundamental factor in order for the improvement on the efficiency of the search processes. The Kuramoto model, based on Nominal Syntagm hierarchy, initially supports this interaction. With the definition of the query search and the Pustejovsky Qualia structure, implicit in the TR+ Gonzalez model, it was possible to obtain a greater relevance of documents recovered through a calculus of weight of describers (terms and relationships) evident in the document. The general stages of the proposed model are: the extraction of Nominal Syntagm and their automatic placement into hierarchy, the pre-processing (tokening and labeling), the naming and capture of the LRBs. After the preliminary outlining of the model, we went on to the gathering of stages and requisite analysis, presented by diagrams and descriptions of the usage cases, finally reaching the development of a conceptual model that culminated in the construction of class diagrams and of a sequence for the proposed application. As we reach the end, we can conclude that the indicated alternative in this work, besides being executable, presents qualitative gains in the results of a search for the retrieval of information and also quantitative gains, when referring to a smaller amount of time spent in the index phase (speed) and a smaller amount of archives generated (memory).

Key-words: Retrieval of Information; Nominal Syntagm; *Qualia* Structure*;* Terms e LRBs.

SUMÁRIO

1. INTRODUÇÃO

O tema “Recuperação de Informação” (RI) é importante para diversas áreas, tais como Biblioteconomia, Lingüística, Ciência da Computação, entre outras. Segundo Baeza-Yates e Ribeiro-Neto (1999), na Ciência da Computação, esse tema diz respeito à recuperação de dados e à recuperação de informação, sendo ambos processos importantes e significativos para a área.

De acordo com os autores, os sistemas de recuperação de informação lidam com objetos lingüísticos (textos) e por isso herdam toda a problemática inerente ao tratamento da linguagem natural. Já a recuperação de dados está associada a sistemas gerenciadores de banco de dados (ou simplesmente banco de dados), que ao organizá-los já especificam, de forma bem definida, a sua estrutura e, por conseguinte, a sua semântica.

Um dos desafios na recuperação de informação, conforme Ferneda (2003), diz respeito a melhorar a relevância dos resultados de uma busca de maneira que o usuário possa encontrar todos os documentos que atendam às suas necessidades de informação. Em outras palavras, isto quer dizer que a busca será precisa se conseguir retornar e/ou listar somente documentos relacionados ao que o usuário expressou na definição da sua busca.

Diversos modelos de RI vêm proporcionando melhorias significativas na relevância dos resultados. De acordo com Baeza-Yates e Ribeiro-Neto (1999), em uma visão centrada no computador, o problema de RI consiste principalmente na construção de índices mais eficientes, no processamento de *querys* de usuários com alta performance e no desenvolvimento de algoritmos de classificação que melhorem a “qualidade” do conjunto de respostas. Apesar disso, os métodos utilizados nesses modelos ainda deixam a desejar, não sendo capazes de recuperar a contento os documentos relevantes a uma consulta do usuário.

Na maioria dos modelos de recuperação de informação existentes hoje, o processo de indexação extrai cada palavra do texto de um documento e insere uma lista de palavras ordenadas, pela freqüência da palavra no texto. Isto desfaz o trabalho intelectual do autor do documento.

Observa-se que diversas pesquisas de RI se focalizam nos algoritmos de busca por documentos relevantes a partir de *querys* estabelecidas. O foco nesses casos é determinar a relevância de documentos. Para isso há várias metodologias, desde medir o tempo de

permanência do usuário no acesso a um documento até a determinação da quantidade de consultas com *querys* semelhantes, entre outras.

Outro aspecto problemático relaciona-se ao fato de que as informações recuperadas dependem também da clareza do usuário ao expressar o que necessita. Ou seja, a dificuldade não se trata apenas de identificar e definir a relevância dos resultados, através dos modelos computacionais de RI, que dão suporte ao processo da busca, mas da capacidade do usuário de formular uma expressão de busca, utilizando as palavras ou expressões de forma clara de modo a representar os documentos desejados, satisfazendo assim a sua necessidade.

As palavras utilizadas pelo usuário possuem um significado claro para ele, mas isso não é suficiente para uma boa recuperação de informação, pois a Língua Portuguesa, segundo Rossi (2003), apresenta muitas palavras iguais com significados diferentes (polissemia), que variam de acordo com o contexto. E há também palavras diferentes em escrita e pronúncia embora com significados iguais (sinonímia). Ocorre ainda a combinação de palavras, que, segundo Martins e Zilberknop (1999), diz respeito a duas ou mais palavras que podem combinar-se em ordem diferente, designando idéias completamente diversas.

Esses aspectos da linguagem natural são obstáculos na obtenção de bons resultados em um procedimento de recuperação de informação. No caso da polissemia e da combinação de palavras pode ocorrer o aumento da taxa de ruídos[1] ou o incremento da taxa de silêncio[2] que acontecem no caso de sinonímia. Isto pode levar a um resultado de busca de documentos que não atenda às necessidades de informação do usuário. Portanto, a existência de uma negociação de significados entre usuário e máquina levaria possivelmente a resultados mais relevantes.

O surgimento das novas tecnologias da informação e da comunicação fez crescer o volume de publicações na Internet. Esse crescimento, segundo Cardoso (2000), tem dificultado ainda mais a recuperação de informações relevantes. Um aspecto positivo é a facilidade de acesso, pela *Web* (*World Wide Web*), aos acervos bibliográficos de diversas universidades brasileiras e, mesmo, do mundo inteiro. Visto que o aumento do acervo torna ainda mais complexa a busca, por isso esperava-se que esses métodos acompanhassem tal desenvolvimento, mas isto ainda não aconteceu de forma satisfatória.

A dificuldade aparece rapidamente nos vários mecanismos da *Web*, como "Google", "Cade", entre outros, que, ao serem acionados para buscar uma determinada informação,

[1] Taxa de ruídos é definida como sendo a relação entre a quantidade de documentos recuperados não pertinentes e a quantidade total de documentos.
[2] Taxa de silêncio é definida como sendo a relação entre a quantidade de documentos recuperados pertinentes, não recuperados e a quantidade total de documentos pertinentes na base de dados.

listam centenas ou mesmo milhares de referências como resposta, sendo normalmente, destas, relevantes apenas as primeiras. Além disso, ao se utilizarem as mesmas palavras em diferentes mecanismos (sites) de pesquisa, os resultados variam, segundo Hill (1999), devido às rotinas automatizadas de pesquisa diferenciadas.

O usuário precisa ainda utilizar palavras-chave para dar foco à sua pesquisa. Segundo Baeza-Yates e Ribeiro-Neto (1999), o interessante seria já poder dizer: "Dê-me dados estatísticos sobre a equipe da seleção brasileira de basquete no ano de 2004". Mas, apesar de a tecnologia da Internet estar progredindo, ainda se está bastante distante desse estágio.

Uma linha de pesquisa que tem como representante o trabalho de Kuramoto (1999) procura abordar a questão da RI desde a perspectiva do apoio ao usuário na formulação da *query* de busca. A expectativa é oferecer já no momento da formulação da *query* um apoio interativo para o estabelecimento de uma chave mais adequada ao contexto real da busca. A proposta de Kuramoto é baseada na determinação dos Sintagmas Nominais (SN) de um domínio de aplicação.

O uso de SN permite um processo de refinamento da busca. A forma de navegar pelos níveis de SN intensifica a interação entre o usuário e o computador (KURAMOTO, 2002). A interface de busca passa a dar um suporte para o usuário na formulação de sua *query* antes de listar todos os documentos.

A proposta de utilização de uma interface de apoio utilizando SN configura-se como inovadora, pois não se tem conhecimento de outra proposição que considere o fato de que nem sempre o usuário é capaz de explicitar a sua necessidade de informação em uma única expressão de busca.

Segundo Kuramoto (2002), as palavras como unidades de um dicionário não contêm qualquer substância. Elas adquirem essa substância no momento em que se inserem no universo do discurso, ou seja, as palavras inseridas no texto de um documento assumem um significado específico.

Percebe-se que essa linha de pesquisa é bastante promissora e que a área de Lingüística pode oferecer alternativas interessantes; uma delas foi vislumbrada na teoria do Léxico Gerativo (LG) de Pustejovsky (1991). Nessa teoria, Pustejovsky, buscando dar conta da polissemia lógica das palavras propondo uma estrutura para a semântica de uma língua da mesma forma que a sintaxe é estruturada. Na estrutura proposta por Pustejovsky, a componente principal é a estrutura de dimensões de significados (denominada de Estrutura de *Qualia*).

Uma palavra escrita pelo usuário pode ser utilizada pelos documentos de um acervo e, portanto, identificada pela máquina através de seus modelos de RI com um sentido completamente diferente do contexto imaginado pelo usuário. Para a palavra "jornal", por exemplo, o usuário pode estar se referindo ao prédio onde fica o jornal, ou ao objeto físico propriamente dito ou até mesmo ao conteúdo do jornal (informação contida).

A Estrutura de *Qualia* auxilia a RI na identificação de qual sentido mais específico o usuário busca, dessa forma esta estrutura poderia classificar os documentos contendo a palavra "jornal" segundo as diferentes *qualia* envolvidas. Isso representaria um refinamento importante na busca, que poderia resultar em mais satisfação para o usuário e, portanto, mais eficiência dos mecanismos de busca. O reconhecimento da importância da teoria de Pustejovsky pode ser constatado na existência de trabalhos relacionados na língua portuguesa, como é o caso da pesquisa de Abrahão (1997) que desenvolveu a modelagem e a implementação de um léxico semântico para a nossa Língua, a partir de um estudo aprofundado da teoria de Pustejovsky.

Além disso, uma outra questão importante a ressaltar é que existem problemas ligados à definição das palavras. Essa crítica, segundo Rossi (2003), se fundamenta no fato de os lexicógrafos[3] parecerem atuar de maneira mais intuitiva do que propriamente fazer uso de teorias semânticas que dêem o devido suporte à tarefa de definir um item lexical. Rossi (2003) reforça que muitos dicionários nem sempre prevêem a polissemia subjacente aos itens lexicais.

Outro trabalho pesquisado que permitiu uma ampliação do modelo proposto nesta dissertação foi o de Gonzalez (2005) com o seu modelo TR+. Este modelo não utiliza sistematicamente a Estrutura de *Qualia*, aparecendo esta apenas implícita, principalmente a parte formal das palavras. As palavras e seus relacionamentos ganham em Gonzalez uma importância contextual pelo cálculo de um peso (peso de descritores) que busca manter sua unidade significativa.

A abordagem proposta para este trabalho orienta-se na melhoria da *query* de busca dos usuários. A pesquisa, síntese e sistematização da proposta de Kuramoto (1999) e do modelo de Gonzalez (2005) possibilitaram o desenvolvimento de um novo modelo chamado de SINTR+. Esse modelo utiliza a formulação de consulta em RI apresentando os Sintagmas Nominais referentes a esta consulta e com isto inicia a interação com o usuário onde o mesmo

[3] Lexicógrafos são autores de dicionários, ou seja, dicionaristas.

escolhe o SN de nível apropriado e a partir daí, há sistematização com o modelo TR+ de Gonzalez.

Pretende-se por um lado ajudar e apoiar o usuário a melhor especificar sua *query* no contexto real da sua busca, por outro lado, potencializa-se o tempo, tanto na fase de indexação como na de busca e reduz-se o espaço utilizado de memória para dados na base.

1.1 Objetivos

1.1.1 Objetivo Geral

Descrever, a partir da identificação das possibilidades de ampliação, de síntese e de sistematização das propostas de Kuramoto e de Gonzalez, um novo modelo para um sistema informatizado de suporte ao usuário na definição da sua *query* de busca durante um processo de recuperação de informação.

1.1.2 Objetivos Específicos

a) Analisar as propostas citadas buscando a sua sistematização e identificação de alternativas de implementação e ampliação.
b) Definir o modelo conceitual do sistema desejado, através da sua análise de domínio representando-o a partir dos seus diagramas de classes e de seqüência.
c) Avaliar exploratoriamente o modelo desenhado a partir da construção de exemplos demonstrativos das suas principais propriedades.

1.2 Metodologia

Para a construção deste trabalho, inicialmente foi realizada uma revisão bibliográfica a partir de livros, artigos e outros materiais disponíveis referentes ao assunto em questão, fundamentalmente sobre a área de Recuperação de Informação. A metodologia utilizada para desenvolver este trabalho baseou-se no cronograma de etapas a serem desenvolvidas descritas a seguir:

a) Estudo e identificação das diferentes alternativas e abordagens atualmente desenvolvidas para a área de recuperação de informações.
b) Formulação da proposta de trabalho: definição do escopo e da fundamentação da proposta.
c) Estudo das teorias de base para a construção do modelo: teoria do Léxico Gerativo de James Pustejovsky e o modelo de Kuramoto. E após um estudo de Abrahão e Gonzalez.
d) Esboço do modelo para o sistema proposto.
e) Especificação dos requisitos do sistema proposto.
f) Construção da análise de domínio: definição do modelo conceitual.
g) Construção dos diagramas de classes e de seqüência para o modelo.
h) Construção de exemplos de aplicação do modelo.
i) Análise e conclusões finais.

1.3 Resultados Esperados e Limitações do Trabalho

A principal contribuição deste trabalho reside no fato de sistematizar as teorias de Kuramoto, Pustejovsky e Gonzalez construindo um novo modelo que amplia as potencialidades das propostas de Kuramoto e Gonzalez melhorando os resultados do processo de recuperação de informações. Esta melhoria ocorre em relação à diminuição do tempo de busca dos documentos e à relevância dos resultados encontrados por meio da junção de diferentes modelos para os processos de indexação e busca.

A princípio o modelo construído é antevisto como aplicável a bases de documentos não distribuídas, e contidas a um determinado domínio de aplicação, mas já é possível perceber formas de adaptá-lo expandindo-o para seu uso na *Web*.

Este trabalho não tem o intuito de gerar uma implementação computacional completa do modelo proposto, propõe-se antes a demonstrar a viabilidade desta implementação, descrevendo os diagramas e as descrições dos casos de uso e a sua modelagem conceitual culminando a construção dos diagramas de classes e de seqüência. A análise das potencialidades e limitações do modelo deverá ser possível a partir da realização de estudos de casos onde se determine a complexidade computacional da implementação requerida.

1.4 Estrutura da Dissertação

O trabalho apresenta um capítulo introdutório que orienta os tópicos do projeto e o desenvolvimento da pesquisa, além de sintetizar os resultados que serão explorados na conclusão.

O Capítulo 2, a seguir, aborda temas e definições da área de RI, mostrando a sua história e também discute o funcionamento e as vantagens e desvantagens dos modelos clássicos de RI.

No Capítulo 3, apresenta-se a fundamentação teórica desta dissertação onde são abordados três autores. Primeiramente, apresenta-se a Proposta de Kuramoto que se baseia nos níveis de Sintagmas Nominais sendo exposto o protótipo de interação entre usuário e máquina desenvolvido por este autor. Na Teoria do Léxico Gerativo de Pustejovsky, deu-se ênfase à apresentação da Estrutura de *Qualia*, pois é a que foi julgada mais adequada para a aplicação no modelo proposto, apresenta-se também uma análise do estudo de Abrahão. Por fim, discute-se e apresenta-se o trabalho de Gonzalez e do seu modelo TR+ que possibilitou, juntamente com a proposta de Kuramoto sistematizar a proposta desta dissertação.

No Capítulo 4 é desenvolvida a proposta do sistema SINTR+ através dos diagramas e das descrições dos casos de uso do modelo, o modelo conceitual, os diagramas de classes e de seqüência juntamente com exemplos demonstrativos das suas propriedades.

No Capítulo 5, têm-se as conclusões referentes ao trabalho bem como as sugestões para continuidade desse foco de pesquisa.

O Capítulo 6 apresenta as referências bibliográficas utilizadas para a realização deste trabalho bem como a bibliografia consultada para a compreensão de conceitos abordados na dissertação finalizando com os anexos.

2. RECUPERAÇÃO DE INFORMAÇÃO

Neste capítulo apresentam-se o histórico e os modelos clássicos da área de recuperação de informação. O objetivo ao abordar esses tópicos é delinear uma visão geral da área a partir de diversos modelos de RI, apontando algumas de suas principais vantagens e desvantagens. Dar-se-á destaque ao fato de que os algoritmos de relevância utilizados para recuperar os documentos desconsideram o contexto da *query* de busca.

2.1 Histórico

Em 1951, segundo Baeza-Yates e Ribeiro-Neto (1999), Calvin Mooers criou o termo "*Information Retrieval"* (Recuperação de Informação) e definiu os problemas a serem abordados por esta nova área de pesquisa, a qual despertou o interesse principalmente de bibliotecários e "experts" da informação.

No contexto da Ciência da Informação, segundo Ferneda (2003, p. 14),

> o termo "Recuperação de Informação" significa, para uns, a operação pela qual se seleciona documentos, a partir do acervo, em função da demanda do usuário. Para outros, "Recuperação de Informação" consiste no fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondentes. O termo pode ainda ser empregado para designar a operação que fornece uma resposta mais ou menos elaborada a uma demanda, e esta resposta é convertida num produto cujo formato é acordado com o usuário (bibliografia, nota de síntese, etc.). Há ainda autores que conceituam a recuperação de informação de forma muito mais ampla, ao subordinar à mesma o tratamento da informação (catalogação, indexação, classificação).

Para alguns autores, segundo Cardoso (2000), RI é dita como uma subárea da Ciência da Computação que estuda o armazenamento e a recuperação automática de documentos, que são objetos de dados, geralmente textos. Para Baeza-Yates e Ribeiro-Neto (1999), o termo "Recuperação de Informação" trata da representação, do armazenamento, da organização e do acesso aos itens da informação.

De acordo com Ferneda (2003) foi a partir dos experimentos de Hans Peter Luhn (Engenheiro pesquisador da IBM) na indexação automática e na elaboração automática de resumos que surgiram os primeiros resultados significativos no tratamento computacional da informação. Com isto "Luhn foi durante vários anos o criador de inúmeros projetos que visavam modificar radicalmente métodos tradicionais de armazenamento, tratamento e

recuperação de informação. Em 1961, já acumulava cerca de 80 patentes nos Estados Unidos" (FERNEDA, 2003, p. 10-11). Estes dados mostram a importância de Luhn no tratamento da recuperação de informações.

Em 1960, segundo Ferneda (2003), foi desenvolvido os princípios básicos do modelo probabilístico para a Recuperação de Informação por Maron e Kuhns, que foi mais tarde definido por Robertson e Jones (1976). A década de 60 foi fundamental em experimentos desta natureza, "em meados dos anos 60 inicia-se uma longa série de experimentos que constitui um marco na Recuperação de Informação: o projeto SMART" (FERNEDA, 2003, p.11). Este autor destaca que este projeto foi desenvolvido por Gerard Salton que se especializou na pesquisa destas evoluções na recuperação de informações, produzindo inúmeros artigos científicos, um modelo de recuperação de informação, a criação e o aprimoramento de diversas técnicas computacionais, além de o sistema SMART.

Estes sistemas de recuperação de informação, geralmente se baseiam na contagem de freqüência das palavras do texto e na eliminação de palavras reconhecidamente de pouca relevância (FERNEDA, 2003). Um exemplo disso são os métodos automáticos de indexação de recuperação de informação que utilizam "filtros" para eliminar palavras de pouca significação (*stopwords*[4] e *noun groups*[5]), além de normalizar os termos reduzindo-os a seus radicais. Esse processo é conhecido como *stemming*[6].

Ferneda evidencia que os trabalhos de Luhn e Salton inicialmente não se preocupavam com a análise semântica das palavras e que seus estudos colaboraram para com a evolução atual das pesquisas.

> Nos trabalhos de Luhn e Salton observa-se inicialmente uma crença de que métodos puramente estatísticos seriam suficientes para tratar os problemas relacionados à recuperação de informação. Porém, no transcorrer de suas pesquisas, percebe-se uma busca por métodos de análise semântica mais sofisticada. Desde os seus primeiros trabalhos, Salton se mostra interessado pela utilização de processos de tratamento da linguagem natural na recuperação de informação. Em livro de 1983, Salton e McGill apresentam em um capítulo intitulado *Future directions in Information Retrieval* a aplicação do processamento da linguagem natural e da lógica *fuzzy* na recuperação de informação, apontando a direção de futuras pesquisas para a Inteligência Artificial. (FERNEDA, 2003, p. 12)

Estas contribuições têm suas principais idéias presentes ainda na maioria dos sistemas de recuperação atuais e nos mecanismos de busca da *Web*. Como aparece na estrutura de componentes de um sistema de recuperação de informação que seguem geralmente um modelo de funcionamento como demonstrado por Cardoso (2000).

---

[4] *Stop Words:* eliminação de artigos e conectivos.
[5] *Noun Groups:* eliminação de adjetivos, advérbios e verbos.
[6] *Stemming:* redução de uma palavra ao seu radical. Exemplo: *Engineering Engineer*.

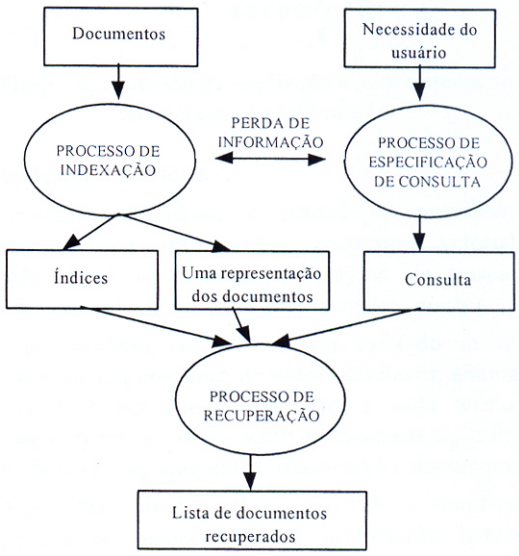

Figura 1: Componentes de um Sistema de Recuperação de Informação
Fonte: GEY apud CARDOSO, 2000.

2.2 Modelos Clássicos de Recuperação de Informação

2.2.1 Modelo Booleano

A álgebra booleana é um sistema binário no qual existem somente dois valores possíveis para qualquer símbolo algébrico: “verdadeiro” ou “falso”. O modelo booleano é um modelo de recuperação simples baseado na teoria dos conjuntos e na álgebra booleana. Além disso, as *querys* são especificadas através de expressões booleanas que têm semânticas precisa.

Segundo Baeza-Yates e Ribeiro-Neto (1999) e Gonzalez (2000), a simplicidade e o formalismo claro do modelo booleano recebiam grande atenção nos anos passados, sendo adotados por muitos sistemas comerciais bibliográficos.

A estratégia de recuperação desse modelo é baseada em um critério de decisão binária, por exemplo, um documento pode ser relevante ou não relevante, sem noção de escala de classificação que previna um bom desempenho na recuperação. Deste modo, o modelo booleano é na verdade muito mais um modelo de recuperação de dados (em vez de informação).

Além disso, conforme Baeza-Yates e Ribeiro-Neto (1999), enquanto expressões booleanas têm semânticas precisas, freqüentemente não é simples traduzir uma informação precisa dentro de uma expressão booleana. O modelo booleano prediz que cada documento é relevante ou irrelevante. Não existe noção de um resultado (*matching*) parcial para as condições da *query*.

As principais vantagens do modelo booleano são o formalismo claro oculto sobre o modelo e sua simplicidade. As principais desvantagens encontram-se no resultado exato, que pode recuperar poucos ou muitos documentos.

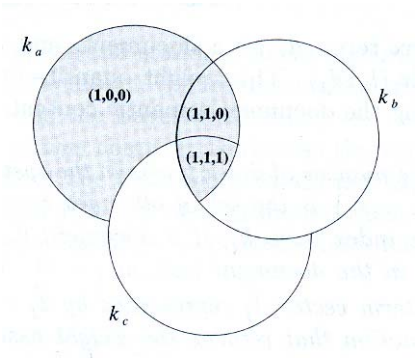

Figura 2: Exemplo dos três componentes conjuntivos para query
Fonte: BAEZA-YATES; RIBEIRO-NETO, 1999.

2.2.1.1 Operadores Booleanos

Os operadores booleanos funcionam através de uma expressão booleana para formulação de buscas. Isto ocorre por meio de operadores lógicos AND, OR e NOT (E, OU e NÃO). Conforme exemplo de Ferneda (2003) a recuperação de informação se dará em uma expressão conjuntiva de enunciado **t1** AND **t2** que recuperará documentos indexados por ambos os termos (t1 e t2). Isso equivale e permite aparecer à intersecção do conjunto dos documentos indexados pelo termo t1 com o conjunto dos documentos indexados pelo termo t2.

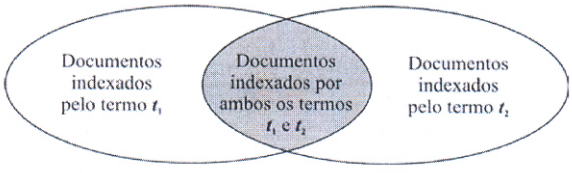

Figura 3: Representação do resultado de uma expressão booleana conjuntiva (AND)
Fonte: FERNEDA, 2003.

O autor demonstra que uma expressão disjuntiva t1 OR t2 recuperará o conjunto dos documentos indexados pelo termo t1 ou pelo termo t2. Isto equivale e possibilita à união entre o conjunto dos documentos indexados pelo termo t1 e o conjunto dos documentos indexados pelo termo t2 (FERNEDA, 2003).

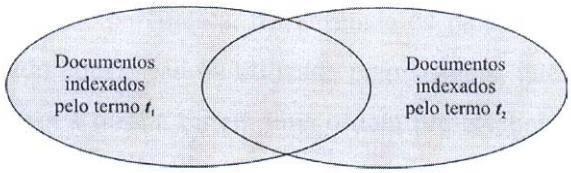

Figura 4: Resultado de uma busca booleana disjuntiva (OR)
Fonte: FERNEDA, 2003.

2.2.1.2 Operadores de Proximidade

No modelo booleano existem os operadores de proximidade que permitem especificar condições relacionadas à distância e à posição dos termos no texto. Um operador de proximidade bastante comum nos sistemas de RI e nos mecanismos de busca da Web é o operador ADJ (FERNEDA, 2003). Esse operador permite pesquisar duas palavras adjacentes no texto de um documento, na ordem especificada na expressão de busca, por exemplo, a expressão **recuperação ADJ informação** terá como resultado os documentos que tiverem a palavra "recuperação", seguida da palavra "informação", ou seja, recuperará documentos que contêm a expressão **"recuperação informação"**. Também pode ser utilizado um termo composto delimitando as suas palavras com aspas, por exemplo, **"recuperação de informação"**.

O modelo booleano, de acordo com Ferneda (2003), possui limitações que o torna pouco atrativo, são elas:

- O resultado de uma busca booleana se caracteriza por dois subconjuntos: os que atendem à expressão de busca e aqueles que não atendem. Presume-se que todos os documentos recuperados são de igual utilidade para o usuário. Não há nenhum mecanismo pelos quais os documentos possam ser ordenados;
- O usuário leigo, se não tiver um treinamento apropriado, formulará somente buscas simples. Para buscas com expressões mais complexas é necessário um conhecimento da lógica booleana;

- Não existe uma forma de atribuir importância relativa aos diferentes termos da expressão booleana. Assume-se implicitamente que todos os termos têm o mesmo peso.

2.2.2 Modelo Vetorial

O modelo vetorial, segundo Baeza-Yates e Ribeiro-Neto (1999), reconhece que o uso de pesos binários é também limitante e propõe uma estrutura em que é possível a resposta (*matching*) parcial. Isto é feito atribuindo-se pesos não binários aos termos indexados em *querys* e em documentos. Esses pesos de termos são, enfim, utilizados para calcular o grau de similaridade entre cada documento armazenado no sistema e a expressão de busca formulada pelo usuário (*querys*). Como a classificação dos documentos recuperados é feita em ordem decrescente desse grau de similaridade, o modelo vetorial leva em consideração documentos que se igualem aos termos de *querys* somente parcialmente.

O modelo vetorial, de acordo com Cardoso (2000) e Gonzalez (2000), representa documentos e consultas como vetores de termos. Os termos são ocorrências únicas nos documentos. Os documentos, retornados como resultado para uma consulta, são representados similarmente, isto quer dizer que o vetor resultado para uma consulta é montado através de um cálculo de similaridade. Aos termos das consultas e dos documentos são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. O ângulo formado por esses vetores determina a proximidade da ocorrência. E o cálculo da similaridade é baseado no ângulo entre os vetores que representam o documento e a consulta.

Cardoso (2000) descreve ainda que os pesos quantificam a relevância de cada termo para as consultas (W_{iq}) e para os documentos (W_{id}) no espaço vetorial. Segundo Cardoso (2000, p. 03) "para o cálculo dos pesos W_{iq} e W_{id}, utiliza-se uma técnica que faz o balanceamento entre as características do documento, utilizando o conceito de freqüência de um termo num documento". Desta forma, se uma coleção possui N documentos e teremos o n_{ti} que é a quantidade de documentos que possuem o termo t_i, com isto o inverso da freqüência do termo na coleção, ou *idf* (*inverse documento frequency*), é dado pela fórmula de Cardoso (2000) abaixo:

$$idf_i = \log (N/n_i)$$

Esse valor é possível usando a fórmula para calcular o peso, $W_{id} = freq(t_i,d) \times idf_i$, que é o produto da freqüência do termo no documento pelo inverso da freqüência do termo na coleção.

No modelo vetorial um documento é representado por um vetor em que cada elemento representa o peso ou a relevância do respectivo termo de indexação para o documento. Cada elemento do vetor (peso) é normalizado de forma a assumir valores entre zero e um. Os pesos mais próximos de um (1) indicam termos com maior importância para a descrição do documento. E termos que não estão presentes em um determinado documento possuem peso igual a zero.

Da mesma forma que os documentos, no modelo vetorial uma expressão de busca, conforme Baeza-Yates e Ribeiro-Neto (1999), também é representada por um vetor numérico em que cada elemento representa a importância (peso) do respectivo termo na expressão de busca.

Diversos documentos e termos de indexação podem ser representados através de uma matriz na qual cada linha representa um documento e cada coluna representa a associação de um determinado termo aos vários documentos.

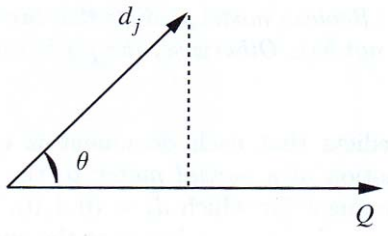

Figura 5: O co-seno do ângulo adaptado como similar (dj, q)
Fonte: BAEZA-YATES; RIBEIRO-NETO, 1999.

Um exemplo de uso do modelo vetorial é o sistema SMART[7], citado anteriormente, este sistema, representa por valor numérico cada documento e seu respectivo termo na descrição do documento. Segundo Ferneda (2003) o sistema SMART fornece um método automático que trata além do cálculo dos pesos dos vetores que representam os documentos também trata os vetores das expressões de busca.

As principais vantagens do modelo vetorial, segundo Baeza-Yates e Ribeiro-Neto (1999), são: (1) esquema de pesos de termos melhora o desempenho da recuperação; (2) estratégias de resposta (*matching*) parcial permitem a recuperação de documentos que se aproximem de condições de *query*; e (3) fórmula de classificação do co-seno ordena os documentos de acordo com o grau de similaridade da *query*. A desvantagem desse modelo, de

acordo com os autores, diz respeito às dependências de termos, prejudicando especialmente o desempenho.

Cardoso (2000) considera como principais vantagens do modelo vetorial a sua simplicidade, a facilidade de se computarem similaridades com eficiência e o fato de que se comporta bem com coleções genéricas.

2.2.3 Modelo Probabilístico

O modelo probabilístico foi introduzido, de acordo com Baeza-Yates e Ribeiro-Neto (1999), em 1976 por Roberston e Sparck Jones, que mais tarde tornou-se como o modelo *Binary Independence Retrieval* (BIR).

Na Matemática, a teoria das probabilidades estuda os experimentos aleatórios, que, conforme Ferneda (2003, p. 35),

> repetidos em condições idênticas, podem apresentar resultados diferentes e imprevisíveis. Isso ocorre, por exemplo, quando se observa a face superior de um dado após o seu lançamento ou quando se verifica o naipe de uma carta retirada de um baralho. Por apresentarem resultados imprevisíveis, é possível apenas estimar a possibilidade ou a chance de um determinado evento ocorrer.
> Para descrever matematicamente um experimento aleatório é necessário inicialmente identificar o conjunto de todos os seus possíveis resultados. A este conjunto dá-se o nome de *espaço amostral*.

Entendendo-se uma busca como um experimento aleatório, segundo Robertson e Jones, é possível descrever o seu espaço amostral como composto de quatro possibilidades, pois, dada uma expressão de busca, pode-se dividir a base de documentos em quatro subconjuntos distintos: o conjunto dos documentos relevantes (Rel), o conjunto dos documentos recuperados (Rec), o conjunto dos documentos relevantes e recuperados (RR) e o conjunto dos documentos não relevantes e não recuperados. O conjunto dos documentos relevantes e recuperados (RR) é resultante da intersecção dos conjuntos Rel e Rec (FERNEDA, 2003).

O conjunto de documentos resultantes da primeira busca é ordenado através de uma forma de ordenação padrão tradicional. Tendo esse conjunto de documentos, o usuário seleciona alguns deles que considera relevantes para a sua necessidade. O sistema utiliza essa informação para tentar melhorar os resultados subseqüentes.

A principal virtude do modelo probabilístico está em reconhecer que a atribuição de relevância é uma tarefa do usuário. É o único modelo que, segundo Baeza-Yates e Ribeiro-

---

[7] SMART (*Sistem for the Manipulation and Retrieval of Text*).

Neto (1999) e Gonzalez (2000), incorpora explicitamente o processo de *Relevance Feedback* como base para a sua operacionalização.

Uma simplificação bastante questionável está no fato de o modelo considerar os pesos dos termos de indexação como sendo binários, ou seja, no modelo probabilístico não é considerada a freqüência com que os termos ocorrem no texto dos documentos.

Em geral, os modelos de RI desconsideram o contexto das palavras informadas pelo usuário, por isso tendem a retornar poucos documentos relevantes em uma consulta. Para isso, pretende-se mostrar no capítulo seguinte, com a ajuda da Lingüística, possíveis abordagens que podem apoiar o usuário, considerando o seu contexto de busca e listando documentos relevantes.

3. FUNDAMENTAÇÃO TÉORICA

Neste capítulo buscou-se apresentar uma síntese dos trabalhos que dão base ao modelo apresentado nesta dissertação. São eles: a Proposta de Kuramoto, a Teoria do Léxico Gerativo e o Modelo de Gonzalez. A Proposta de Kuramoto baseia-se em uma hierarquização em níveis de Sintagmas Nominais. Na Teoria do Léxico Gerativo de Pustejovsky, mostram-se as estruturas compostas e deu-se destaque à Estrutura de *Qualia*, julgada mais adequada para a aplicação no trabalho proposto. Analisou-se o estudo de Abrahão a partir de Pustejovsky. A terceira teoria, de Gonzalez, apresenta uma proposta automatizada com o modelo TR+.

3.1 A Proposta de Kuramoto

Neste capítulo apresentam-se os conceitos e as características da proposta de Kuramoto, que se baseia na determinação de Sintagmas Nominais (SN) de uma *query*. A sua proposta preocupa-se em buscar os SN, uma vez que são considerados como importante elemento de uma frase, sendo entendidos como o núcleo significativo (cerne) de uma oração.

Em sua tese de doutorado, Kuramoto relata que todo o trabalho de reconhecimento e extração de SN dos documentos foi realizado de forma não automatizada. Isto auxiliou na elaboração de um modelo para reconhecimento, extração e indexação de SN, inseridos na amostra do protótipo desenvolvido.

O modelo proposto por Kuramoto refere-se ao aproveitamento dos SN organizado hierarquicamente em "árvores", criando um novo conceito de indexação que pode introduzir inovação em termos de uma interface de busca.

Esse modelo de interface, de acordo com Kuramoto (2002), permitiria que o usuário navegasse no conjunto de SN até encontrar o que melhor atendesse à sua necessidade de informação. Somente após esse procedimento, o usuário teria então acesso aos documentos de onde foram extraídos os SN. Tal processo proporcionaria ao usuário um maior conhecimento sobre a base de dados que está sendo consultada, uma vez que lhe permitiria reconhecer a estrutura de sintagmas nominais presentes nos documentos pertencentes ao sistema.

Os processos de indexação automática, utilizados em modelos de RI, segundo Michel Le Guern (1984 apud KURAMOTO, 1995), deveriam extrair dos documentos informações

que facilitassem a recuperação para o usuário e não símbolos sem referência, como considera que são as palavras.

Para Silva e Koch (1993), toda frase de uma língua constitui uma organização, ou seja, uma combinação de elementos lingüísticos agrupados conforme certos princípios, que a caracterizam como uma estrutura. Para Baeza-Yates e Ribeiro-Neto (1999), grande parte da semântica do documento ou da requisição do usuário é perdida quando se substitui o texto completo por um conjunto de palavras.

Aparentemente, um conjunto de frases de nossa língua, de acordo com Silva e Koch (1993), tem pouco em comum, variando quanto à extensão, ao sentido, às palavras de que se compõem e à ordem em que essas se apresentam. Apesar da aparente diversidade, as frases possuem uma organização interna que segue princípios gerais bem definidos de modo que o falante será capaz de dizer se uma seqüência de palavras: a) se está de acordo com o sistema gramatical da língua; b) se se apresenta completa ou incompleta; c) se é passível de interpretação semântica.

Conforme Silva e Koch (apud ABREU et al., 2004, p.03), "o sintagma consiste num conjunto de elementos que constituem uma unidade significativa dentro da oração e que mantêm entre si relações de dependência e de ordem". As palavras se combinam em conjuntos em torno de um núcleo. Esses conjuntos, os sintagmas, desempenham uma função no conjunto maior, que é a frase. Para Liberato (apud PARREIRAS, 2003), o SN é a parte do enunciado que representa um conceito ou referente.

Assim, por exemplo, nos conjuntos de sintagmas – David, o estudante; a menina doente; e minha filha –, o núcleo é um elemento nominal (nome ou pronome), tratando-se, portanto, de sintagmas nominais. Nos conjuntos – viajou de carro; dormiu; e levará a encomenda – o elemento fundamental é o verbo, de modo que se têm, nesses casos, sintagmas verbais.

A natureza do sintagma depende, portanto, do tipo de elemento que constitui o seu núcleo; além do sintagma nominal (SN) e do sintagma verbal (SV), existem os sintagmas adjetivais (SA), que têm por núcleo um adjetivo, e os sintagmas preposicionais (SP), formados normalmente de preposição mais sintagma nominal (SILVA; KOCH, 1993).

Na estrutura da oração, em sua forma de base, aparecem como constituintes obrigatórios o SN e o SV. Por exemplo: **Os garotos** (SN) **empinavam papagaios de papel** (SV). Pode-se dizer que as regras básicas de estrutura frasal são as seguintes: O = SN + SV (SP) (o elemento O significa Oração).

3.1.1 Extração dos Sintagmas Nominais

O trabalho de Kuramoto compreendeu o desenvolvimento de um protótipo de interface de busca utilizando os sintagmas nominais como forma de acesso à informação. Para testar esse protótipo foram examinados e extraídos, segundo Kuramoto (2002), cerca de 8.800 sintagmas nominais de uma amostra de 15 artigos selecionados aleatoriamente da revista Ciência da Informação.

Kuramoto (1995, p. 6) relata que:

> a extração dos sintagmas nominais foi realizada de forma manual, simulando uma extração automática. Este procedimento foi adotado em função da não-existência ainda de um sistema de extração automática de SN em acervos contendo documentos em Língua Portuguesa.

Como os SN nem sempre se apresentam de forma clara, Kuramoto aponta a ocorrência normal em todo texto em linguagem natural de anáforas[8] e de elipses[9] que dificultou a identificação dos SN. Essas dificuldades, segundo Kuramoto (1995), aumentam em um processo automatizado. Algumas das dificuldades encontradas por Kuramoto no procedimento de extração dos SN são descritas a seguir.

a) SN escondidos em frases com fatoração

Para Kuramoto (1995, p. 06) as "frases com fatoração são aquelas que contêm uma seqüência de palavras que precedem um outro conjunto de palavras coordenadas pelas conjunções **e/ou**, por exemplo, **o processo de negociação dos setores privado e público".**

Percebe-se, nesse exemplo, que o SN de nível 1 compreende tanto os setores privado e público, visto que a referência dos dois adjetivos está contida na palavra em plural "setores". Existem outros exemplos de frases com fatoração nas quais as palavras coordenadas aparecem entre parênteses, significando um complemento combinatório do termo ou da frase que precede o parêntese, por exemplo, **profundas transformações (políticas, econômicas, sociais, tecnológicas)**.

b) Artigo Zero

---

[8] Em Lingüística, segundo Ducrot e Todorov (1972 apud KURAMOTO, 1995), um segmento do discurso é dito anafórico quando, para interpretá-lo (inclusive do ponto de vista literário), for necessário se reportar a um outro segmento do mesmo discurso.

[9] A figura de sintaxe "elipse" é definida por Cunha e Cintra (1991 apud KURAMOTO, 1995) como sendo a omissão de um termo que o contexto ou a situação permitem facilmente suprimir.

Um outro fator de dificuldade na extração dos SN é a freqüente ausência de determinantes[10] na língua portuguesa, diferente da língua francesa na qual são raros os SN com ausência de um determinante. Motivo pelos quais algumas regras estabelecidas para a língua francesa não foram utilizadas. De acordo com Kuramoto (1995, p. 7), "no procedimento de extração dos SN, constatou-se que 28,89% dos SN não eram precedidos de qualquer determinante. Em uma amostra de 6.010 SN, 1.736 SN não são precedidos por nenhum determinante". Estes números demonstram que o modelo necessário deve considerar este fator.

c) Cálculo das anáforas

Quando uma entidade é referenciada pela primeira vez em um texto, segundo Gasperin, Goulart e Vieira (2003), a expressão que a descreve é dita nova no discurso. Quando tal entidade é retomada no texto, a expressão que a descreve é dita anafórica, sendo considerado o seu antecedente a expressão anterior correferente.

Para Kuramoto (1995, p. 7-8), "os elementos anafóricos, em português, aparecem freqüentemente mediante partículas como os pronomes". No entanto, na proposta do autor, não foi possível resolver dois casos de anáforas.

Um primeiro caso de anáfora ocorre nas palavras sem fonte explícita no texto, tais como "nesse sentido**"** (em que sentido?), "nossa experiência" (de quem? do autor? dos técnicos de informação?) etc. Como a interpretação das idéias está contida no documento não fica evidente a solução desse tipo de anáfora.

O segundo caso é constituído de termos cujas fontes se encontram, como por exemplo, na história dos acontecimentos, como "esse período pré-industrial*,* esse sistema de comunicação" etc. Por este motivo os SN foram extraídos da mesma forma como se encontravam no texto.

d) Cálculo das elipses

Outra questão que necessita um entendimento do contexto de uma frase é o problema ligado a este tipo de figura de sintaxe. Visto que, depende da capacidade de percepção da falta de alguma palavra no contexto de uma frase. Segundo Kuramoto (1995), é preciso, para identificá-la, analisar não somente as frases precedentes, mas também as frases seguintes. Como neste exemplo: "uma visão de longo prazo que assegure não só a sobrevivência (?)

[10] Segundo Silva e Koch (1993), o determinante, quando simples, é representado por um artigo, numeral ou pronome adjetivo.

como também o crescimento da organização". Que promove o questionamento de "qual o complemento do termo 'sobrevivência'? 'Sobrevivência' de quem?" A solução encontrada poderia estar na frase seguinte: "o crescimento da organização".

Para promover a extração completa da frase o SN seria: "uma visão de longo prazo que assegure não só a sobrevivência da organização como também o crescimento da organização".

3.1.1.1 Extração Automática de Sintagmas Nominais

A extração automática de SN é considerada importante para a área de RI, pois, segundo Chishman et al (2000), agiliza este processo, e gera um percentual baixo de erros. Já foi desenvolvido um extrator automático de sintagmas nominais para a língua portuguesa no âmbito do projeto VISL chamado "Palavras"[11], que vem sendo usado pelo grupo de pesquisa da UNISINOS.

Segundo Abreu, Goulart e Vieira (2004), para obter a análise das sentenças dos textos, utiliza-se o analisador sintático "Palavras", que é considerada uma ferramenta robusta para a análise sintática do português.

A partir da saída do analisador sintático, segundo Gasperin, Goulart e Vieira (2003), a ferramenta "Xtractor" gera três arquivos XML. O primeiro é o arquivo de palavras; o segundo inclui as categorias morfossintáticas; e o terceiro é o arquivo com as estruturas sintáticas das sentenças.

Assim, após todo esse processo é possível extrair de modo automático os sintagmas nominais das sentenças de um texto, ressaltando-se que estes não estão ainda organizados segundo a estrutura de níveis que propõe Kuramoto.

3.1.2 A determinação de uma estrutura para os SN

A essência da proposta de Kuramoto (1995) reside na percepção que o autor teve de que os SN organizam-se naturalmente numa estrutura de níveis encadeados. Kuramoto percebeu nessa organização em níveis um caminho para propiciar ao usuário mais facilidade

[11] O analisador Palavras faz parte de um grupo de analisadores sintáticos (softwares) do projeto VISL - Visual Interactive Sintax Learning, do Institute of Language and Communication da University of Southern Denmark. Disponível em: <http://visl.sdu.dk/visl/pt/parsing/automatic/>. (ABREU; GOULART; VIEIRA, 2004).

no uso de um SRI, levando também a resultados mais precisos. Para compreender a estrutura proposta pelo autor, apresenta-se a seguir o exemplo usado pelo próprio Kuramoto:

As Características do Meio Ambiente do Mundo dos Negócios
SN1: os negócios
SN2: o mundo dos negócios
SN3: o meio ambiente do mundo dos negócios
SN4: as características do meio ambiente do mundo dos negócios

Figura 6: Exemplo da estrutura de níveis de Sintagmas Nominais
Fonte: KURAMOTO, 1995.

Esse exemplo mostra o potencial da estrutura de relações de encadeamento de um conjunto de SN. Para o autor,

> a análise do sintagma nominal, no exemplo, permitiu a extração do SN – o meio ambiente do mundo dos negócios. A partir desse SN, pode-se visualizar um outro SN embutido – o mundo dos negócios – que, por sua vez, possui um quarto SN – os negócios – que representa o nível mais inferior[12]. Percebe-se, nesse exemplo, a existência de quatro SN encadeados que, enumerados em ordem crescente (do SN mais simples ao mais complexo), levam à classificação do SN original como sendo de nível 4 (KURAMOTO, 1995, p.04).

Com base nessas características apresentadas por Kuramoto (1995), os SN podem ser organizados sob uma estrutura de árvore. Esta estrutura possibilita que o Sistema de Recuperação de Informação (SRI), possa atender às necessidades de consultas do usuário. Para atender esta demanda é preciso fornecer um centro de SN de seu interesse (como o exemplo do autor: "negócios").

Para isso, apresentam-se todos os SN1 relativos a essa busca, inclusive o SN "os negócios". A partir da lista encontrada de SN1, o usuário poderá restringir o seu perfil de busca escolhendo um SN1, por exemplo, "os negócios", e solicitar os SN2 relacionados a esse SN1. O SRI apresenta todos os SN2, inclusive o SN "o mundo dos negócios" e assim sucessivamente (KURAMOTO, 1995).

Este autor afirma que esta passagem por vários níveis promove um refinamento no processo.

> O processo de refinamento é realizado por meio da passagem pelos vários níveis de uma estrutura arborescente de SN[13], dado que o SN vai se tornando mais específico

[12] Segundo Kuramoto (1995), os sintagmas nominais, à medida que são extraídos de um outro SN, são classificados por níveis. Assim, o sintagma mais simples é denominado SN de nível 1. Constitui SN de nível 2 aquele a partir do qual foi extraído o de nível 1 e assim sucessivamente.

[13] Constatou-se empiricamente, utilizando a maquete desenvolvida nesta experimentação, de acordo com Kuramoto (1995), que a quantidade de SN de segundo nível em relação a um dado SN de primeiro nível pode ser maior que o total de SN de primeiro nível. Por exemplo: a resposta à demanda do centro de SN "informação" foi de 122 SN de primeiro nível, e a resposta à demanda do SN de primeiro nível "a informação" foi de 172 SN de segundo nível. Por outro lado, verificou-se que

> à medida que se atingem os níveis mais elevados da estrutura. Ao percorrê-la, o usuário está, na realidade, delimitando, ou melhor, qualificando a sua necessidade de informação. Cabe, portanto, ao usuário identificar o nível em que as suas necessidades de informação serão atendidas. (KURAMOTO, 1995, p. 04-05)

Esta possibilidade de hierarquia permite uma interação entre o usuário e máquina e uma escolha individual de refinamento.

3.1.3 Protótipo: Desenho da Interface de Busca

A Figura 7 descreve de maneira esquemática a interação entre o usuário e o protótipo de Kuramoto (1995).

O protótipo viabiliza a primeira interação, pois há uma tela em que permite ao usuário fazer a sua solicitação de informação fornecendo uma palavra (centro de SN1). A partir dessa palavra surgem outras interações, como mostra o esquema de Kuramoto (1995) na Figura 7 que ocorrem nas ações abaixo.

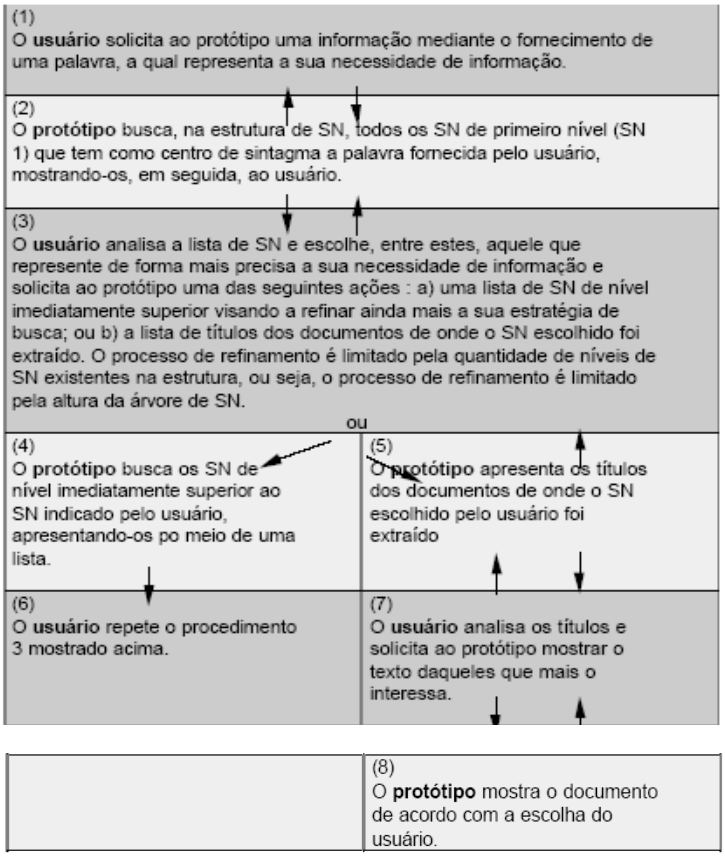

Figura 7: Procedimentos de interação usuário–protótipo
Fonte: KURAMOTO, 1995.

---

o SN "a informação" indexava 15 documentos na base, enquanto o SN de segundo nível "a análise da informação" indexava apenas 1 (um) documento. Confirma-se, nesse exemplo, que a passagem de um dado nível a um superior na árvore de SN proporciona maior refinamento no processo de seleção dos documentos.

3.1.4 Organização dos Sintagmas Nominais como Estrutura de Busca

Na proposta de Kuramoto (1995) foram desenvolvidas as seguintes estruturas de busca:

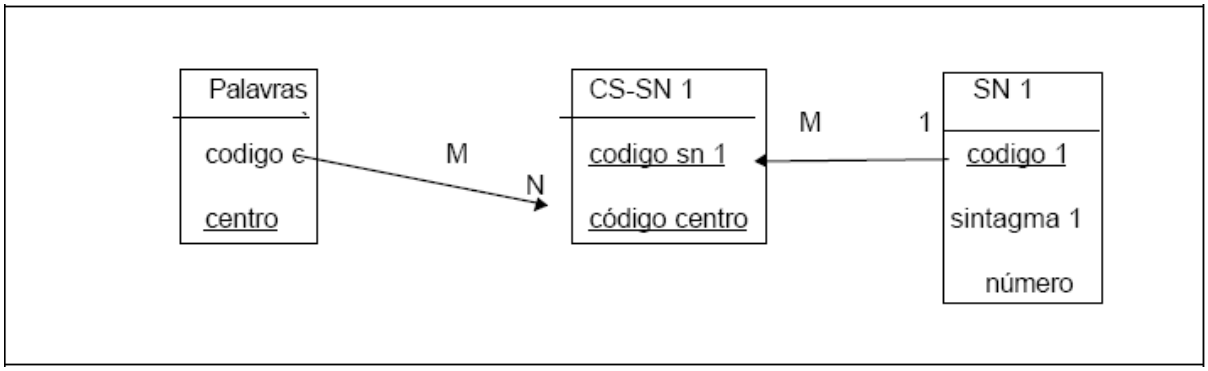

Figura 8: Estrutura de dados para acessar os Sintagmas Nominais de primeiro nível a partir de uma palavra
Fonte: KURAMOTO, 1995.

Kuramoto (1995) mostra na Figura 8 a associação das tabelas Palavras, CS-SN1 e SN1. Cada dado tem nomes dos elementos que estão sublinhados e representam as chaves de cada tabela. Na tabela Palavras, observa-se que o autor agrupa todas as palavras (centro) que representam os centros de SN1. Há uma atribuição de código para cada "centro" chamado "código c". A tabela CS-SN1 é uma tabela de associação dos códigos dos centros de SN1 com os códigos dos SN1.

> Essa figura mostra que para cada centro de SN1 existem vários SN1. A indicação na seta da associação da tabela Palavras com a tabela CS-SN1 define que, na tabela Palavras, podem existir M ocorrências de um código de centro de SN1. O mesmo pode ocorrer na tabela CS-SN1, em que esse código pode verificar-se N vezes. Essa indicação traduz a idéia de que para cada SN1 pode existir mais de um centro de SN1. Isto se explica pela existência, no contexto de um SN, de palavras que são tão importantes quanto o centro de sintagma. (KURAMOTO, 1995, p. 11)

Observa-se o exemplo "o sistema de informação**".** Nesse o autor define o SN1, de "sistema**".** Todavia, esta não é a única palavra fundamental, pois a palavra "informação**"** tem tanta importância quanto o próprio centro de sintagma (sistema).

Kuramoto (1995, p. 11) mostra ainda que existe associação entre o centro de SN1 e a vários SN de nível 1.

> Cada centro de SN1 pode estar associado a mais de um SN1. Essa indicação é dada pela seta que associa a tabela SN1 à tabela CS-SN1, onde o número 1 significa que, na tabela SN1, existe uma só ocorrência de um determinado código de SN1, enquanto, na tabela CS-SN1, existem M ocorrências desse código.

Outro elemento de dados importante na tabela SN1, é chamado "número", que, segundo Kuramoto (1995, p. 11-12) "indica a quantidade de artigos de onde um determinado

SN1 foi extraído". O número de referências de onde o SN foi extraído aparece para cada apresentação de SN1 relacionado com um centro de SN1, escolhido pelo usuário.

Kuramoto (1995) ilustra numa outra figura (Figura 9) a estrutura de dados construída para a busca dos SN2 a partir de um SN1 selecionado pelo usuário.

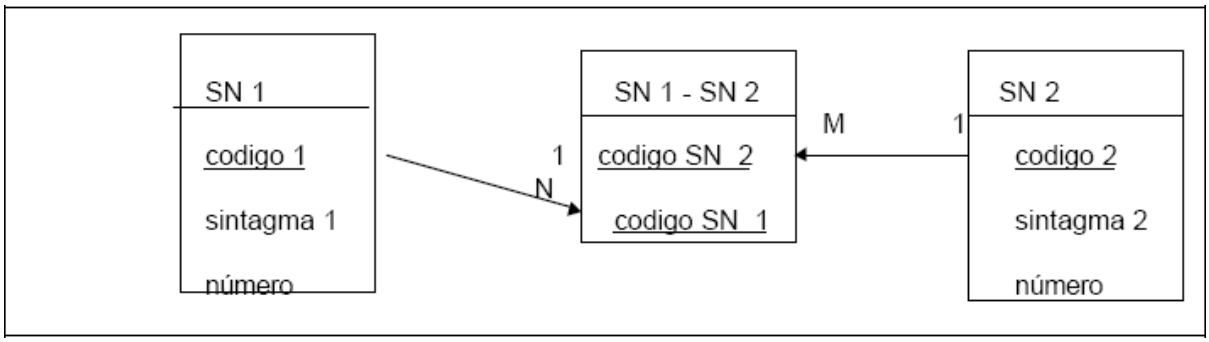

Figura 9: Estrutura de dados para acessar os Sintagmas Nominais de segundo nível a partir de Sintagmas Nominais de primeiro nível
Fonte: KURAMOTO, 1995.

Nessa ilustração, observa-se que se mantém a estrutura da Figura 8, em uma associação de tabelas que busca facilitar a busca dos SN2 a partir de um SN1 escolhido pelo usuário. Segundo Kuramoto (1995, p. 12), "percebe-se, analogamente, que um dado SN1 pode estar associado a vários SN2, e vice-versa. Isto traduz a idéia de que um SN2 pode ter embutido mais de um SN1. Essa estrutura atende às características dos SN listados no início desta seção".

A busca de informações se mantém na mesma estrutura para os SN de nível 3 e 4 que são semelhantes às Figuras acima (SN1 e SN2) com diferença apenas no nome de cada elemento que é correspondente ao número dos SN.

O acesso aos documentos está representado na Figura 10 que exemplifica uma escolha no SN1.

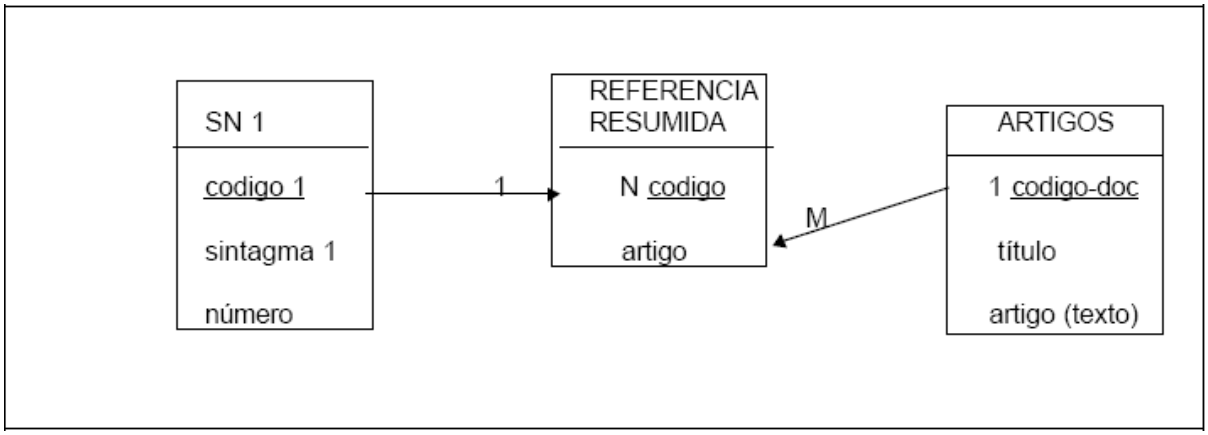

Figura 10: Estrutura de dados para o acesso aos títulos e textos dos artigos
Fonte: KURAMOTO, 1995.

Essa estrutura foi desenvolvida para que o protótipo atenda a uma demanda do usuário, viabilizando a visualização de todos os títulos e textos de documentos de onde um SN1 foi extraído. Há outras associações semelhantes a essas da Figura 17 que servem para o acessar os documentos a partir de SN de qualquer um dos quatro níveis previstos no protótipo.

Kuramoto (1995, p. 12-13) ressalta ainda as ações do código numérico.

> É importante observar que todas as tabelas contendo os SN nos seus vários níveis têm como chave de acesso um código numérico único de SN. Para tanto, construiu-se uma tabela contendo os SN onde estes são identificados por meio de um código numérico. Não existe nenhum impedimento técnico por parte do sistema Access quanto ao uso do próprio texto dos SN como chave de acesso às informações. Deve-se ressaltar que, apesar da lentidão que este tipo de chave de acesso provoca, as estruturas de dados seriam mais simples e fáceis de manusear. Contudo, optou-se pela utilização das chaves numéricas identificando cada SN com o intuito de obter maior velocidade de acesso aos SN e às informações.

Finalizando esta apresentação do modelo de Kuramoto, cabe destacar que a utilização da árvore de SN por níveis permite uma visualização mais fácil do conteúdo da base de dados, e mantém o que há de mais significativo nos documentos: sua semântica.

As estruturas de *Qualia* e de Herança Lexical do Léxico Gerativo de Pustejovsky, a serem apresentadas na próxima seção, permitem também da mesma forma, considerar a semântica dos itens lexicais através da criação de uma malha/rede de relações de palavras e seus significados, através dos papéis que compõem a EQ.

3.2 A Teoria do Léxico Gerativo de Pustejovsky

Pustejovsky defende a idéia de que assim como a gramática tem uma estrutura (sintaxe), a semântica (significado) também tem uma estrutura básica. Na estrutura básica da sintaxe das línguas em geral, segundo Souza e Silva (1993), as orações são compostas de Sintagma Nominal (SN) mais Sintagma Verbal (SV), basicamente. Na busca da estrutura semântica, Pustejovsky (1991) delineia a teoria do Léxico Gerativo (LG) como uma abordagem na área da semântica lexical que pretende dar conta da criatividade semântica do uso das palavras em contexto.

Segundo Rossi (2003), Ullmann concorda com essa dificuldade do uso das palavras em contexto quando declara que "não são raros os casos em que ocorre uma polivalência das palavras, acarretando, por conseqüência, fenômenos semânticos inerentes às línguas naturais, entre eles, a ambigüidade lexical". Essa ambigüidade é provocada em decorrência de fatores

lexicais denominados de polissemia e de homonímia, ou, na terminologia de Weinreich, conforme Rossi (2003), de ambigüidade complementar e ambigüidade contrastiva, respectivamente.

No primeiro caso, trata-se da polissemia, que, de um modo geral, conforme Moura (2001), "é definida como um fenômeno que permite associar a um mesmo item lexical mais de um sentido, os quais mantêm alguma relação semântica entre si". Assim, a palavra "livro", por exemplo, é polissêmica, pois expressa ao menos dois sentidos diferentes que possuem entre si algum tipo de laço semântico: (a) objeto físico e (b) informação.

Já, no segundo caso, o da ambigüidade contrastiva, trata-se de homonímia, definida por Pustejovsky como a situação na qual um item lexical é associado com ao menos dois sentidos diferentes e sem relação entre si. Desse modo, a palavra "manga", por exemplo, é uma palavra homônima, pois não há nenhuma relação semântica evidente entre os sentidos de "fruta" e "parte da blusa".

Segundo Rossi (2003, p. 14) Ullmann salienta que "é difícil, em casos particulares, determinar onde termina a polissemia e onde começa a homonímia, uma vez que não é fácil e nem sempre possível medir intuitivamente o grau de proximidade dos significados".

A polissemia lógica é denominada por Pustejovsky (1991) para restringir a ambigüidade complementar, abordada anteriormente, nos casos em que ocorre uma relação lógica, portanto previsível entre os sentidos de uma palavra polissêmica. Havendo mais de um sentido, é importante ressaltar que pode existir sobreposição desses sentidos em um mesmo contexto.

Além de ter sido tratada como polissemia lógica por Pustejovsky, segundo Rossi (2003), desde Weinreich esse fenômeno da complementaridade dos sentidos tem sido abordado como polissemia regular e polissemia sistemática.

A teoria do Léxico Gerativo (LG) de Pustejovsky aponta o problema da multiplicidade de significados das palavras e enfatiza um tratamento relacionado ao problema da polissemia das palavras. Segundo Neto (2003), nessa perspectiva, Pustejovsky desenvolveu o LG, que é um modelo de processamento de língua natural que trata da explicação semântica de itens lexicais, tanto isolados quanto em contexto.

Assim como a gramática caracteriza o comportamento sintático específico de uma certa categoria de palavras, Pustejovsky propõe uma teoria gerativa do significado da palavra. E, ainda, pretende mostrar que seu modelo, segundo Rossi (2003, p. 47), "é contrário a

estaticidade presente em duas concepções semânticas teóricas das décadas de 60 e 70: as baseadas em redes conexionistas e as baseadas em primitivos fixos[14]".

Rossi (2003, p. 47) afirma que a teoria de redes conexionistas organiza a semântica das palavras através de relações e elos, para esta autora isso "dificulta a representação de sentidos que exibem polissemia regular, haja vista a distância na rede entre os sentidos que mantêm relação sistemática entre si". Por exemplo, os sentidos de "objeto físico" e "informação" são naturalmente distantes, no entanto, mantêm entre si relação sistemática no caso de "livro" e de outras palavras.

Já, no segundo caso, o das teorias baseadas em primitivos semânticos fixos, o léxico é tratado como uma lista enumerativa de sentidos. Por isso mesmo tais modelos são denominados por Pustejovsky (1991) de *Sense Enumeration Lexicon (SEL)* - léxico de enumeração de sentidos. O problema, segundo Pustejovsky (1991), é que essa caracterização dos possíveis sentidos de uma palavra postulada pelo modelo SEL é aplicada tanto para a ambigüidade contrastiva como para a polissemia lógica.

Fica evidente, segundo Rossi (2003), que Pustejovsky se opõe aos modelos SEL, pois apesar de eles proverem uma enumeração exaustiva dos sentidos de um item lexical, ainda se mostram limitados, não dando conta dos objetivos básicos da teoria semântico-lexical, ou seja, o uso criativo de palavras, a permeabilidade dos significados e as múltiplas formas sintáticas das expressões.

O objetivo principal do LG, segundo Pustejovsky (1991), é prover uma descrição formal da língua que seja expressiva e flexível o suficiente para apreender a natureza gerativa da criatividade lexical e extensão de sentido. Caracteriza assim o LG como um sistema semântico de perspectiva lógica que envolve quatro níveis de representação, um sistema de tipos semânticos e três tipos de mecanismos gerativos.

No decorrer deste capítulo, serão especificadas as noções teóricas básicas do modelo gerativo de Pustejovsky que estruturam o léxico em quatro níveis de representação (argumentos, eventos, *qualia* e herança) sobre os quais atuam dispositivos gerativos (a coerção de tipo, a co-composição e a ligação seletiva).

3.2.1 Estruturas do Léxico Gerativo

[14] Conforme Pustejovsky (1995), a teoria de primitivos fixos é defendida por autores como Lakoff (1971), Wilks (1975), Schank (1975), Katz (1977). Já a teoria de redes conexionistas é defendida por Carnap (1956), Collins e Quillian (1969), Fodor (1975), Brachman (1979).

Para capturar o significado lexical, estudou-se as estruturas de Pustejovsky (1991) que propõe quatro níveis de representação: estrutura de argumento, estrutura de evento, estrutura de *qualia* e estrutura de herança lexical descritos abaixo.

3.2.1.1 Estrutura de Argumento

Para Pustejovsky (1991) essa estrutura é uma especificação mínima que agrupa os itens lexicais em quatro argumentos:

- ***verdadeiros*** – parâmetros do item lexical que têm a necessidade de serem expressos sintaticamente. Ex.: Marta morou em Paris;
- ***apagados*** – parâmetros que não têm necessidade de serem realizados sintaticamente, são argumentos opcionais. Ex.: Joana coseu uma saia sem linha;
- ***sombreados*** – parâmetros que já estão semanticamente presentes no item lexical e só devem ser expressos através de operações de subtipo ou especificação de discurso. Ex.: Paulo salgou a carne com sal grosso;
- ***adjuntos verdadeiros*** – parâmetros que, mesmo sendo parte da interpretação situacional, modificam uma expressão lógica sem, contudo, estarem ligados à representação semântica de algum item lexical específico. Esses parâmetros introduzem expressões adjuntivas de modificação temporal ou espacial. Ex: David dormiu cedo.

3.2.1.2 Estrutura de Evento

Essa estrutura para Pustejovsky (1991) refere-se a organização de um conjunto de eventos no que tange à ordenação temporal de seus subeventos e a designação de qual deles será considerado o principal em relação ao evento matriz.

- ***Evento de estado –*** aquele cujo(s) argumento(s) não sofre(m) alteração durante o intervalo temporal do evento. Ex.: Kátia mora em Florianópolis.
- ***Evento de processo –*** aquele cujo(s) argumento(s) sofre(m) alteração de estado ou indica(m) o início de alguma atividade sem uma culminação precisa. Ex.: Heloisa canta bem.
- ***Evento de transição*** - aquele cujo(s) argumento(s) sofre(m) alguma ação de temporalidade determinada e resulta(m) em um estado diferente do inicial. Ex.: Tereza fez uma boneca.

A estrutura, a seguir, apresenta os atributos semânticos essenciais dos itens lexicais (como, por exemplo, a categoria, a composição, a função e a origem) através dos papéis: formal, constitutivo, télico e agentivo. É a estrutura principal responsável pela explicação da polissemia lógica abordada no texto (Pustejovsky, 1991).

3.2.1.3 Estrutura de *Qualia*

Devido a sua proximidade com o SN, visto que trabalha por conceitos (nomes), esta estrutura foi utilizada no desenvolvimento do modelo proposto pela pesquisa. Trata de um conjunto formado por quatro *qualia* que visam guiar o processo de entendimento a respeito de um objeto ou uma relação no mundo, dando, por conseqüência, um modo de especificar a denotação de tal objeto ou relação. É dividida em quatro papéis, os quais são descritos na seqüência.

a) Quale formal - faz a distinção de determinado item dentro de um domínio maior levando em consideração sua:

- orientação;
- magnitude;
- forma;
- dimensão;
- cor;
- posição.

b) Quale constitutivo - estabelece a relação entre um objeto e suas partes constituintes ou próprias a partir das propriedades:

- material;
- peso;
- partes e elementos componentes.

Além disso, o *quale* constitutivo informa também de que classe um item é parte, caso haja tal relação, ou seja, ele informa tanto uma relação de hiperonímia[15] quanto de meronímia[16].

[15] Hiperonímia: ocorre quando o significado de um lexema (palavra) abrange o significado de outro lexema. O significado de um é mais genérico que o significado de outro. Por exemplo, “aeronave” é um hiperônimo de “teco-teco”.

Estas relações são utilizadas na modelagem do protótipo de Gonzalez (2005). Este protótipo foi pesquisado e apreendido como parte integrante da proposta desta dissertação e será descrito na seção 3.3 do capítulo 3.

c) Quale télico - explicita a finalidade e a função de um objeto.

- Propósito de um agente ao realizar um ato.
- Função interna ou objetivo que descreve certas atividades.

d) Quale agentivo - determina os fatores envolvidos na origem ou causa de um objeto, partindo de considerações sobre:

- criador;
- artefato;
- tipo natural;
- cadeia causal.

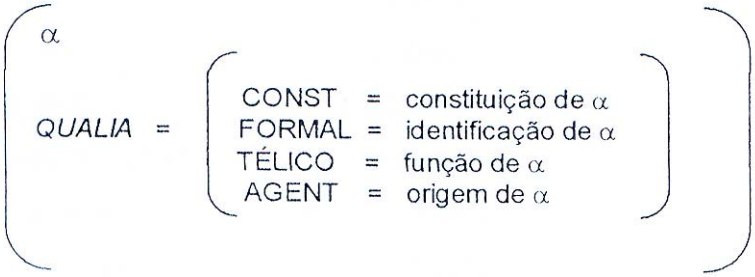

Figura 11: Representação da matriz de um item lexical
Fonte: ROSSI, 2003.

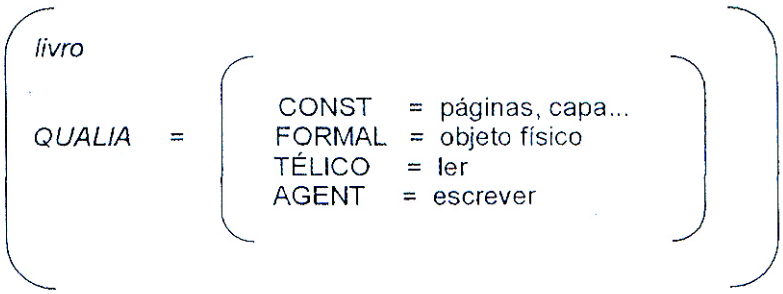

Figura 12: Matriz superficial da Estrutura de Qualia do item lexical "livro"
Fonte: ROSSI, 2003.

[16] Meronímia: ocorre quando o significado de um lexema (palavra) faz parte ou é uma porção do significado de outro lexema. Por exemplo, as palavras "capa" e "folhas" são merônimos de "livro".

É importante salientar que, segundo Neto (2003a), a Estrutura de *Qualia* não deve ser considerada apenas como uma lista de fatos interessantes sobre um item lexical e sim como um conjunto de propriedades que leva a uma explicação mais clara de tal item.

Isto equivale dizer que o objetivo da Estrutura de *Qualia* é abarcar o significado de uma palavra e explicitar como se relaciona com o uso da língua. Assim, essa estrutura salienta a explicação do uso da criatividade lingüística contextual não como uma estrutura isolada, mas em conjunto com os mecanismos gerativos que serão apresentados mais adiante.

Seguem alguns exemplos da Estrutura de *Qualia*.

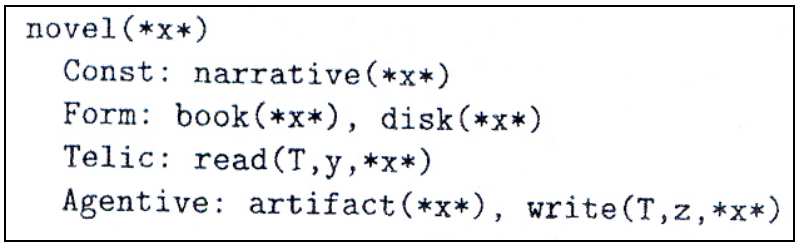

Figura 13: Exemplo da Estrutura de Qualia do item lexical “romance”
Fonte: PUSTEJOVSKY, 1991.

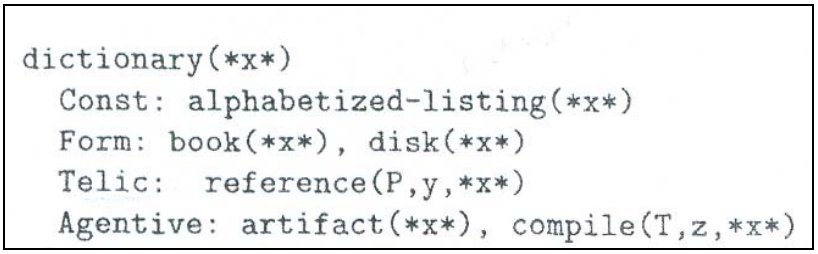

Figura 14: Exemplo da Estrutura de Qualia do item lexical “dicionário”
Fonte: PUSTEJOVSKY, 1991.

3.2.1.4 Estrutura de Herança Lexical

Esta estrutura também é de fundamental importância porque nesta ocorre a relação das qualias, ou seja, são estruturas lexicais que podem se organizar com outras estruturas em uma grade de tipo e assim ajudar na organização geral do léxico. Por exemplo, na figura abaixo, o LG relaciona “dicionário”, “livro” e “peça” através de suas estruturas de *qualia* em que se observa que os três itens lexicais são diferentes entre si, no entanto, mantêm relações semânticas.

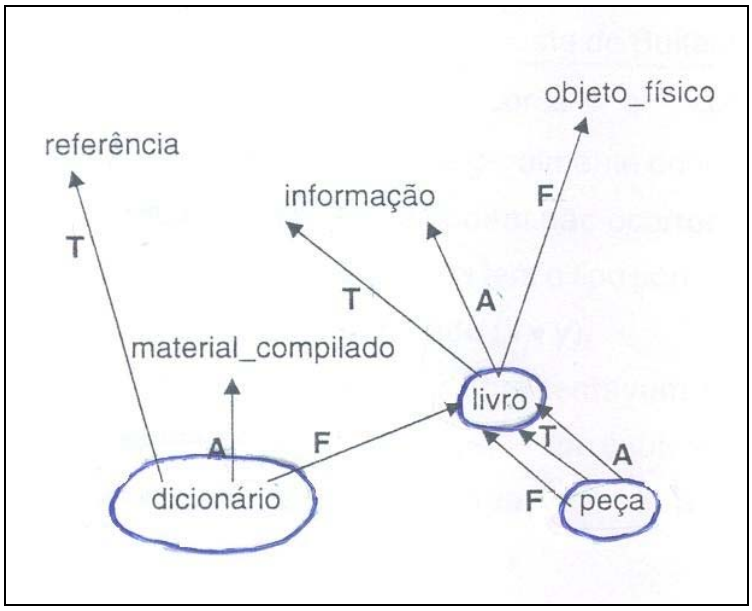

Figura 15: Exemplo do LG relacionando "dicionário", "livro" e peça através de suas EQ
Fonte: NETO, 2003a.

3.2.2 Sistema de Tipos Semânticos

Um sistema de tipos semânticos analisa o comportamento polissêmico e lógico de nomes implicitamente relacionais como, por exemplo, porta, janela. Pustejovsky mostra como o léxico gerativo faz uso de estruturas de aspectos típicos e afirma que esses nomes têm dois sentidos relacionais ("objeto físico" e "abertura") que são logicamente parte do significado do nome. Essa habilidade que um item lexical tem de agrupar vários sentidos é chamada "paradigma léxico-conceptual (plc ou lcp)". O plc é como um construtor de tipo, por exemplo, em palavras como "porta", e_1 significa objeto_físico, e_2 abertura, e o tipo resultante é "objeto_físico.abertura_plc = {objeto_físico.abertura, objeto_físico,abertura}".

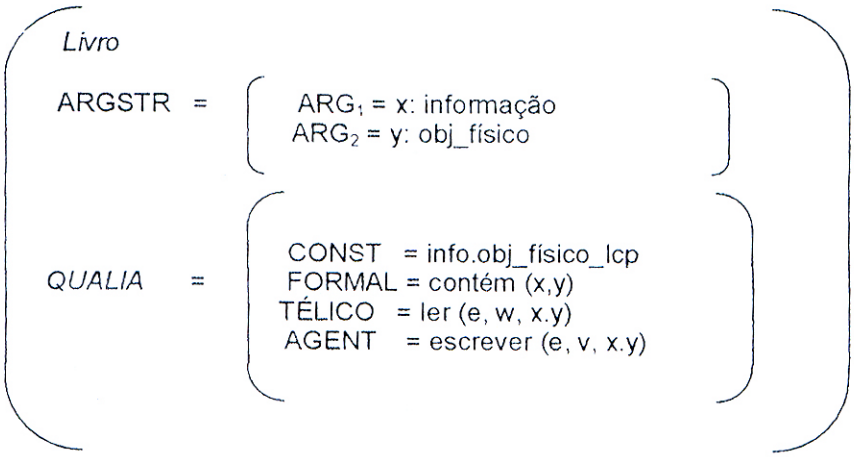

Figura 16: Exemplo de polissemia lógica na representação matricial da palavra "livro"
Fonte: ROSSI, 2003.

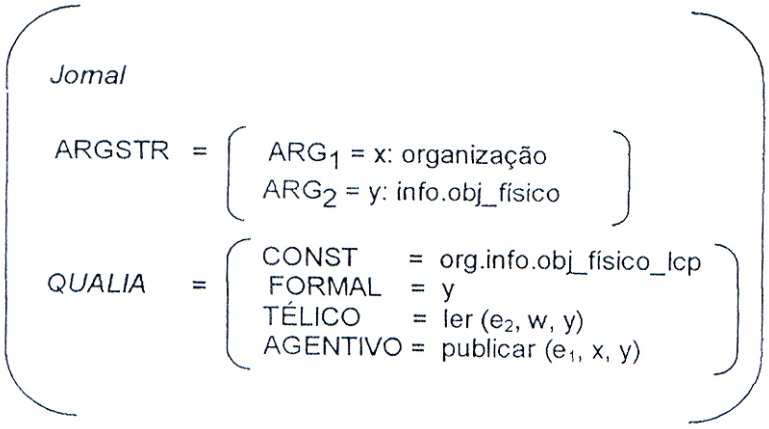

Figura 17: Exemplo de polissemia lógica na representação matricial da palavra "jornal"
Fonte: ROSSI, 2003.

3.2.2 Mecanismos gerativos

O Léxico Gerativo apresenta ainda um conjunto de três mecanismos que fazem uso das estruturas "evento", "argumento" e "*qualia*", os quais são ditos gerativos, pois relacionam diferentes itens lexicais possibilitando a interpretação composicional de palavras em contexto.

3.2.2.1 Coerção de tipo

Autoriza a mudança de tipo e, por extensão, de denotação de nomes e expressões de acordo com o contexto a que pertencem. A coerção de tipo reconstrói a semântica do complemento e só terá sucesso se o item lexical em questão tiver um atalho para o tipo desejado. O exemplo clássico dado por Pustejovsky é "João começou um livro", em que o predicado *começar* requer um tipo diferente do apresentado por livro, ou seja, o verbo requer um complemento do tipo "evento" que não é satisfeito por "livro". O termo "começar um livro" é interpretado como começar a ler (ou escrever) um livro.

3.2.2.2 Ligação seletiva

Rege a relação semântica que um modificador tem com o seu núcleo, ou seja, ela trata do problema da polissemia adjetival, uma vez que os adjetivos são interpretados a partir da semântica do núcleo. Exemplos:

(1) Um passeio rápido
(2) Um motorista rápido

(3) Um digitador rápido
(4) Um computador rápido

O primeiro problema está claramente exemplificado com (1) em oposição a (2), (3) e (4), ou seja, o primeiro trata de uma adjetivação sobre um evento e os demais, de uma adjetivação sobre indivíduos. Já para o segundo problema diz-se que a interpretação do adjetivo vai ser selecionada por algum dos *qualia* do núcleo do sintagma nominal, ou seja, pela ligação seletiva. Esse mecanismo vai buscar a interpretação de *rápido*, para os exemplos acima, no *quale* télico dos núcleos.

3.2.2.3 Co-composição

Os itens lexicais componentes de um determinado sintagma influenciam-se mutuamente, e um complemento pode adicionar um sentido ao seu núcleo. Pustejovsky começa exemplificando esse mecanismo com a polissemia de verbos como o "assar" que apresenta dois sentidos: uma mudança de estado e outra de criação do objeto. Os exemplos clássicos são:
(a) Letícia assou as batatas.
(b) Letícia assou o bolo.

Observa-se que em (1) houve apenas uma mudança de estado, pois as batatas já existiam antes de serem assadas; em (2) um sentido de criação de objeto é atribuído ao verbo, uma vez que antes da assadura o bolo não existia. Contudo, Pustejovsky (1991) afirma que ordinariamente só há um sentido para "assar", o de mudança de estado, pois tal verbo tem seu tipo de evento modificado devido a informações que são trazidas pelo complemento, ou seja, essas leituras só são possíveis a partir de mecanismo de co-composição em que os complementos co-especificam o verbo.

Por buscar formalizar a estrutura semântica de uma língua, o trabalho de Pustejovsky é de grande importância para a área de recuperação de informação. Uma tentativa de implementação computacional da sua teoria foi realizada por Abrahão (1997) envolvendo a modelagem e a implementação de um léxico semântico para a Língua Portuguesa. Inicialmente este autor realizou um estudo de conceitos básicos relacionados à semântica. Durante a sua pesquisa foram apresentadas técnicas de representação do conhecimento e do significado que auxiliaram a seleção e o entendimento do modelo proposto por Pustejovsky.

Como subsídio para a implementação de um léxico semântico para o português, Abrahão (1997) fez um estudo aprofundado da teoria de Pustejovsky onde salienta que os problemas mais comuns à representação do significado das palavras como "ambigüidade lexical polissêmica", por exemplo, são solucionados de forma eficiente e computacional.

Como o modelo de Pustejovsky é voltado ao Inglês foram encontradas semelhanças e diferenças entre a língua origem do modelo e o Português:

> Variações verbais - facilita o mapeamento direto, os verbos são inseridos numa forma canônica (básica ou infinitiva) no léxico; variações de grau nos substantivos como alternativa de solução são armazenados em uma forma canônica; palavras que se comportam como verbo e substantivo; palavras que se comportam como adjetivo e substantivo também são mapeadas através do uso da estrutura de lcps de Pustejovsky; mapeamento de expressões - expressões devem ser inseridas no léxico, pois expressam um significado específico; substantivos compostos por mais de uma palavra; acentuação – itens lexicais do Inglês não apresentam acentos. Esta característica do Português deve ser inserida no léxico, pois diferencia o significado de suas palavras. Deste modo, esta informação foi atribuída aos registros de informações semânticas através de uma variável que contém o tipo e a posição na palavra em que o acento aparece (ABRAHÃO, 1997, pgs 78-80).

Abrahão (1997) construiu sua implementação do léxico sobre uma estrutura em árvore *Trie*[17] que proporciona um maior poder de representação na busca de informações e baixa quantidade de dados armazenados. As informações semânticas associadas aos itens lexicais são armazenadas em listas encadeadas, a partir de uma estrutura denominada de Descritor Semântico. Um item lexical pertence ao léxico semântico se este item possui um Descritor Semântico associado ao seu último caractere na árvore. E, ainda, um Descritor Semântico abrange os ponteiros essenciais para a busca das informações semânticas relativas ao item lexical.

De acordo com este autor as informações semânticas associadas aos itens lexicais seguem o modelo de Pustejovsky (1991), sendo dividida em três estruturas básicas: de argumentos, de eventos e de *Qualia*. As estruturas de argumentos e de eventos são implementadas através de uma lista de argumentos e uma lista de eventos. A estrutura de *Qualia* é composta de quatro listas de informações, uma para cada papel (formal, constitutivo, télico e agentivo).

Segundo este autor, todas as estruturas do léxico semântico foram desenvolvidas em vetores. A manipulação destes vetores dá-se sobre estruturas denominadas cabeçalhos. Estes cabeçalhos fornecem informações sobre a alocação de vetores em memória, ponteiros para os vetores de informação, tamanhos dos vetores e os arquivos associados ao sistema. O núcleo de

[17] Segundo Abrahão (1997), "é um tipo especial de estrutura onde cada caractere dos itens lexicais determina um nodo da árvore".

dados do sistema é constituído de dois cabeçalhos: cabeçalho da árvore *Trie* e o cabeçalho das informações semânticas.

A biblioteca de funções contém os procedimentos necessários para manutenção do banco de dados lexical, bem como procedimentos de busca de informações semânticas. Juntamente com a biblioteca, uma interface gráfica foi construída possibilitando a manutenção do banco de dados e facilitando a visualização da semântica dos itens lexicais. Esta interface gráfica é implementada na linguagem de programação em C para as estações de trabalho *SUN* sobre o sistema de janelas *XVIEW*[18] (ABRAHÃO, 1997).

Esta seção mostrou a importância da teoria de Pustejovsky e suas possibilidades. O LG é fundamental para compreensão semântica, pois considera o contexto da palavra, sendo capaz de estruturar um domínio específico através da EQ e também de identificar dentro de um domínio quando determinada palavra aparece em tal contexto. Pelo desenvolvimento do trabalho de Abrahão pode-se perceber a dimensão e os elementos necessários para o significado de uma palavra, reforçando-se assim o valor e a viabilidade da teoria de Pustejovsky.

A próxima seção apresenta o trabalho de Gonzalez (2005), que estudou Pustejovsky[19] e posteriormente desenvolveu sua própria concepção de uma estrutura de RI (toda automatizada).

3.3 O Modelo TR+ de Gonzalez

O modelo TR+ é considerado um modelo para RI que utiliza duas fases para o desenvolvimento de sua estrutura: fase de indexação e fase de busca.

[18] *XVIEW* "é um sistema de janela orientado a objeto que permite ao programador criar e utilizar objetos tais como janelas, textos, painéis, ícones entre outros para construir uma aplicação. Seus objetos são predefinidos e são ricos em funcionalidade, o que permite que o código necessário para manipular essas janelas seja pequeno, simples e muito fácil de se compreender". (ABRAHÃO, 1997, p. 86)

[19] Realizou um trabalho individual no doutorado denominado "O Léxico Gerativo de Pustejovsky sob o enfoque da Recuperação de Informações", de 2000a.

Indexação de textos, segundo Baeza-Yates e Ribeiro-Neto (1999) e Gonzalez (2005) é o processo que estipula descritores[20] dos conteúdos dos textos de uma coleção de documentos, com objetivo de busca e classificação dos mesmos para atender consultas em sistemas de RI. Descritores podem descrever conceitos atômicos, sendo 'termos', ou conceitos complexos, sendo 'relacionamentos'. O conjunto de descritores concebido na indexação favorece uma visão lógica dos documentos, com o propósito de unir esses descritores, termos e relacionamentos, a conceitos presentes nos textos dos documentos.

Para os relacionamentos este autor classifica três tipos explicando-os através do exemplo "... têm preocupado os pesquisadores". O primeiro tipo é o par modificado-modificador como 'pesquisador-preocupado'. O segundo é o bigrama (preocupado, pesquisador) e o terceiro é o Sintagma Nominal que, para ele, significa 'pesquisador preocupado' e que para a pesquisa de Gonzalez ficaria na sua forma natural 'preocupado os pesquisadores'. O autor ainda cita que há outros formatos de relacionamentos como a expressão ternária (preocupação-de-pesquisador) e a relação binária (preocupação,pesquisador).

Gonzalez (2005) aponta dois tipos de relacionamentos como problemas: os bigramas por não poderem descrever o conceito ("ferro, sopa" para "panela de ferro com sopa") e os termos com palavras comuns, mas coadjuvantes importantes ("sentar, banco" e "depositar, banco"); os sintagmas nominais que para o autor representam tanto o conceito atômico quanto o complexo ("noite" e "boca da noite"). É importante perceber que, a partir dessas características e aspectos acima definidos, Gonzalez (2005) propôs um novo modelo de espaço de descritores (união do conjunto de termos com o conjunto de relacionamentos). Este novo modelo surgiu a partir de outros cinco modelos de descritores já existentes:

1. Unigrama: conjunto de termos não relacionados.
2. N-grama (NG): conjunto de relacionamentos estatísticos.
3. Termo-Termo (TT): conjunto de termos relacionados estatística ou sintaticamente.
4. Termo-Relacionamento (TR): conjunto de termos e relacionamentos sintáticos.

[20] A palavra *descritores* é usada para se tratar dos termos e relacionamentos enquanto os índices se referem apenas aos termos. O descritor 'termo' significa uma unidade lexical formada por uma única palavra ou por mais de uma denominada de 'termo composto'. E o descritor 'relacionamento' ocorre entre termos, ou seja, são relações de construções sintaticamente diferentes que têm o mesmo significado (semântica). Exemplo: 'defesa eficiente' é igual a 'defender eficientemente' e 'feira de domingo' é igual a 'feira dominical'. Alguns autores, como Baeza-Yates e Ribeiro-Neto (1999), utilizam a palavra 'índice' ao invés de descritores, contudo Gonzalez ressalta que esta palavra refere-se apenas aos 'termos' não dando conta da semântica que envolve os 'relacionamentos'.

5. Relacionamento-Termo (RT): conjunto de relacionamentos sintáticos e seus componentes. “Os Sintagmas Nominais constituem os principais descritores neste caso”. (GONZALEZ, 2005, p.41)

O modelo TR+ proposto por este autor combina aspectos dos modelos TR e RT.

A **Figura 18** dá uma visão geral do modelo TR+ de Gonzalez (2005), na fase de indexação com suas etapas essenciais e na fase de busca para a classificação por relevância dos documentos em relação à consulta.

Figura 18: Visão Geral do modelo TR+
Fonte: Gonzalez, 2005.

O espaço de descritores do modelo TR+ construído na fase de indexação é composto de quatro processos principais:

a) Pré-processamento (*toquenização* e etiquetagem);
b) Nominalização;
c) Captura de RLBs;
d) Termos e RLBs.

Na etapa “a” de pré-processamento ocorrem duas ações fundamentais: *Toquenização* e Etiquetagem. A *toquenização* é a identificação de cada item lexical (palavra e pontuação); Na etiquetagem existe um etiquetador gramatical (*part-of-speech tagger - parser*) que identifica, através de uma etiqueta (*tag*), a categoria gramatical de cada palavra do texto (adjetivo, substantivo, verbo entre outras). Geralmente é morfológico (identifica somente a

categoria morfológica) ou morfossintático (identifica também as funções sintáticas). Estes processos são realizados de forma automatizada[21].

Antes da nominalização é realizada a geração de espaço dos descritores que se constitui na: seleção e normalização dos descritores e, ainda a contagem de freqüência de ocorrência dos descritores - termos (para o cálculo de seus pesos), que será usada na etapa "d".

Faz parte do processo de seleção de descritores a eliminação de *stopwords*[22], que podem ser descartadas na fase de indexação e na consulta. Essa exclusão justifica-se, segundo o autor, porque as *stopwords* são consideradas palavras com pouca representatividade. A seleção dos descritores, a quantidade dos mesmos e o peso de cada um podem ser afetados pela normalização lingüística.

A normalização, segundo Gonzalez (2005), apresenta três tipos conhecidos como:

- ***Sintática*** - que transforma frases semanticamente equivalentes, mas sintaticamente diferentes ("eficiente processo rápido" e "processo rápido eficiente").
- ***Léxico-semântico*** – que utiliza relacionamentos semânticos (como a sinonímia) para substituir palavras morfologicamente distintas por uma única forma que representa o conceito evidenciado.
- ***Morfológica*** – reduz as formas flexionais de uma palavra por meio da conflação[23].

No modelo TR+ foi utilizada a normalização lexical para o processo de nominalização. Este processo de nominalização constitui a etapa "b" e significa a transformação de uma palavra (advérbio, adjetivo ou verbo), existente no texto, em um substantivo semanticamente equivalente, constituído com regras válidas de formação de palavras (GONZALEZ, 2005).

A tabela abaixo mostra exemplos de termos nominalizados. Nesta etapa de nominalização é utilizada a ferramenta CHAMA[24].

---

[21] A ferramenta FORMA (*Toquenização* e Etiquetagem Morfológica) foi utilizada por Gonzalez. O autor cita o nome desta ferramenta no seu site http://www.inf.pucrs.br/~gonzalez/tr+/ Acesso em 14 de fevereiro de 2006.
[22] *Stopwords* são palavras como preposições, artigos e conjunções.
[23] Conflação são processos realizados por algoritmos que combinam a representação de duas ou mais palavras em um único termo. Há dois métodos mais comuns: *stemming*, que reduz a palavra para a parte fundamental semelhante ao radical; e lematização, que reduz a palavra variável à correspondente forma "canônica".
[24] A ferramenta CHAMA (nominalização de adjetivos, verbos e advérbios) foi desenvolvida por Marco Antonio Insaurriaga Gonzalez (doutor em Ciência da Computação pela UFRGS). Em sua tese de doutorado intitulada "Termos e Relacionamentos em Evidência na Recuperação de Informação", 2005.

palavra original	classe	substantivo abstrato	substantivo concreto
saltar	verbo	salto	saltador
emendado	particípio	emenda	emendador
puro	adjetivo	pureza	ε
facilmente	advérbio	facilidade	ε
oval	adjetivo	ε	ε
fluvial	adjetivo	ε	rio
jovem	adjetivo	juventude	jovem

ε = ausência de nominalização

Tabela 1: Exemplos de nominalização
Fonte: Gonzalez, 2005.

Devido às diferentes variações que a nossa Língua Portuguesa apresenta, este autor trabalha em seu modelo com palavras sem acentuação e em letras minúsculas, ocorrendo um comprometimento do significado das palavras como, por exemplo, é citado por ele, pública e publica.

A etapa "c" de captura de Relações Lexicais Binárias (RLBs) é, segundo Gonzalez (2005), o relacionamento entre termos nominalizados, ou seja, sintaticamente diferentes, mas semanticamente iguais[25]. Uma RLB pode ser classificada, também, quanto à nominalização de seus componentes. Este autor sistematiza e classifica esta questão conforme aparece em seus exemplos abaixo (2005, p. 47):

- ***Original***, onde o termo não recebeu o processo de nominalização.

Exemplos: calma do local $\xrightarrow{rlb}$ de(calma,local)
rapidez da saída $\xrightarrow{rlb}$ de(rapidez,saida)
representação do ator $\xrightarrow{rlb}$ de(representação, ator)
cantor Zeviola $\xrightarrow{rlb}$ =(zeviola,cantor)

- ***Derivada***, onde um dos termos, pelo menos, resulta do processo de nominalização.

Exemplos: local calmo $\xrightarrow{rlb}$ de(calma,local)
saiu rapidamente $\xrightarrow{rlb}$ de(rapidez,saida)
ator representou $\xrightarrow{rlb}$ de(representação, ator)
Zeviola cantou $\xrightarrow{rlb}$ =(zeviola,cantor)

Uma RLB, de acordo com Gonzalez (2005), apresenta a seguinte aparência:

[25] Gonzalez desenvolveu o software RELLEX para o reconhecimento de relações lexicais binárias em sua tese de doutorado, 2005.

id (*t*1,*t*2) onde:

id significa o identificador de relação e,

*t*1 e *t*2 são os termos nominalizados.

Este autor aponta os três tipos de RLBs, quanto ao identificador *id*:

- ***Classificação***: onde *id* é especificado com um sinal de igual (=), *t*1 representa uma subclasse ou uma instância de *t*2 e *t*2 representa uma classe.

  Exemplos: =(cao,animal)

  =(PET, garrafa). Exemplo desenvolvido nesta dissertação.

- ***Restrição***: onde *id* é uma preposição, *t*1 representa um elemento modificado e *t*2 representa um elemento modificador.

  Exemplos: de(equipe,atletismo)

  com(supervisor,experiencia)

  por(orientacao,ministro)

- ***Associação***: onde *id* representa um evento, *t*1 é um sujeito e *t*2 é um objeto (direto ou indireto) ou um adjunto.

  Exemplos: superacao(aluno,dificuldade)

  interesse.a(proposta,negociante)

  moradia.em(presidente,brasilia)

As Relações Lexicais Binárias, conforme Gonzalez (2005), são inseridas no espaço de descritores para ampliar o seu universo. As RLBs descrevem relações semânticas lexicais como as que são apresentadas na estrutura de *Qualia* da teoria do Léxico Gerativo de Pustejovsky (GONZALEZ, 2000, PUSTEJOVSKY, 1991). O estudo desta teoria motivou o Gonzalez a desenvolver a proposta das RLBs como parte integrante de seu trabalho como um modo de adequá-la a aplicações na área de RI.

Como já foi descrita na seção 3.2, a Estrutura de *Qualia*, da teoria do Léxico Gerativo, descreve um item lexical através de quatro papéis: formal, constitutivo, agentivo e télico. O ***papel formal*** distingue um item lexical em um domínio maior. Em uma RLB, segundo Gonzalez (2005), do tipo classificação, como "=(computador,maquina)" por exemplo, o computador seria distinguido como uma máquina ou, em "=(ipmf,tributo)", o ipmf seria um tributo. Portanto a RLB do tipo classificação corresponde ao ***papel formal*** da estrutura de *Qualia*.

O ***papel constitutivo*** estabelece a relação entre um item lexical X e suas partes constituintes. Em uma RLB do tipo restrição, como "de(mesa,madeira)" por exemplo, haveria a indicação de que a mesa é feita de madeira ou, em "com(massa,alho)", de que há alho na massa. O ***papel agentivo*** especifica os fatores envolvidos na origem ou causa de um item lexical. Em uma RLB, para este autor, do tipo restrição, como "por(publicacao,autor)" por exemplo, seria especificado que a publicação se deve ao autor ou, em "por(impedimento,lei)", que a lei é a razão do impedimento.

O ***papel télico*** explica qual a função ou finalidade do item lexical. Em uma RLB do tipo associação, como "conserto(encanador,vazamento)" por exemplo, explica que a função do encanador é o conserto do vazamento ou, em uma RLB do tipo restrição como "para(leitura,aprendizado)", que a finalidade da leitura é o aprendizado (GONZALEZ, 2005).

Este autor salienta que não se quer que as RLBs "interpretem" o texto com distinções, indicações, especificações ou explicações dos tipos apresentados. O propósito é de que as RLBs sejam descritores de tais fatos, mas sem classificação (etiquetas). Por isto os identificadores de relação não são rotulados com os papéis descritos. A única exceção é o identificador das RLBs do tipo classificação. O indicador "=" é o rótulo inevitável para o clássico "é um" porque não há outro papel possível nesse tipo de relação.

No modelo TR+, está envolvido além da coleção de documentos constituída por descritores (termos e relacionamentos) também os seus respectivos pesos que dependem de uma formulação matemática denominada de 'cálculo de representatividade' dos descritores em cada documento, que é um diferencial deste modelo e está na fase "d" onde os termos e RLBs serão armazenados.

Para ocorrer o cálculo do peso dos descritores é aplicado o conceito de evidência[26]. Este conceito não depende apenas da freqüência de ocorrência de um descritor, mas de um outro mecanismo: "a representatividade de um descritor depende, além de sua freqüência de ocorrência no texto, da ocorrência de mecanismos de coesão frásica" (GONZALEZ, 2005, p.48). A **coesão frásica** determina uma **junção** significativa entre os componentes de uma **frase**[27]. Esta junção aliada com a freqüência de ocorrência constitui o conceito de evidência como um dos aspectos essenciais da Tese de Gonzalez (2005).

A evidência dos termos é realizada de forma direta com a freqüência e a coesão frásica, mas a evidência de um relacionamento não, pois esta é dependente primeiramente das

[26] Evidência significa qualidade daquilo que é evidente, que é incontestável, que todos vêem ou podem ver e verificar (Dicionário Eletrônico Michaelis). Como descreve Gonzalez (2005) "é aquilo não oferece ou não dá margem à dúvida".
[27] Site http://acd.ufrj.br/~pead/tema09/coesaogramatical.html

evidências de seus termos. Este conceito está inserido no cálculo de representatividade de um descritor.

O cálculo da representatividade é um cálculo de relevância do termo ou relacionamento que varia de acordo com as abordagens (booleana, vetorial e probabilística) e pode ser realizado apenas com a freqüência da palavra no documento, ou ainda, com a freqüência vinculada com a sua informação morfológica ou sintática (GONZALEZ, 2005).

Para realizar o cálculo da representatividade dos descritores há duas estratégias de determinação que são: os modelos com unigramas, que tratam os termos de forma independente (abordagens vetorial e probabilística) e os modelos com dependência entre termos. Estas dependências envolvem conjuntos diferentes de conhecimentos que são os estatísticos e os lingüísticos[28]. Os conhecimentos lingüísticos são "léxico, morfológico, fonológico, sintático, semântico e pragmático" (ABRAHÃO, 1997, p.11).

Estes dois modelos descritos acima são apresentados como mais significativos, porém ainda utilizam a abordagem booleana. Isto porque Gonzalez (2005) define como o caminho mais promissor a combinação da abordagem booleana (individualmente limitadora) com a união dos conhecimentos estatísticos e lingüísticos entre si, que permitem mais interação com o usuário.

O cálculo da representatividade, ao mesmo tempo, que é uma propriedade básica de um descritor apresenta diferentes formas de acordo com as abordagens vetorial e probabilística (capítulo 2) e gera diversas interpretações. Por isto, Gonzalez (2005) propõe um novo cálculo que compreenda a importância do contexto nas fórmulas inseridas no seu modelo TR+.

O outro momento de seu modelo (Figura 18) compreende a 'fase de busca' que inclui Pré-Processamento (*toquenização* e etiquetagem), Nominalização e Captura de RLBs. Estas etapas ocorrem da mesma maneira que na fase de indexação. Inclui também as etapas: Formulação de consulta booleana, Busca e Classificação.

Na etapa "e" (Formulação de Consulta Booleana), Gonzalez (2005) explica que se a consulta *q*, em linguagem natural, formulada pelo usuário for, por exemplo, "pintura restaurada", então será formulada no formato Booleano, conforme o modelo TR+, a seguinte consulta *qb*:

[28] Estes conhecimentos envolvem níveis léxico-morfológico e sintático, sintagmas nominais (sujeito, objeto direto e indireto e adjunto adnominal). A vantagem destes é a capacidade de identificar relacionamentos entre palavras não adjacentes, como "algoritmos" e "concorrentes" em "algoritmos seqüenciais e concorrentes".

*r*1 *OU* *r*2 *OU* **((**n1 (*p*1) *OU* n2(*p*1) **)** *E* **(**n1 (*p*2) *OU* n2(*p*2) **))** onde: *r*1 = de(restauracao,pintura), *r*2 = *r*1' = diferente_de(restauracao,pintura), n1(*p*1) = (elemento vazio), n2(*p*1) = pintura, n1(*p*2) = restauracao, n2(*p*2) = restaurador, *p*1 = pintura, e *p*2 = restaurada.

Tabela 2: Exemplo de uma consulta *qb*
Fonte: Gonzalez, 2005, p. 51.

Na fase de busca, a etapa "f", ocorre uma relação entre a etapa "e" e a etapa "d". Esta última acontece ainda na fase de indexação visto que "estando os termos e as RLBs definidas e calculados os pesos, a classificação dos documentos depende do valor de relevância dos mesmos e da formulação Booleana da consulta". (GONZALEZ, 2005, p. 50).

A etapa "g" (Classificação) é resultado de um cálculo, sobre os dados obtidos no procedimento anterior, que identifica o valor de relevância de cada documento recuperado-os em ordem decrescente. Um exemplo de classificação é indicado por Gonzalez (2005) através da fórmula de uma consulta denominada *q*. Nesta consulta encontram-se os termos *t*1 e *t*2 e a RLB *r* e, se estes dois termos estão relacionados através de *r* em um documento *d*, estes terão dupla contribuição no cálculo do valor de relevância de *d*, porém se *t*1 e *t*2 ocorrem em *d*, mas não estão relacionados através de *r*, o autor considera que esta contribuição será simples e, assim, *d* tende a perder posições na classificação por relevância a *q*.

Os documentos recuperados classificam-se em dois grupos:

> (a) grupo superior, de maior relevância: documentos que atendem às condições estabelecidas na consulta Booleana, ou seja, possuem pelo menos uma das RLBs da consulta ou, na falta de todas elas, possuem obrigatoriamente todos os termos conforme especificado;
> (b) grupo inferior, de menor relevância: documentos que não atendem a todas as condições estabelecidas na consulta Booleana, mas possuem pelo menos um dos termos da consulta.
> Os documentos são classificados em ordem decrescente do valor de relevância, tanto nos grupos superior como inferior. (GONZALEZ, 2005, p. 51)

É importante ressaltar que toda a proposta de Gonzalez (Modelo TR+) foi automatizada, testada e aprovada. Foi utilizado o software FORMA para a etapa de pré-processamento e os demais softwares como CHAMA (nominalização) e RELLEX (regras de

identificação de RLBs) foram desenvolvidos pelo autor. Diversos algoritmos juntamente com abordagens de RI (booleana, probabilística e vetorial) foram desenvolvidos para as fases posteriores do seu trabalho, como o cálculo do peso dos descritores, a busca e a classificação de documentos.

As experimentações desenvolvidas por Gonzalez (2005), em seu trabalho, lograram comprovar que: o processo de nominalização, como processo de normalização lexical, proporciona melhores resultados de recuperação que os produzidos pelos processos tradicionais (lematização e *stemming*); a identificação de RLBs (obtenção de informação lingüística) contribui de forma positiva para a descrição de dependências de termos, ampliando o espaço de descritores; o cálculo da representatividade dos descritores baseado em evidência melhora a classificação de relevância dos documentos, com vantagem sobre o cálculo baseado em freqüência de ocorrência; o uso de consultas com operadores Booleanos trata-se de uma forma eficaz de complementar a especificação de dependências de termos; e, também, a inclusão de conhecimento lingüístico, como a realizada no modelo proposto pelo autor, apresenta relação custo/benefício viável dentro do atual estágio de desenvolvimento da pesquisa em RI.

O próximo capítulo descreve o novo modelo proposto para esta dissertação baseado na identificação das possibilidades de ampliação, de síntese e de sistematização do modelo de Kuramoto com a estrutura de Gonzalez. Pode ser considerada uma solução híbrida de um modelo de RI que une três teorias: Sintagmas Nominais de Kuramoto, Léxico Gerativo de Pustejovsky e Modelo TR+ de Gonzalez. Apresentar-se-á os parâmetros gerais norteadores e justificadores do modelo, a descrição narrativa da sua funcionalidade, os resultados dos testes e a descrição formal UML do modelo.

4. APRESENTAÇÃO E DISCUSSÃO DO MODELO PROPOSTO

A proposta desta dissertação é de integrar a aplicação prática do projeto dos Sintagmas Nominais de Kuramoto sistematizando e associando com o modelo TR+ de Gonzalez (2005).

Na descrição do modelo do sistema proposto foi utilizado o método denominado de Processo Unificado (UP), que envolve as fases de concepção, elaboração, construção e transição e utilizou-se a Linguagem de Modelagem Unificada (UML), que é fortemente relacionada com a metodologia utilizada segundo Wazlawick (2004).

Neste capítulo desenvolve-se o modelo conceitual da aplicação proposta para a qual foram realizadas as etapas: de levantamento e análise de requisitos, representada pelo diagrama e pela descrição dos casos de uso, e de construção dos diagramas de classes e de seqüência relacionados.

4.1 Procedimentos desenvolvidos utilizando o modelo de SN de Kuramoto e a proposta Gonzalez - "Estrutura SINTR+"

Esta dissertação optou por realizar uma relação entre propostas diferenciadas: utilizar o modelo de SN de Kuramoto para a organização dos conceitos mais significativos dos documentos e a proposta de Gonzalez para a busca dessas informações que estarão estruturadas através da dependência entre termos. Esta relação foi desenvolvida na criação da "Estrutura SINTR+", que tem como especificidade a busca nos documentos, a partir do banco de dados dos Sintagmas Nominais. Esta escolha, de unir em uma estrutura própria os SN e o Modelo TR+, pautou-se pelo intuito de orientar mais objetivamente o usuário na definição da sua *query* de busca, através de uma navegação sobre a estrutura de SN presentes no documento e de posterior apresentação de lista de documentos efetivamente relevantes.

O objetivo é trabalhar com os Sintagmas Nominais evidenciando e potencializando uma união com o modelo TR+ de Gonzalez (2005). O modelo abaixo (**Figura 19**) apresenta uma nova proposta pautada na junção sistematizada e analítica da extração dos SN na Estrutura de Kuramoto (1999) com o Modelo TR+ de Gonzalez (2005): "Estrutura SINTR+".

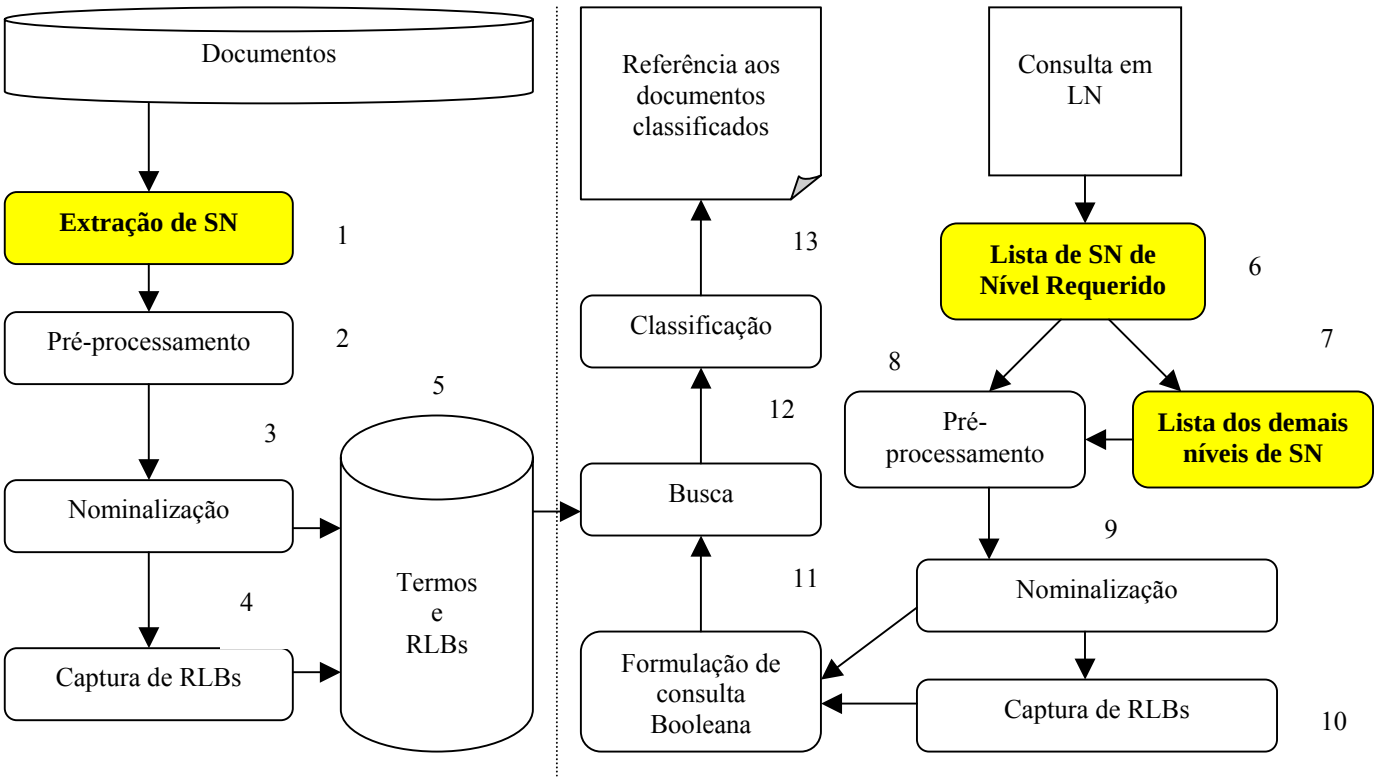

Figura 19: Visão Geral do Modelo Proposto "Estrutura SINTR+"

O modelo proposto se inicia a partir dos documentos a serem inseridos, com a extração de todos os seus Sintagmas Nominais (***Etapa 1***). Extraídos os SN, na ***Etapa 2*** é feito o pré-processamento onde ocorrem a *Toquenização* e a Etiquetagem que. Essa etapa ao invés de acontecer com todas as palavras do documento, como ocorre no modelo TR+, é realizada de forma mais objetiva e rápida, apenas diretamente sobre os termos constantes nos SN. O foco de análise, somente sobre os termos inclusos nos SN, permanece para todas as etapas subseqüentes.

Antes do processo de nominalização, na ***Etapa 3*** é executada a geração de espaço dos descritores constituída na: seleção e normalização dos descritores e, ainda na contagem da freqüência de ocorrência dos descritores - termos (para o cálculo de seus pesos), a ser usada na Etapa 5.

Em seguida, ocorre o processo de nominalização que constitui a Etapa 3 e significa a mudança de uma palavra (advérbio, adjetivo ou verbo), existente nos SN, em um substantivo concreto e/ou abstrato. Na ***Etapa 4*** ocorre a identificação das RLBs nos SN que significa o

relacionamento entre termos nominalizados. Estas etapas acima são constituídas para a geração do espaço de descritores (termos e RLBs) referentes à ***Etapa 5***.

Na 'fase de busca', primeiramente o usuário digita uma palavra, por exemplo, "plásticos". A resposta para o usuário ocorrerá, pois internamente foi feita uma programação (a ser implementada) para identificar o nível do SN solicitado pelo usuário para que posteriormente apareça para este a lista de todos os SN do nível apresentado, contendo a *query* solicitada.

No caso do exemplo "plásticos", o processo avança na ***Etapa 6,*** listando todos os sintagmas nominais de primeiro nível (SN1) dos documentos (uma vez que a solicitação referia-se ao nível 1). Nesta etapa o usuário poderá escolher um dos sintagmas de primeiro nível ou confirmar a sua escolha (*query*) inicial. O processo continua com a escolha de uma dentre as opções de: i) ver a lista de documentos relacionados ao SN1 definido, ou ii) solicitar a relação de sintagmas de seu segundo nível. A visualização da lista de sintagmas de nível superior permitiria ao usuário filtrar mais a sua consulta. Para a determinação da lista de SN de segundo nível como, por exemplo, "a reciclagem de plásticos", "a indústria de plásticos" (Figura 20) também foi feita uma programação específica que será descrita posteriormente.

Na continuidade do processo, o usuário pode prosseguir o refinamento da sua busca através da seleção de SN de maior nível ou pode dar-se por satisfeito com o resultado (***Etapa 7***), solicitando diretamente a lista dos documentos associados ao SN definidos. Nesse caso a lista é apresentada na ordem de classificação oportunizada pela Estrutura TR+, conforme o descrito nas próximas etapas.

O processamento proposto para a determinação da relação dos sintagmas de um determinado nível, foi pensado com vista a gerar economia de espaço de memória utilizada, uma vez que serão armazenados na base de dados os documentos e seus SN de últimos níveis e manipulados apenas os últimos níveis da estrutura de SN. Os níveis anteriores relativos ao SN serão determinados na programação desenvolvida, a partir da identificação do número de preposições que o SN apresenta. Nesta programação se houver apenas um termo (ou mesmo apenas um termo composto), o SN é considerado um SN de 1º nível. A presença de um termo composto com mais uma preposição, indica a existência de um SN de 2º nível. Já três termos com duas preposições vão indicar a presença de um SN de 3º nível e, finalizando, quatro ou mais termos com 3 (ou mais) preposições remetem ao SN de 4º nível.

Ao optar pela apresentação da lista de documentos, serão desenvolvidas (internamente) na programação, conforme o proposto pela Estrutura TR+ de Gonzalez, as etapas de Pré-processamento (*toquenização* e etiquetagem – ***Etapa 8***), Nominalização (***Etapa***

9), Captura de RLBs (***Etapa 10***), Formulação de consulta Booleana (***Etapa 11***), Busca (***Etapa 12***) e por fim Classificação (***Etapa 13***).

Na ***Etapa 11*** é trabalhado no formato Booleano uma consulta formulada pelo usuário, conforme o modelo TR+. A ***Etapa 12*** ocorre uma relação entre a ***Etapa 11*** e a ***Etapa 5*** (esta etapa ocorre ainda na fase de indexação). A ***Etapa 13*** é a última e resulta do cálculo que identifica o valor de relevância de cada documento recuperando-os em ordem decrescente.

É importante reforçar que o sistema irá verificar o pré-processamento, nominalização e a captura de RLBs já realizadas na fase de indexação, comparando-as. Após esta identificação, o sistema usa a formulação de consulta Booleana para a busca, chegando à classificação dos documentos de acordo com o peso dos descritores (termos e RLBs) formulados na fase de indexação e definidos na fase de busca (de acordo com o termo escolhido e a coleção dos documentos).

Exemplificando o parágrafo acima, a **Figura 20** mostra o funcionamento inicial desta estrutura no que se refere aos Sintagmas Nominais:

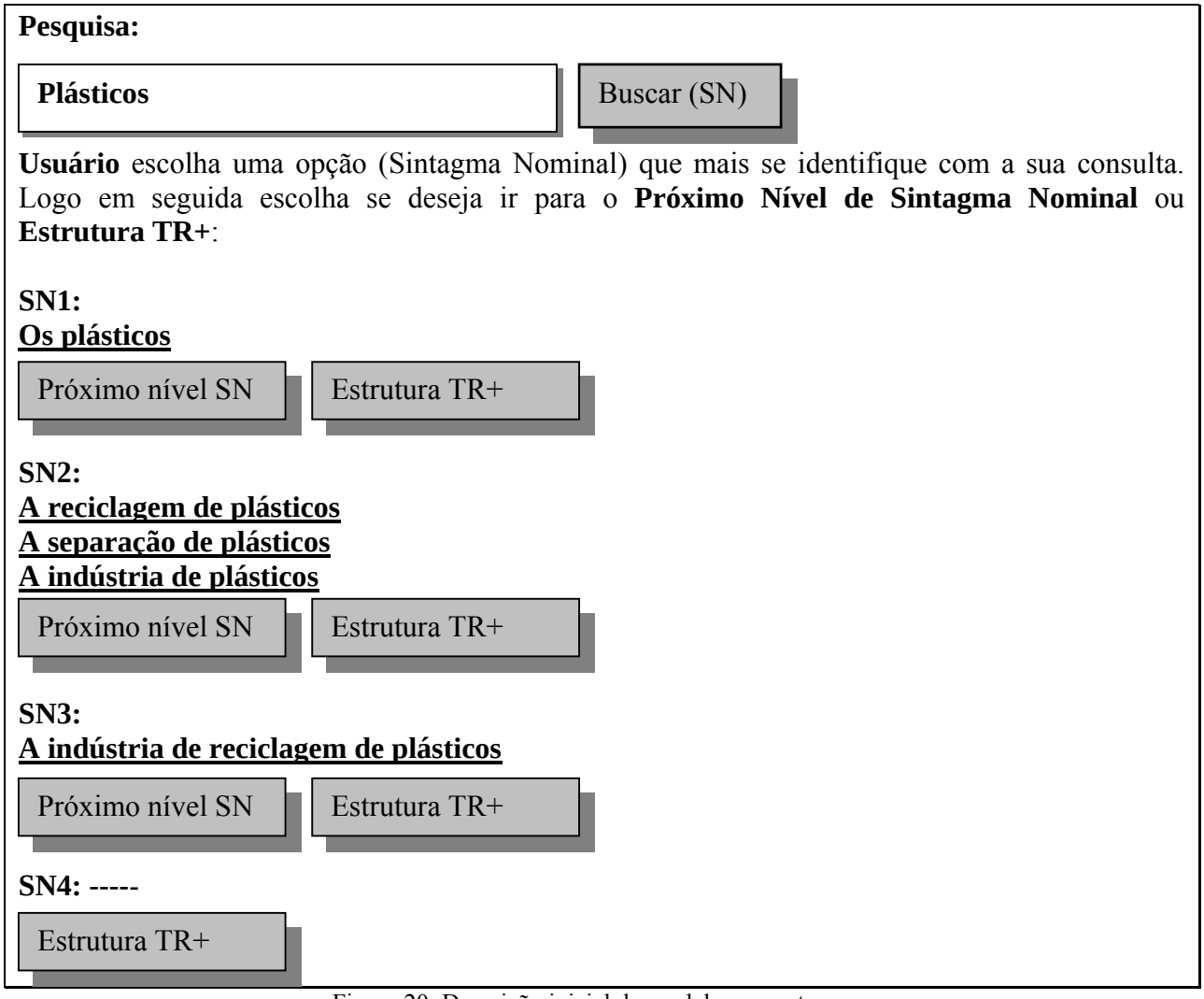

Figura 20: Descrição inicial do modelo proposto

Buscando analisar as vantagens que a proposta do modelo SINTR+ apresenta, vale lembrar que o modelo TR+ de Gonzalez já apresenta benefícios como:

- O processo de nominalização propicia melhores resultados de recuperação do que os produzidos pelos processos tradicionais (lematização e *stemming*);
- A identificação de RLBs colabora para a descrição de dependência de termos que ampliam o espaço de descritores;
- O cálculo da representatividade dos descritores baseado em evidência melhora a classificação da relevância de documentos em relação àquela obtida através da extração e do cálculo por freqüência de ocorrência;
- O uso de consultas com operadores Booleanos oferece uma forma eficaz de complementar a especificação de co-dependência semântica entre termos.

As vantagens antevistas na elaboração da proposta SINTR+ expandem as já obtidas pelo modelo de Gonzalez[29] pois une a elas a vantagem do modelo de hierarquia de níveis de SN de Kuramoto. Estas vantagens são: a "Estrutura SINTR+" executa em um menor tempo na fase de indexação dos documentos; a "Estrutura SINTR+" contém um tamanho menor de arquivos de índice; e a "Estrutura SINTR+" proporciona facilidade na fase de nominalização, visto que os SN são o núcleo de maior significação de um texto[30].

Os documentos (textos) usados como campo empírico desta dissertação foram artigos retirados da Internet sobre o tema "Lixo". Neste contexto, fazem parte da coleção de documentos temas como: "Cuidados com o Lixo", "Lixo Industrial", "O destino do lixo químico", entre outros. Como ainda não havia disponíveis extratores automáticos de SN por hierarquia em níveis, foi feita uma leitura dos textos dos quais se retirou manualmente seus sintagmas. Os SN significativos com o tema "Lixo" foram extraídos de dois (2) documentos (que estão nos ANEXOS A e B) e são apresentados no Anexo C.

Após esta etapa foram extraídos todos os sintagmas nominais (somente do documento1 - ANEXO A) que estão sublinhados no texto independentes do tema para exemplificar a extração da consulta.

Para avaliar preliminarmente a extensão com que as vantagens antevistas no modelo proposto realmente se verificariam, foi realizado um teste com o documento1 (ANEXO A) composto de 9 parágrafos e 1006 palavras (**Figura 21**).

---

[29] Este modelo foi testado e aprovado na sua proposta de doutorado que está inserida no contexto do grupo de pesquisa da PUCRS no qual o autor participa de estudos na área há mais de uma década.

[30] Isto pode ser observado do Anexo A (Documento1) em que os SN são destacados no texto.

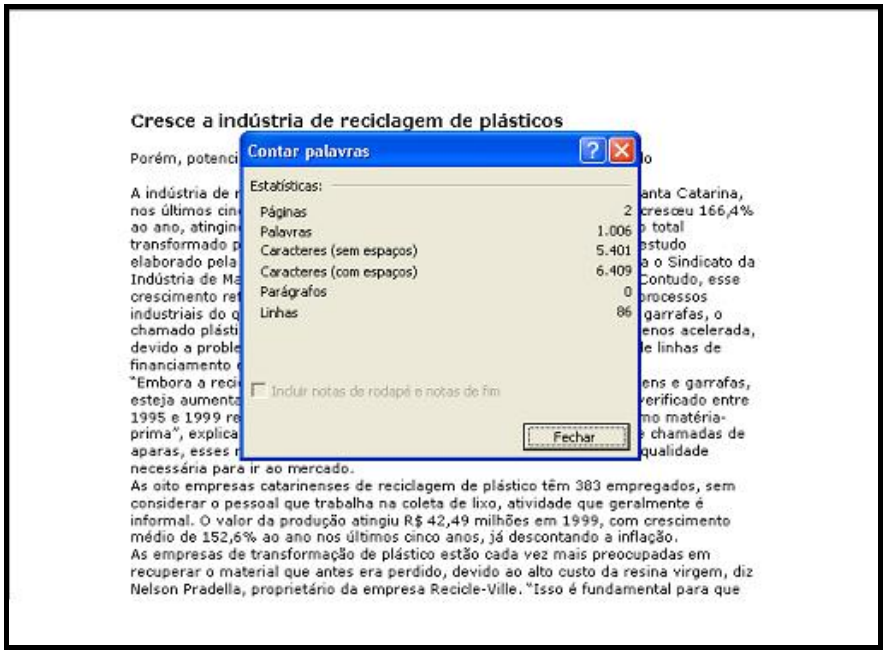

Figura 21: Número de palavras do Documento1

O documento1 **(ANEXO A)** foi o escolhido, para dimensionar a redução no total de palavras/termos a serem incluídos na base de dados, demonstrando a importância do modelo apresentado, conforme tabela abaixo:

Categorias	Texto Total	SNs
Total de palavras/termos	1006	640
Substantivos	369	334
Advérbios	41	04
Verbos	133	Ausência de verbos
Adjetivos	73	55

Figura 22: Tabela comparativa Texto Total e SNs

O texto possui um total de **1006** palavras/termos sendo destes **369** substantivos, **41** advérbios, **133** verbos e **73** adjetivos (**Figura 22**). Do texto todo foi extraído um total de **139** sintagmas nominais. E destes, o número total de palavras/termos é de **640**, sendo **334** substantivos, **04** advérbios e **55** adjetivos.

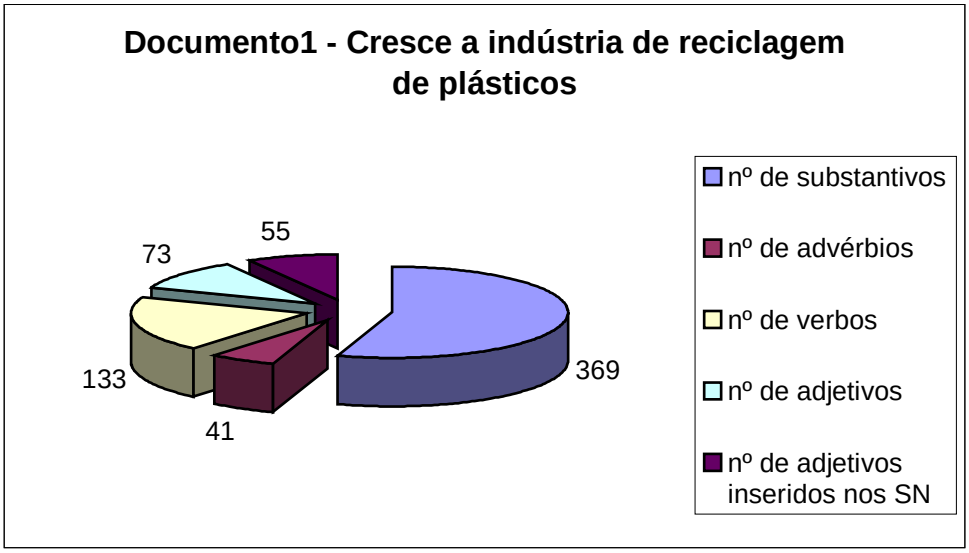

Figura 22: Número de substantivos, advérbios, verbos e adjetivos do Documento1

Relacionando o número de adjetivos do texto todo e os adjetivos inseridos nos SN, pode-se notar um ganho expressivo, pois se tem uma redução de 18 adjetivos. Destes dados, 133 verbos foram descartados (novamente afirma-se da importância dos SN que representam a unidade significativa do texto). Também se observa que 37 advérbios não foram incluídos, diminuindo assim o número de descritores.

Estes dados apontam aspectos positivos que consolidam a importância da utilização dos SN na diminuição de descritores, com conseqüente redução do uso de memória, e ainda, melhora na fase de busca pelo tempo de resposta.

A **Figura 23** apresenta o comparativo entre o percentual do número de palavras do texto com o percentual do número de palavras dos Sintagmas Nominais. Isto mostra que o percentual de SN de 64% tem um valor reduzido, colaborando para um número menor de descritores, desta forma, restringe-se também o uso de memória (neste caso ocupado na fase de indexação), reduzem-se os descritores e diminui-se o tempo de resposta na fase de busca. Estes dados não são somente relevantes frente a um modelo de RI, mas corroboram para a manutenção do seu funcionamento.

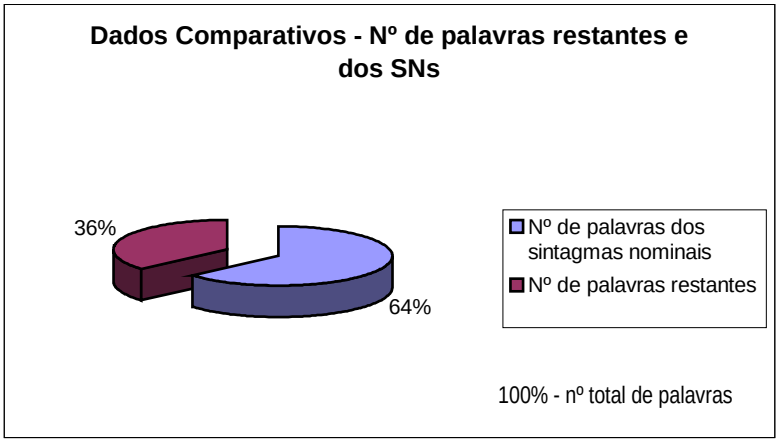

Figura 23: Número de palavras restantes x Sintagmas Nominais

A **Figura 24** mostra que existe um percentual de 28% de adjetivos inseridos nos Sintagmas Nominais. Esses adjetivos, durante o processo de nominalização, conforme Gonzalez (2005) são transformados em substantivos concretos e/ou abstratos (se houver). Isto aponta um número bem inferior comparado a um texto inteiro o que promove uma diminuição de substituições, de um adjetivo por um substantivo concreto e/ou abstrato, que pode inferir no significado do documento e a redução destas substituições evita possíveis erros de interpretação.

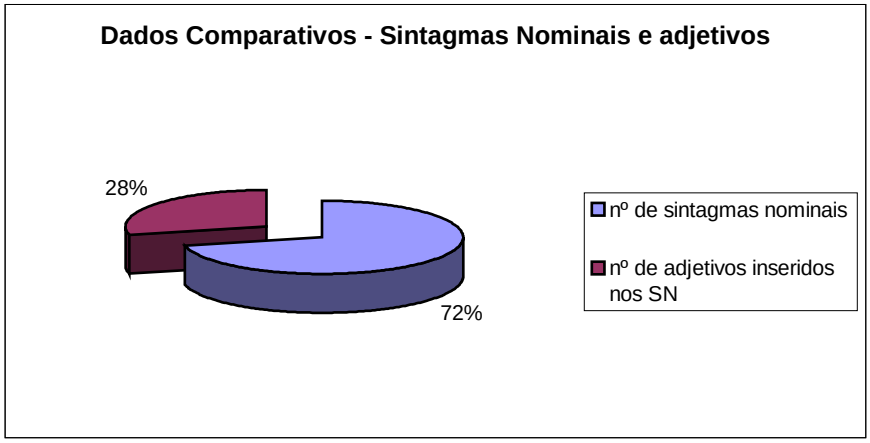

Figura 24: Sintagmas Nominais e adjetivos inseridos nos SN

A extração dos Sintagmas Nominais corresponde à primeira etapa. Depois desta extração manual se agrupou os SN em quatro níveis: 1, 2, 3 e 4 **(ANEXO C)**.

Para o desenvolvimento das demais etapas (*toquenização*, etiquetagem morfológica, nominalização e as relações lexicais binárias), foi escolhido o parágrafo 6 do documento1 **(ANEXO A)**.

A indústria da reciclagem do plástico no Brasil tem crescido bastante em função do reaproveitamento do PET, que é usado no segmento de monofilamentos, em artigos como vassouras e na indústria têxtil. Conforme Ana Flores, a reciclagem gera 250 mil empregos no País, dos quais 70% são informais. Porém, a maior parte do potencial de mercado ainda está sendo desperdiçada, avalia. "Cerca de 15% do total de plástico que é industrializado no País é reciclado. Em dez anos poderíamos chegar a 60%, como nos Estados Unidos, desde que fosse implementado um conjunto de medidas incentivando essa prática", assegura.

Tabela 3: Parágrafo 6 do documento1

Na etapa de *toquenização* e etiquetagem são identificadas classes de palavras, como: substantivos, adjetivos, advérbios, preposições, artigos, conjunções e inclusive ponto. No **Anexo D**, é possível visualizar essas informações em duas ferramentas de extração disponíveis nos sites do Projeto de Lingüística Computacional Hermes da Fundação Universidade Federal do Rio Grande (FURG/Brasil) e do Programa de LAEL da PUC-SP - Programa de Estudos Pós-Graduados em Lingüística Aplicada e Estudos da Linguagem da Pontifícia Universidade Católica de São Paulo[31].

A partir desta identificação, adjetivos, advérbios e verbos, são transformados em substantivos (concreto e/ou abstrato), quando for possível. Ou até mesmo o adjetivo seja o mesmo nome (grafia) para substantivos. Esse processo de nominalização, no trabalho de Gonzalez (2005), foi realizado através da ferramenta CHAMA, desenvolvida por ele mesmo.

Após o processo de nominalização são identificadas as RLBs (Relações Lexicais Binárias), conforme descrito nesta seção. Gonzalez (2005) desenvolveu também a ferramenta RELLEX para identificação das RLBs. Para o caso do teste, optou-se por fazer manualmente[32] **(ANEXO E)** devido à indisponibilidade destas duas ferramentas. Esta etapa tem uma importância muito grande, onde são reconhecidos os relacionamentos das palavras no texto através de identificadores. A **tabela 4** mostra as RLBs identificadas do parágrafo 6 do documento1 **(ANEXO A)**, de forma manual.

RLBs classificação	=(textil, industria)
RLBS restrições	de (industria, reciclagem) de (reciclagem, plastico) de (reaproveitamento, PET) de (segmento, monofilamento) de (mercado, potencialidade) de (plastico, totalidade) de (conjunto, medida)

Tabela 4: RLBs identificadas no parágrafo 6 do documento1

[31] As páginas disponíveis são: hermes.sourceforge.net/hermesweb.html e http://www2.lael.pucsp.br/corpora/etiquetagem/index.html

[32] Dicionários consultados: MICHAELIS. **Dicionário Eletrônico**. Acesso em: mar de 2006 e FERREIRA, Aurélio Buarque de Holanda. **Novo Aurélio Século XXI**: o dicionário da língua portuguesa. 1999.

4.2. Descrição Formal do Modelo Proposto: SINTR+

Os Sintagmas Nominais de Kuramoto em conjunto com as abordagens utilizadas no modelo TR+ de Gonzalez promovem a utilização de conceitos orientados a objetos (O.O.) porque é considerada a melhor metodologia para projeto de software, permite uma organização aprimorada do código, tem uma proximidade com a UML (Linguagem de Modelagem Unificada), proporciona uma facilidade de manutenção do código, apresenta menor grau de replicação do código e possibilita uma aplicação em camadas, o MVC[33], um padrão de projeto através da Linguagem Orientada a Objetos. Para compreender estes conceitos e o desenvolvimento da modelagem proposta ressaltaram-se alguns aspectos básicos de seus fundamentos.

A Linguagem UML, segundo Larman (2000), expressa a modelagem de sistemas e utiliza os conceitos orientados a objetos. Como na aplicação proposta trabalhar-se-á especificamente apenas nas etapas de análise e projeto, considera-se importante o uso da linguagem UML por ser esta uma linguagem poderosa para expressar de modo claro e preciso o processo de geração de projetos de software. Para Wazlawick (2004), esta linguagem dá suporte a que esse processo gere uma estrutura fácil de ser compreendida. Para o autor isto ocorre quando se utiliza um software autodocumentado e de fácil entendimento tanto em nível macro quanto em detalhes.

Este autor define que o Processo Unificado (UP) está associado à notação UML e indica que suas fases são: concepção, elaboração, construção e transição. Conforme Wazlawick (2004), é na primeira fase que se faz o levantamento dos principais requisitos e compreende-se o sistema de forma abrangente. A fase de elaboração é constituída de análise e projeto e a fase de construção corresponde à implementação e testes.

A análise de requisitos, ainda segundo este autor (2004, p. 24), "está associada ao processo de descobrir quais são as operações que o sistema deve realizar e quais são as restrições que existem sobre elas". Já a análise de domínio, "está relacionada à descoberta das informações gerenciadas pelo sistema, ou seja, à representação e transformação da informação" (2004, p. 26).

No caso de um sistema de informações sobre uma instituição de ensino (Módulo controle de alunos), por exemplo, possivelmente a análise de requisitos permitiria descobrir que o sistema deveria controlar a data, o curso e a turma em que o aluno foi matriculado, o início e término do curso, calcular automaticamente os pagamentos, gerar relatórios de

contrato especificando as cláusulas legais de direito e dever do aluno na Instituição etc. Essas operações são chamadas de "requisitos funcionais".

Há, também, relacionados a um sistema em construção os requisitos não funcionais, que dizem respeito à operação e à usabilidade do sistema. Um exemplo de requisito não-funcional seria a necessidade de fazer a matrícula via Internet. Essa é uma restrição de operação. Um outro exemplo seria uma central de acidentes de trânsito onde o registro de um dado acidente devesse ser feito em no máximo 10 segundos, o que demandaria um processamento e uma interface bastante eficiente, constituindo-se esse em um requisito de usabilidade.

Para as etapas de levantamento e análise de requisitos costuma ser utilizado, o diagrama de casos de uso. Segundo Guedes (2004), esse diagrama possibilita a compreensão do comportamento externo do sistema por qualquer pessoa. Entendem-se aqui, casos de uso, segundo Larman (2000), como um documento narrativo que descreve a seqüência de eventos (ações) de um ator (um agente externo) que usa um sistema para completar um processo e descreve também as respostas do sistema. Pode se dizer que caso de uso é um cenário com atores e ambientes. Criam-se as cenas e as narrativas das mesmas, ajudando a entender o que se quer do sistema. O interessante dos casos de uso é que os mesmos permitem que o projeto seja construído de forma participativa por um grupo de pessoas, uma vez que sua descrição se dá em uma linguagem textual e diagramática.

A partir dos casos de uso é possível construir o modelo conceitual. Conforme Larman (2000, p. 99) "o modelo conceitual ilustra os conceitos significativos em um domínio de problema". Para Wazlawick (2004, p. 102), "...o modelo conceitual deve descrever a informação que o sistema vai gerenciar... trata-se de um artefato do domínio do problema e não do domínio da solução".

É importante ressaltar que o modelo conceitual representa somente o aspecto estático da informação. Os elementos que representam informação são: conceitos (representados por classes), atributos (informações alfanuméricas ligadas diretamente aos conceitos) e associações (tipo de informação que liga diferentes conceitos entre si).

O diagrama de casos de uso do sistema proposto foi desenvolvido no software JUDE *Community*: Ferramenta de Modelagem UML. Um software *freeware*, muito utilizado para a criação deste tipo de diagramas. Neste software podem também ser desenvolvidos os outros tipos de diagramas do UML tais como: de classes, seqüência, colaboração, gráficos de estados.

---

[33] A sigla significa *Model*, *View* e *Controller*.

Os casos de uso identificados para esta aplicação foram descritos em duas situações. A primeira é referente à pesquisa do usuário e a segunda ao gerenciamento e operação do banco de dados (BD) no nível de administrador. Para descobrir estes casos de uso foi necessário, primeiramente, identificar os atores envolvidos com o sistema (usuário e administrador). E, na seqüência, a cada grande processo reconhecido correspondeu a um caso de uso do sistema.

As **Figuras 25** e **26** são diagramas na UML que representam casos de uso e seus atores. As elipses significam casos de uso e os bonecos representam atores. Para cada uma das situações (pesquisa e gerenciamento de operação do BD no nível de administrador) foram identificados os seguintes casos de uso:

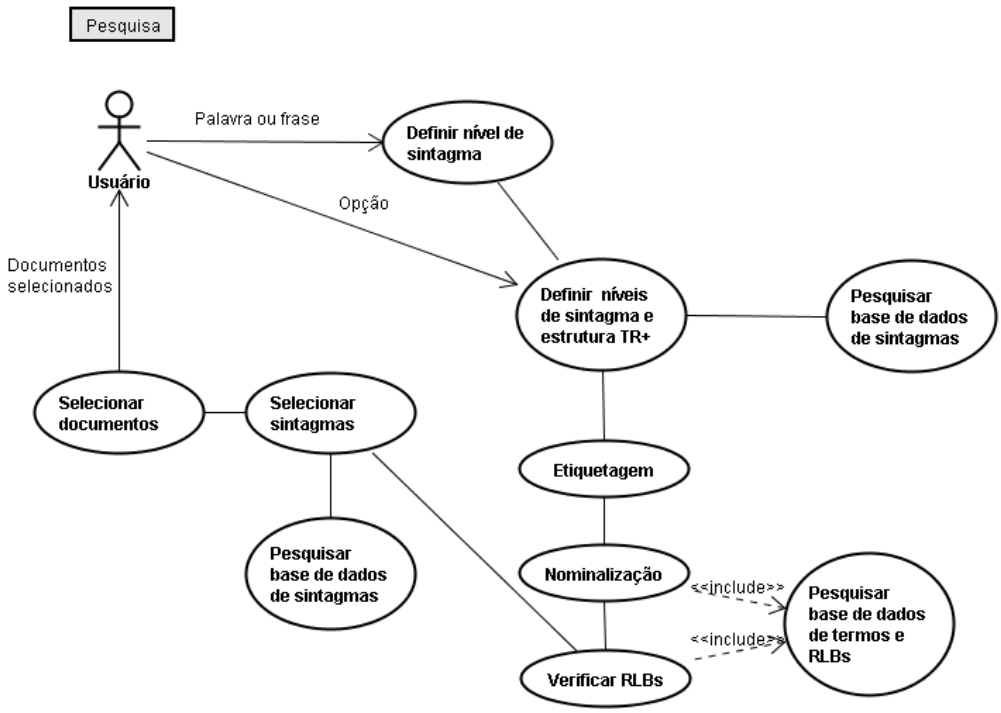

Figura 25: Diagrama de casos de uso da UML do sistema proposto – Pesquisa do Usuário

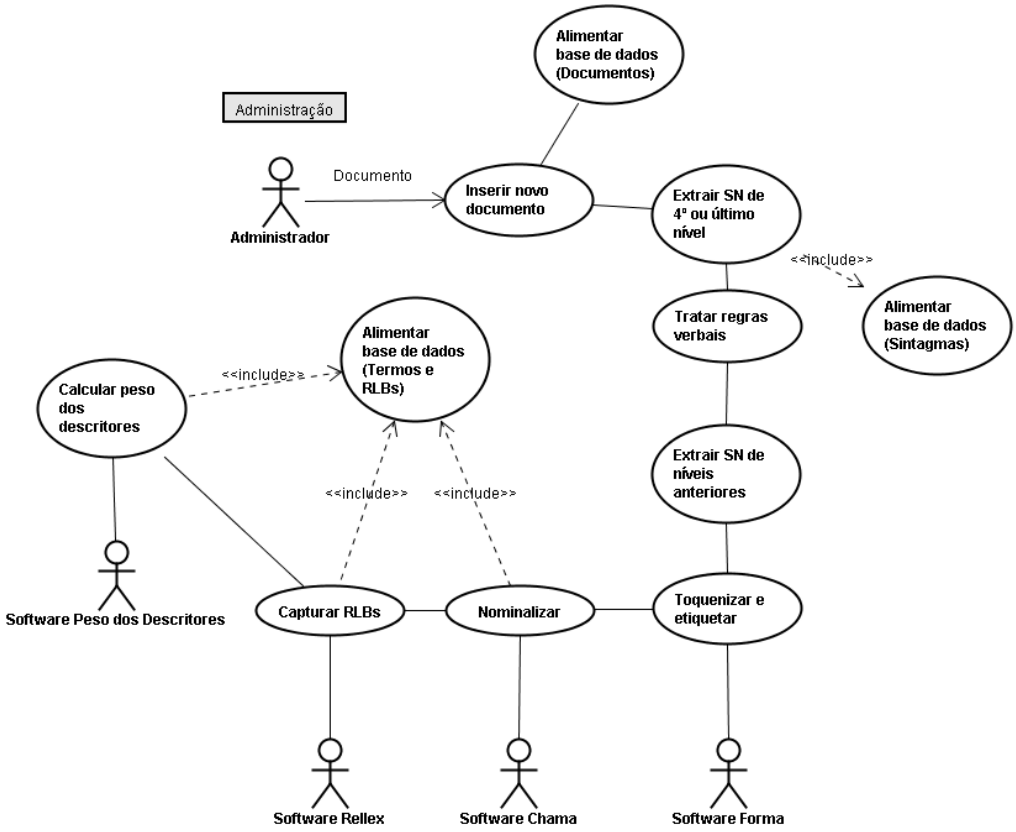

Figura 26: Diagrama de casos de uso da UML do sistema proposto – Gerenciamento e Operação do BD no nível de administrador

Deve-se lembrar que, na proposta deste trabalho, para economia de espaço de memória foram sistematizados dois momentos: o 1º em um armazenamento na base de dados do documento apenas para a lista final do usuário; e outro com os Sintagmas Nominais que serão armazenados na base de dados no 4º ou no último nível apresentado (Figura 19). Os níveis anteriores relativos ao SN serão procurados por uma programação desenvolvida relacionada diretamente com os Sintagmas. Com isto, não haverá necessidade de acesso à memória da base de documentos em todas as ações e esta servirá somente na última escolha do usuário tendo um ganho significativo quanto à rapidez de acesso aos dados da base e a não existência de duplicação de dados.

Os casos de uso costumam ser documentados, conforme Guedes (2004), por meio de uma linguagem bastante simples, fornecendo a função em linhas gerais dos casos de uso, quais atores interagem com os mesmos, quais etapas devem ser executadas pelo ator e pelo sistema, quais parâmetros devem ser fornecidos e quais restrições o caso de uso deve possuir. As **Tabelas** abaixo (5 a 15) apresentam as descrições dos casos de uso do sistema proposto referente ao gerenciamento e operação do BD no nível de administrador.

Nome do Caso de Uso: Inserir novo documento	
Caso de Uso Geral: não possui	
Ator Principal: Administrador	
Atores secundários: não possui	
Resumo: Permite ao administrador do sistema inserir arquivos na base de dados de documentos, iniciando o processo de alimentação de todas as demais bases de dados.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
1) Anexar um documento	
	2) Verificar se documento já não existe na base de dados
	3) Inserir o documento
Restrições/validações: Apenas documentos válidos[34] deverão ser aceitos	

Tabela 5: Descrição do caso de uso – Inserir novo documento

Nome do Caso de Uso: Alimentar base de dados (Documentos)	
Caso de Uso Geral: não possui	
Ator Principal: não possui	
Atores secundários: não possui	
Resumo: Armazenar em meio físico e com segurança os documentos inseridos pelo Administrador através do sistema.	
Pré-condições: Administrador anexa um documento válido	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Armazenar em base de dados os documentos anexados
Restrições/validações: não possui	

Tabela 6: Descrição do caso de uso – Alimentar base de dados (Documentos)

Nome do Caso de Uso: Extrair SN de 4º ou último nível	
Caso de Uso Geral: não possui	
Ator Principal: não possui	
Atores secundários: não possui	
Resumo: Extrair do documento inserido na base de dados todos os sintagmas nominais de 4º ou último nível.	
Pré-condições: o documento estar devidamente validado e inserido na base de dados	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) realizar a análise do documento inserido, extraindo todos os sintagmas nominais de 4º ou último nível, enviando informações para alimentação de base de dados de sintagmas
Restrições/validações: não possui	

Tabela 7: Descrição do caso de uso – Extrair SN de 4º ou último nível

[34] Documentos válidos são considerados aqui apenas os documentos em formato de texto (como doc, txt).

As ações do sistema da **tabela 7** seguem as regras estabelecidas na seção 4.1 da página 66.

Nome do Caso de Uso: Tratar regras verbais	
Caso de Uso Geral: não possui	
Ator Principal: não possui	
Atores secundários: não possui	
Resumo: Realizar o tratamento de regras verbais dos sintagmas nominais de 4º ou último nível extraídos do documento.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) aplicar rotinas de tratamento de regras verbais e palavras no infinitivo
Restrições/validações: não possui	

Tabela 8: Descrição do caso de uso – Tratar regras verbais

Nome do Caso de Uso: Extrair SN de níveis 3, 2 e 1 (níveis anteriores)	
Caso de Uso Geral: não possui	
Ator Principal: não possui	
Atores secundários: não possui	
Resumo: Aplicar regras de extração de sintagmas de níveis 3, 2 e 1 (níveis anteriores)	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Definir o nível apropriado de cada sintagma a partir do 4º ou último nível, enviando informação para o usuário
Restrições/validações: não possui	

Tabela 9: Descrição do caso de uso – Extrair SN de níveis 3, 2 e 1 (níveis anteriores)

A **tabela 9** segue a mesma regra da **tabela 7**.

Nome do Caso de Uso: Alimentar base de dados (Sintagmas)	
Caso de Uso Geral: não possui	
Ator Principal: não possui	
Atores secundários: não possui	
Resumo: Persistir as informações extraídas nos casos de uso “Extrair SN de 4º ou último nível”.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Armazenar na base de dados o 4º ou último nível de sintagma extraído do documento inserido
Restrições/validações: não possui	

Tabela 10: Descrição do caso de uso – Alimentar base de dados (Sintagmas)

Nome do Caso de Uso: *Toquenizar* e etiquetar	
Caso de Uso Geral: não possui	
Ator Principal: Software Forma	
Atores secundários: não possui	
Resumo: Submeter os sintagmas extraídos ao software Forma.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Aplicar o conceito de *Toquenização* e Etiquetagem dos sintagmas extraídos e armazenados em base de dados
Restrições/validações: não possui	

Tabela 11: Descrição do caso de uso – Toquenizar e etiquetar

Nome do Caso de Uso: Nominalizar	
Caso de Uso Geral: não possui	
Ator Principal: Software Chama	
Atores secundários: não possui	
Resumo: Submeter as informações resultantes do processo de *Toquenização* e Etiquetagem ao software Chama.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Aplicar o conceito de Nominalização das informações do documento
Restrições/validações: não possui	

Tabela 12: Descrição do caso de uso – Nominalizar

Nome do Caso de Uso: Capturar RLBs	
Caso de Uso Geral: não possui	
Ator Principal: Software Rellex	
Atores secundários: não possui	
Resumo: Submeter as informações resultantes do processo de Nominalização ao software Rellex.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Realizar o processo de captura de RLBs a partir das informações extraídas do documento
Restrições/validações: não possui	

Tabela 13: Descrição do caso de uso – Capturar RLBs

Nome do Caso de Uso: Calcular peso dos descritores	
Caso de Uso Geral: não possui	
Ator Principal: Software Peso dos Descritores	
Atores secundários: não possui	
Resumo: Submeter as informações resultantes do processo de Captura de RLBs ao software Peso de Descritores.	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Calcular o peso dos descritores ao resultado obtido através da captura de RLBs do documento
Restrições/validações: não possui	

Tabela 14: Descrição do caso de uso – Calcular peso dos descritores

Nome do Caso de Uso: Alimentar base de dados (Termos e RLBs)	
Caso de Uso Geral: não possui	
Ator Principal: não possui	
Atores secundários: não possui	
Resumo: Persistir as informações obtidas nos casos de uso “Nominalizar”, “Capturar RLBs” e “Calcular peso dos descritores” na base de dados de Termos e RLBs	
Pré-condições: não possui	
Pós-condições: não possui	
Ações do ator	Ações do sistema
	1) Armazenar as informações relativas aos Termos e RLBs extraídos do documento em base de dados
Restrições/validações: não possui	

Tabela 15: Descrição do caso de uso – Alimentar base de dados (Termos e RLBs)

Após a identificação dos casos de uso e suas descrições partiu-se para o modelo conceitual da aplicação proposta.

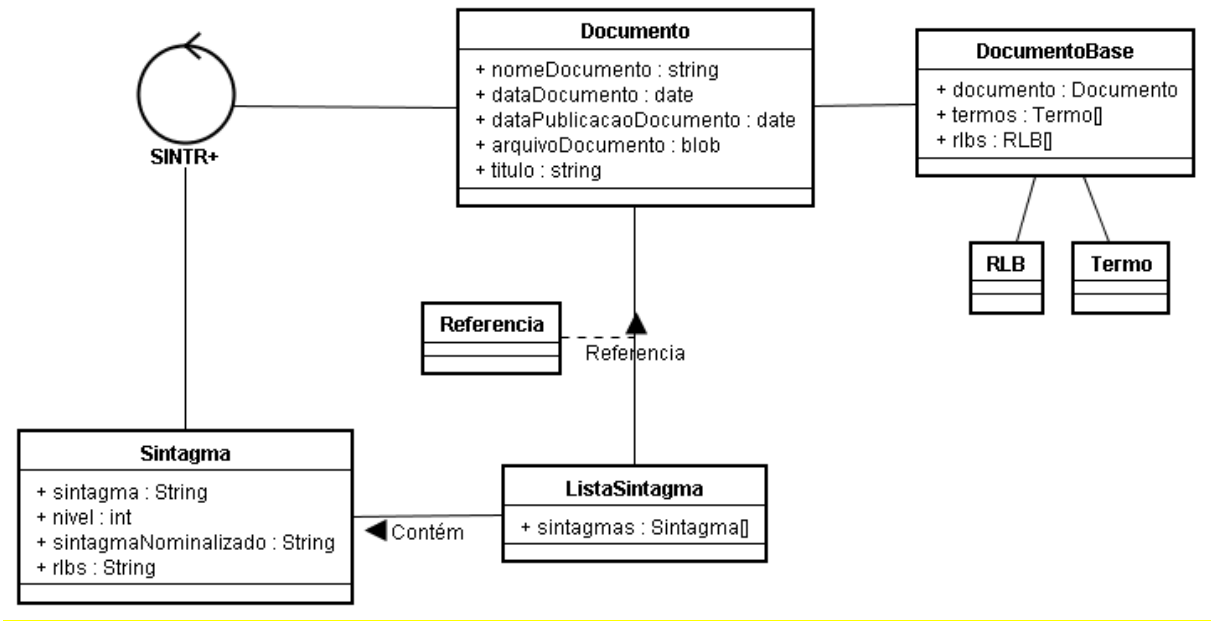

Figura 27: Modelo Conceitual do sistema proposto

O diagrama de classes, segundo Guedes (2004), é considerado o mais importante e o mais utilizado diagrama da UML. É o diagrama de classes que permite a visualização das classes que irão compor o sistema com os seus respectivos atributos e métodos. Demonstra como as classes se relacionam, complementam e transmitem informações entre si. Pode-se dizer que esse diagrama serve ainda como base para a construção de outros diagramas da linguagem UML.

A **Figura 28** apresenta o diagrama de classes do modelo proposto referente à Pesquisa do usuário.

Foi construído um diagrama de classes (Pesquisa de Usuário) seguindo estas definições/ações:

- Página de Consulta: refere-se a uma página HTML de pesquisa (ou seja, uma linguagem para *Web*) ou também a uma interface gráfica (GUI) para computador *desktop* (cliente).
- Controlador da Página: contém a lógica de negócio da aplicação.
- Classe Sintagma: *bean* responsável por instanciar e classificar sintagmas de diferentes níveis, usa o método setSintagma para receber informações vindas da página, passando pelo controlador.
- Classe ListaSintagma: cria instância de array de Sintagma, associando-os a instâncias de Documento. Realiza a busca e classificação destes, retornando ao controlador e posteriormente à página através do método getDocumentos.

- Classe Documento: instância de Documento armazenado em base de dados de documentos.

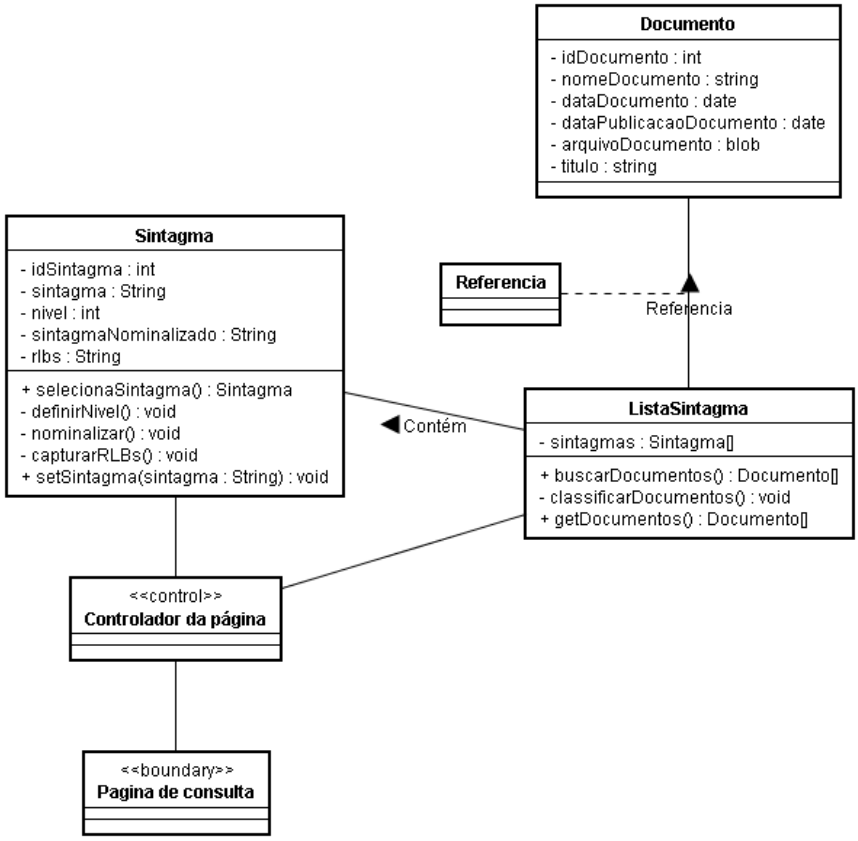

Figura 28: Diagrama de classes do sistema proposto – Pesquisa de Usuário

A **Figura 29** apresenta o diagrama de classes do modelo proposto referente ao Gerenciamento e Operação do BD no nível de administrador.

Foi construído um segundo diagrama de classes seguindo estas definições/ações:

- Página de Consulta refere-se a uma página HTML de inclusão de documentos
- Controlador da Página contém a lógica de negócio da aplicação
- Classe DocumentoBase: *bean* responsável por instanciar um objeto que irá conter o documento a inserir, bem como realizar os processos de *toquenização* e etiquetagem (trocando mensagens com o software FORMA), nominalização (trocando mensagens com o software CHAMA), gerando termos e RLBs (trocando mensagens com o software RELLEX), e por fim inserindo as informações nas bases de dados.
- Classes Termo e RLB: indicam as instâncias de objetos termos e RLBs e deverão ser modeladas conforme especificação do software RELLEX.

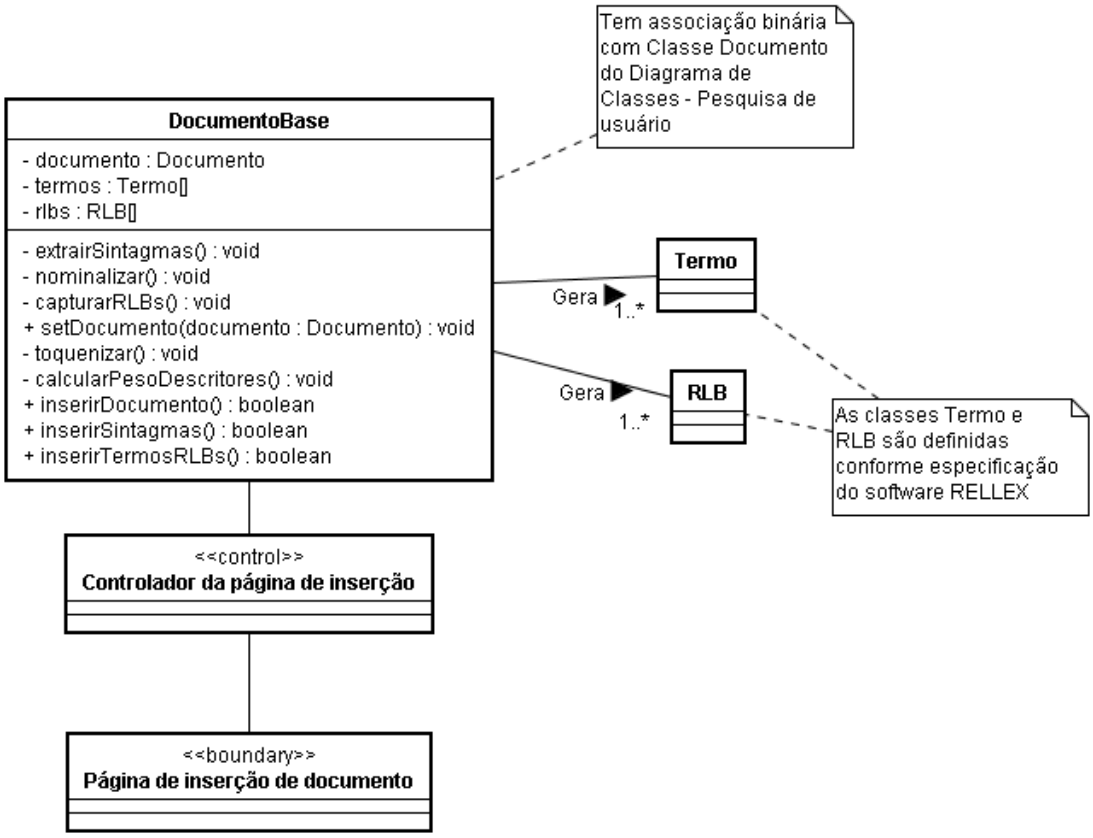

Figura 29: Diagrama de classes do sistema proposto – Gerenciamento e Operação do BD no nível de administrador

O diagrama de seqüência, segundo Guedes (2004), procura determinar a seqüência de eventos que ocorrem em um determinado processo, isto é, quais métodos devem ser disparados entre os objetos envolvidos, quais condições devem ser satisfeitas e em que ordem durante o processo específico. Foram construídos os diagramas de seqüência abaixo (**Figuras 30** e **31**) da aplicação proposta.

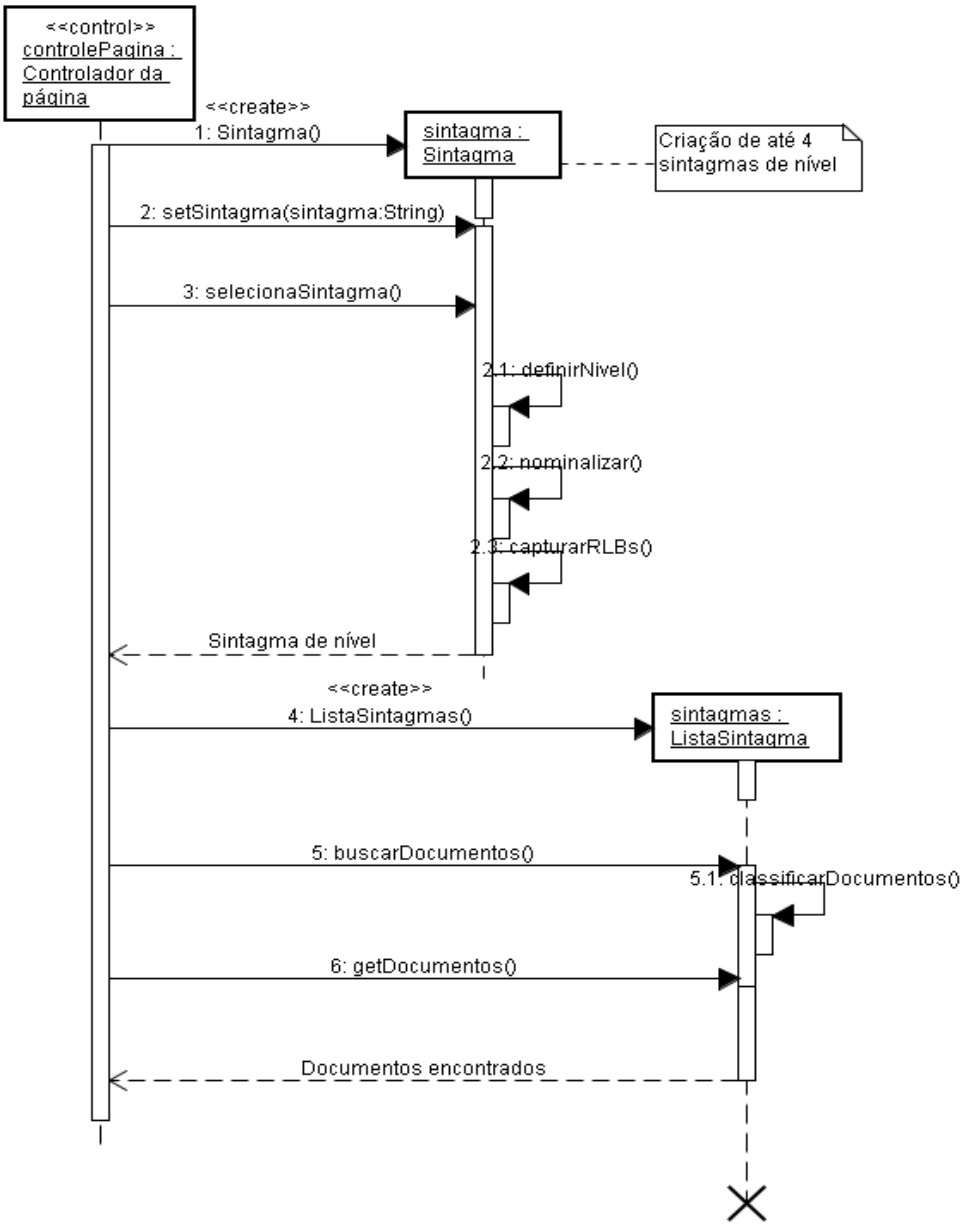

Figura 30: Diagrama de Seqüência do sistema proposto – Pesquisa de Usuário

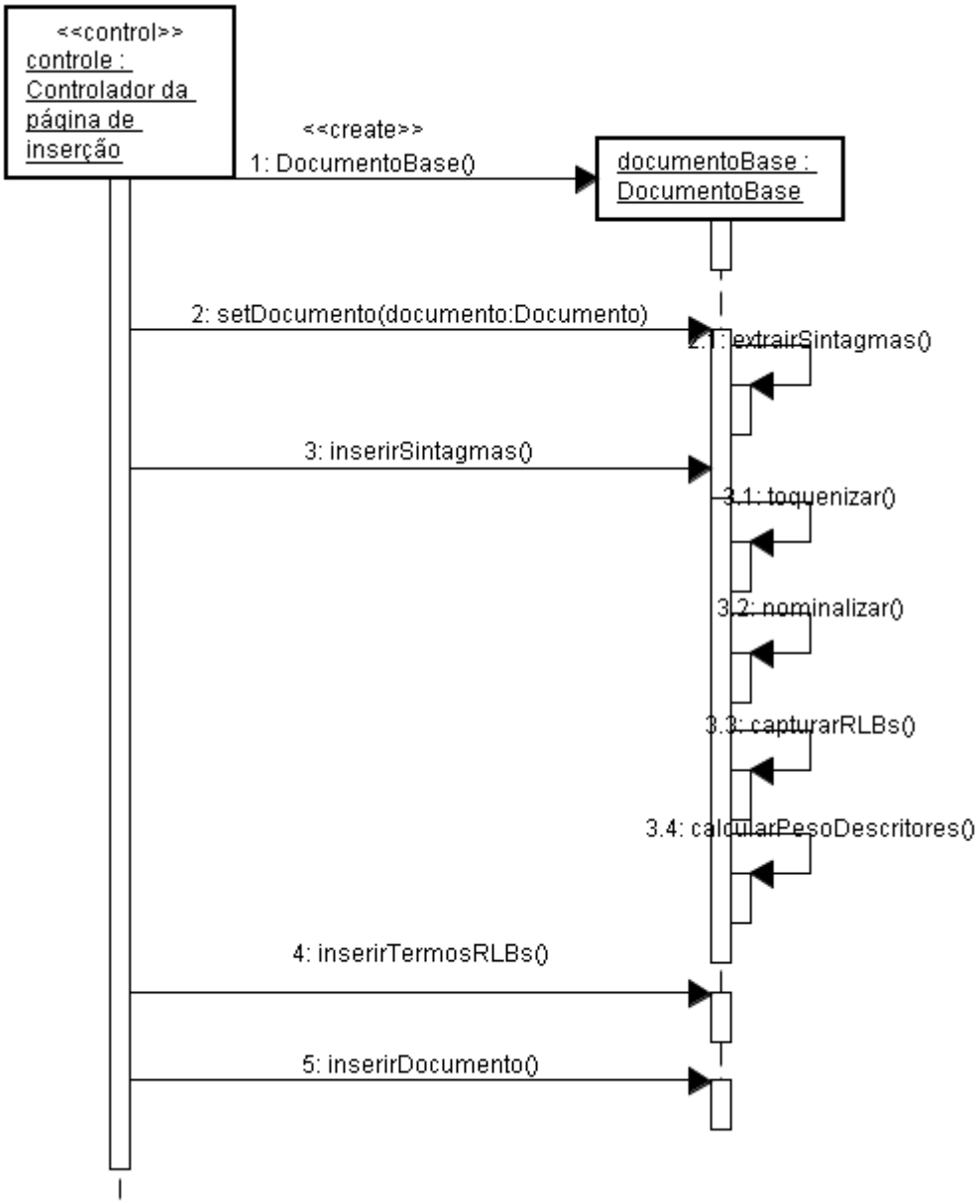

Figura 31: Diagrama de Seqüência do sistema proposto – Gerenciamento e Operação do BD no nível de administrador

Na elaboração dos diagramas e descrições dos casos de uso e dos diagramas de classes e de seqüência observou-se a importância do modelo conceitual porque permitiu orientar as etapas de desenvolvimento do modelo proposto. Visto que, no modelo conceitual foram criados conceitos, atributos e associações referentes à particularidade da pesquisa que puderam ser utilizados para a construção das etapas dos diagramas.

5. CONCLUSÃO

Neste capítulo apresentam-se as considerações finais, incluindo os aspectos relativos às dificuldades, aos progressos e limitações encontradas durante o desenvolvimento da pesquisa bem como as sugestões para a continuidade deste trabalho.

O objetivo geral que norteou este trabalho levou ao estudo dos modelos de busca e ao desenvolvimento de uma proposta para a melhoria dos processos de recuperação de informações.

Centrando-se no tema Recuperação de Informação, foram analisados os modelos de Kuramoto (1999), e, posteriormente, de Gonzalez (2005). O modelo de Kuramoto, baseado em uma estrutura hierárquica de sintagmas nominais, possibilita ao usuário definir melhor a sua *query* de busca. A Estrutura de *Qualia* do Léxico Gerativo de Pustejovsky contribuiu para o entendimento das relações e da estrutura de construção de significado entre as palavras, permitindo o tratamento de questões semânticas como a polissemia lógica. A proposta de Gonzalez, apropriando-se dos resultados de Pustejovsky, evidencia características morfológicas e relações de coesão importantes na descrição de conceitos presentes em um texto, propiciando que um texto possa computacionalmente significar mais do que uma seqüência de palavras.

Buscou-se uma síntese dessas propostas identificando as possibilidades de ampliação do modelo de Kuramoto pela junção da teoria do Léxico Gerativo de Pustejovsky utilizadas, nesta dissertação, a partir do modelo de Gonzalez que se manteve adequado devido ao fato de que o autor apresenta processos para as fases de indexação, busca e classificação de RI. Os termos e relacionamentos inseridos na base de dados, do modelo TR+ de Gonzalez, estão implicitamente relacionados com a Estrutura de *Qualia* do LG.

O novo modelo SINTR+, além do suporte ao usuário, envolve a análise, a sistematização e a ampliação do modelo de Kuramoto com a utilização da estrutura TR+ de Gonzalez (2005) para a melhoria e a otimização do processo de seleção dos documentos recuperados em uma busca.

O estudo e a descrição do modelo em UML permitiu, por ser uma linguagem poderosa, expressar de modo mais claro e preciso o modelo SINTR+. Foi construída a análise de domínio do sistema desejado, incluindo o desenvolvimento de diagramas de casos de uso bem como suas descrições, do modelo conceitual, de diagramas de classes e de seqüência. As

fases de análise e projeto, desenvolvidas para a aplicação proposta dão suporte à continuidade do seu desenvolvimento.

O novo modelo desenvolvido foi projetado como um sistema de recuperação de informação (SRI) aplicável a bases de dados não distribuídas abrangendo a um determinado domínio de aplicação, a sua adequação e expansão para uso na *Web,* constitui-se em uma importante linha de continuidade de pesquisa.

A principal contribuição deste trabalho está na sistematização e síntese das teorias de Kuramoto com Gonzalez, indicando o uso dessas teorias como uma nova alternativa para a melhoria da busca de recuperação de informações. Os modelos de recuperação simplesmente buscavam as informações solicitadas pelo usuário. O novo modelo proposto SINTR+ baseia-se na interação entre o usuário e a máquina através de Sintagmas Nominais por níveis e, também, nas relações das palavras conforme o modelo de Gonzalez.

Com este trabalho não se pretendeu desenvolver uma implementação completa do modelo construído. Mas o trabalho conseguiu mostrar a exeqüibilidade desta implementação computacional, descrevendo os diagramas e as descrições dos casos de uso e a sua modelagem conceitual culminando com a construção dos diagramas de classes e de seqüência. A próxima etapa, que permitiria detalhar as potencialidades e limitações do modelo de forma ampla, poderia se constituir em amplos estudos de casos onde se determinaria a complexidade computacional da implementação requerida.

Os dados apresentados no capítulo 4 já indicam aspectos positivos que consolidam a importância da utilização dos Sintagmas Nominais na diminuição de descritores para manipulação, com um ganho bastante significativo porque os índices possuem informações relevantes dos documentos (conceitos significativos de uma sentença) e com isto, agiliza-se a pesquisa na base de dados. Quer-se crer aqui, e um estudo mais amplo poderia determinar, que essa redução de descritores não deve ter nenhum impacto na qualidade da busca realizada.

Outro aspecto significativo é a redução do uso de memória tanto na fase de indexação como na de busca tornando mais rápido o processo interno.

Outro aspecto positivo se refere à melhoria de desempenho como um todo, pois quanto menor o tráfego em uma rede, menos informações o servidor vai processar e estará mais disponível. E quanto melhor for o processo de indexação, menos memória o servidor vai utilizar. E com isto, o tempo de resposta na fase de busca diminui e o resultado qualitativo da pesquisa se amplia.

Uma outra vantagem é que no modelo SINTR+ serão armazenados na base de dados os documentos e seus SN de últimos níveis e manipulados apenas os últimos níveis da estrutura de SN. Será só através de uma programação que serão classificados por níveis diminuindo assim o volume duplicado de dados na manipulação.

Os diagramas construídos referentes ao gerenciamento e operação do BD no nível do administrador são fundamentais para o entendimento do funcionamento e da manutenção do banco de dados facilitando processos como a inserção de novos documentos e outras ações, contribuindo também para o diferencial deste trabalho.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ABRAHÃO, Paulo Ricardo Carneiro. **Modelagem e Implementação de um Léxico Semântico para o Português**. Dissertação (Mestrado). Porto Alegre: PUCRS, 1997.

ABREU, Sandra C.; GOULART, Rodrigo; VIEIRA, Renata (2004). **Identificação de Expressões Anafóricas e Não Anafóricas com Base na Estrutura do Sintagma**. 2.º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2004) - Salvador/BA - 05 e 06 de agosto de 2004. Disponível em http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/tilsandra04.pdf. Acesso em: nov de 2004.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval**. New York: Addison-Wesley, 1999.

CARDOSO, Olinda N. P. Recuperação de Informações. In: **Infocomp-Journal of Computer Science**. vol. 2. n. 1. Lavras, MG: 2000. p.33-38. Disponível em: http://www.dcc.ufla.br/infocomp/artigos/v2.1/olinda.pdf. Acesso em: mar de 2004.

CHISHMAN, Rove. et al. Extração de Sintagmas Nominais para o Processamento de Co-Referência. In: **V Encontro para o processamento computacional do Português escrito e falado** (PROPOR 2000). Atibaia - São Paulo. Anais do V Encontro para o processamento computacional do Português escrito e falado. São Carlos: ICMC/USP, 2000. Disponível em: http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/propor00.pdf. Acesso em: jan de 2005.

FERNEDA, Edberto. **Recuperação de Informação: análise sobre a contribuição da ciência da computação para a ciência da informação**. Tese (Doutorado). São Paulo: USP, Escola de Comunicação e Artes, 2003. Disponível em http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/. Acesso em: set de 2004.

FERREIRA, Aurélio Buarque de Holanda. **Novo Aurélio Século XXI**: o dicionário da língua portuguesa. 3.ed. Rio de Janeiro: Nova Fronteira, 1999.

GASPERIN, C.; GOULART, R.; VIEIRA, R. **Uma ferramenta para Resolução Automática de Co-referência**. Anais do Encontro Nacional de Inteligência Artificial (ENIA), Campinas: SP, 2003. Disponível em http://www.exatec.unisinos.br/~renata/laboratorio/publicacoes/art1.pdf. Acesso em: set de 2004.

GONZALEZ, Marco Antônio Insaurriaga. **Representação Semântica de Sentenças em Linguagem Natural e sua aplicação na Recuperação de Informação**. Trabalho Individual 2. Doutorado. Porto Alegre: PPCC da PUCRS, 2000.

_____. **O Léxico Gerativo de Pustejovsky sob o enfoque da Recuperação de Informações**. Trabalho Individual 1. Doutorado. Porto Alegre: PPCC da PUCRS, 2000a.

_____. **Termos e Relacionamentos em Evidência na Recuperação de Informação**. Tese (Doutorado). Porto Alegre: PPGC da UFRGS, 2005.

GUEDES, Gilleanes T. A. **UML**: uma abordagem prática. São Paulo: Novatec, 2004.

HILL, Brad. **Pesquisa na Internet**. Rio de Janeiro: Campus, 1999.

KURAMOTO, Hélio. **Proposition d'um Système de Recherche d'Information Assistée par Ordinateur**. Tese (Doutorado). L'Université Lumière – Lyon - França, 1999.

_____. **Uma abordagem alternativa para o tratamento e a recuperação de informação textual**: os sintagmas nominais. Ciência da Informação (Brasília), v.25, n.2, 1995. Disponível em: http://dici.ibict.br/archive/00000169/01/Ci[1].Inf-2004-476.pdf. Acesso em: mar de 2004.

_____. **Sintagmas Nominais: uma nova proposta para a recuperação de informação**. DataGramaZero Revista de Ciência da Informação. v.3, n.1, fev. 2002. Disponível em: http://www.dgzero.org/fev02/Art_03.htm. Acesso em: mar de 2004.

LARMAN, Craig. **Utilizando UML e Padrões: uma introdução à análise e ao projeto orientado a objetos**. Porto Alegre: Bookman, 2000.

MARTINS, Dileta Silveira; ZILBERKNOP, Lúbia Scliar. **Português Instrumental**. 20ª ed. Porto Alegre: Sagra Luzzatto, 1999.

MOURA, Heronides M. de M. **A determinação de sentidos lexicais no contexto**. Cadernos de Estudos Lingüísticos. v. 41. Campinas, SP: 2001.

NETO, Magdiel Medeiros Aragão. **A polissemia em palavras designativas de objetos físicos e eventos**. 2003. Disponível em http://www.abralin.org.br/anais.htm. Acesso em: mai de 2004.

_____.**A Polissemia de acordo com a Teoria do Léxico Gerativo**. São Miguel do Oeste, SC: Revista do Centro de Ciências da Comunicação e Artes. n.6 mai/ago. 2003a.

PUSTEJOVSKY, James. **The Generative Lexicon**. Association for Computational Linguistics. Computer Science Department. Brandeis University. Cambridge, MA: The MIT Press, 1991. Disponível em http://portal.acm.org/citation.cfm?id=176324. Acesso em: set de 2004.

ROSSI, Albertina. **Palavras Polissêmicas entre evento e informação e seu tratamento nos dicionários Aurélio e Houaiss**. Tese (Doutorado) Florianópolis: USFC, Centro de Comunicação e Expressão - Programa de Pós-Graduação em Letras/Lingüística, 2003.

SILVA, Edna Lúcia da. **Metodologia da pesquisa e elaboração de dissertação**. Edna Lúcia da Silva, Estera Muszkat Menezes. – 2a ed. rev.– Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001. Disponível em: http://projetos.inf.ufsc.br/arquivos/Metodologia%20da%20Pesquisa%203a%20edicao.pdf. Acesso em: mai de 2005.

SILVA, Maria C. de S.; KOCH, Ingedore V. **Lingüística aplicada ao português: sintaxe**. 5.ed. São Paulo: Cortez, 1993.

WAZLAWICK, Raul Sidnei. **Análise e Projeto de Sistemas de Informação Orientados a Objetos**. Rio de Janeiro: Elsevier, 2004.

6.1 Bibliografia Consultada

BRÄSCHER, Marisa. **A Ambigüidade na Recuperação da Informação**. Revista Ciência da Informação (Brasília), v.3, n.1, 2002. Disponível em: http://www.dgz.org.br/fev02/Art_05.htm. Acesso em: abr de 2004.

CARVALHO, Nívea M. de Melo. **Recuperação da informação: implementação e avaliação de sistema de recuperação de informação utilizando o modelo vetorial**. Dissertação (Mestrado) Amazonas: Universidade Federal do Amazonas, Programa de Pós-Graduação em Informática, 2002. Disponível em: http://pos.facom.ufu.br/~rene/acervo/sri/RI-ModeloVetorial-NiveaCarvalho.pdf Acesso em: Ago de 2004.

FODOR, Jerry; LEPORE, Ernie. **The emptiness of the Lexicon:** Critical Reflections on J. Pustejovsky's The Generative Lexicon. Rutgers University Center for Cognitive Science.

GOMES, Andréia de Fátima R. **O singular nu e a sentença genérica no português brasileiro**. Dissertação (Mestrado) Florianópolis: UFSC, Programa de Pós-Graduação em Lingüística, 2001.

GONZALEZ, Marco; LIMA, Vera L. S. de. Sintagma Nominal em Estrutura Hierárquica Temática na Recuperação de Informação. **Anais**. ENIA 2001, Fortaleza: 2001. Disponível em: http://www.inf.pucrs.br/~gonzalez/docs/sneht.pdf Acesso em: dez 2005.

_____. T-Lex: Thesaurus com Estruturação Semântica e Operações Gerativas. XXVII Conferencia Latinoamericana de Informatica (CLEI'2001). Ciudad de Mérida, Venezuela, 2001. Disponível em: http://www.inf.pucrs.br/~gonzalez/docs/artigotlex.pdf. Acesso em: jan de 2006. (http://www.inf.pucrs.br/~gonzalez/pesqq.htm)

_____. Recuperação de Informação e Processamento da Linguagem Natural. XXIII Congresso da Sociedade Brasileira de Computação, Campinas, 2003. Anais do III Jornada de Mini-Cursos de Inteligência Artificial. Disponível em: http://www.inf.pucrs.br/~gonzalez/docs/minicurso-jaia2003.pdf. Acesso em: jan de 2006.

HEIDE, Ann. **Guia do Professor para a Internet: completo e fácil**. 2.ed. Porto Alegre: Artes Médicas Sul, 2000.

MOURA, Heronides M. de M. **Linguagem e cognição na interpretação de metáforas**. Universidade Federal de Juiz de Fora: Editora UFJF, 2003. Disponível em: http://www.revistaveredas.ufjf.br/volumes/v6n1/cap11.pdf Acesso em: jan de 2006.

PARREIRAS, Fernando. **O uso de sintagmas nominais como fonte de descritores para textos de periódicos científicos**. Escola de Ciência da Informação. Belo Horizonte, 2003. Disponível em: http://www.fernando.parreiras.nom.br/publicacoes/sn.pdf. Acesso em: set de 2004.

PÉREZ, Cláudia C. C.; GASPERIN, Caroline; VIEIRA, Renata. **Extração Semi-Automática de Conhecimento a partir de Textos**. 2003. Disponível em: http://www.inf.unisinos.br/~renata/laboratorio/ publicacoes/enia2003-submitted.pdf Acesso em: ago de 2005.

PIZZATO, Luiz A. Estrutura Multitesauro para Recuperação de Informações. Dissertação (Mestrado) Porto Alegre: PUCRS, Faculdade de Informática - Pós-Graduação em Ciência da Computação, 2003. Disponível em: http://www.pucrs.br/uni/poa/info/pos/dissertacoes/arquivos/pizzato.pdf Acesso em: ago de 2004.

PUSTEJOVSKY, James. **Type Construction and the logic of concepts**. Disponível em http://www.cs.brandeis.edu/~jamesp/articles/index.html Acesso em: set de 2004.

_____. **The Metaphysics of Words in Context**. (2000) Disponível em http://www.cs.brandeis.edu/~jamesp/articles/index.html Acesso em: set de 2004.

_____. **The Semantics of Agentive Nominals**. Disponível em http://www.cs.brandeis.edu/~jamesp/articles/index.html Acesso em: set de 2004.

WORDNET a lexical database for the English language. Cognitive Science Laboratory. Princeton University. Disponível em: http://wordnet.princeton.edu/ Acesso em: jan de 2006.

ANEXOS

ANEXO A - DOCUMENTO1

Endereço na *Web*: http:// www.reciclaveis.com.br/anamg.htm

Segunda-feira, 28 de agosto de 2000 - Número 599

Cresce a indústria de reciclagem de plásticos

Porém, potencial do lixo doméstico ainda é pouco aproveitado no estado

A indústria de reciclagem foi a que mais cresceu no setor plástico de Santa Catarina, nos últimos cinco anos. No período, o volume reprocessado no estado cresceu 166,4% ao ano, atingindo 16,9 mil toneladas em 1999. Isso equivale a 3,7% do total transformado pelo setor em Santa Catarina. Os dados fazem parte de estudo elaborado pela empresa de consultoria MaxiQuim, de Porto Alegre, para o Sindicato da Indústria de Material Plástico no Estado de Santa Catarina (Simpesc). Contudo, esse crescimento reflete mais o reaproveitamento de resíduos gerados em processos industriais do que a reciclagem de lixo doméstico, como
1 embalagens e garrafas, o chamado plástico "pós-consumo". Este segmento cresce de maneira menos acelerada, devido a problemas como a necessidade de escala de produção, falta de linhas de financiamento e ausência de legislação que estimule a atividade. "Embora a reciclagem do material pós-consumo, como sacos, embalagens e garrafas, esteja aumentando em Santa Catarina, a maior parte do crescimento verificado entre 1995 e 1999 refere-se a empresas que utilizam resíduos industriais como matéria-prima", explica o diretor da MaxiQuim, João Luiz Zuñeda. Normalmente chamadas de aparas, esses resíduos incluem também as peças que não atingiram a qualidade necessária para ir ao mercado.

As oito empresas catarinenses de reciclagem de plástico têm 383 empregados, sem considerar
2 o pessoal que trabalha na coleta de lixo, atividade que geralmente é informal. O valor da produção atingiu R$ 42,49 milhões em 1999, com crescimento médio de 152,6% ao ano nos últimos cinco anos, já descontando a inflação.

As empresas de transformação de plástico estão cada vez mais preocupadas em recuperar o material que antes era perdido, devido ao alto custo da resina virgem, diz Nelson Pradella,
3 proprietário da empresa Recicle-Ville. "Isso é fundamental para que elas sejam competitivas, pois vendendo os resíduos do processo industrial como sucata, as empresas obtêm menos de 20% do valor da resina virgem". Cobrando 30% do preço da resina virgem, a Recicle-Ville devolve para a indústria seus resíduos em condições de serem utilizados normalmente no processo produtivo", explica.

A empresa de Joinville foi uma das firmas que ajudou a elevar os índices desta indústria no
4 estado. Até agora ela estava trabalhando apenas com matéria-prima gerada nos processos industriais, mas isso deve mudar a partir desta semana. Criada há um ano, a empresa reprocessa cerca de 220 toneladas de plástico por mês e está aumentando a sua capacidade para 310 toneladas. Ela ainda opera, basicamente, como terceirizada de empresas de processamento de plásticos, reprocessando para elas os resíduos que geram e devolvendo essa matéria em forma granular, mesmo estado da resina virgem. Como a matéria prima reciclada será utilizada para fazer o mesmo produto que originou a

apara, a qualidade final não é afetada. Mas, a Recicle-Ville está ingressando também no segmento de reciclagem do plástico pós-consumo. A partir desta semana, a empresa coloca em funcionamento um sistema de coleta junto a escolas do município para recolher materiais plásticos como sacos, garrafas e tampinhas, apostando principalmente no PET. Com isso, ela tem a vantagem de receber material mais limpo.

A contaminação do plástico pelo lixo orgânico é justamente um dos principais problemas para o crescimento da indústria da reciclagem do lixo doméstico. A simples separação do lixo orgânico do seco já traria um impulso importante para o setor, diz Ana Flores, diretora do
5 departamento de meio ambiente e desenvolvimento sustentado da Federação das Indústrias do Estado de São Paulo (Fiesp), e autora do livro "O dinheiro está no lixo – recicle essa idéia". "Deveriam ser criados mecanismos de estímulo para a reciclagem. Na Holanda, por exemplo, uma Coca-Cola custa US$ 2,20. Devolvendo a garrafa acontece o reembolso de US$ 1. Você acha que alguém vai jogá-la no lixo?", diz.

A indústria da reciclagem do plástico no Brasil tem crescido bastante em função do reaproveitamento do PET, que é usado no segmento de monofilamentos, em artigos como vassouras e na indústria têxtil. Conforme Ana Flores, a reciclagem gera 250 mil empregos no
6 País, dos quais 70% são informais. Porém, a maior parte do potencial de mercado ainda está sendo desperdiçado, avalia. "Cerca de 15% do total de plástico que é industrializado no País é reciclado. Em dez anos poderíamos chegar a 60%, como nos Estados Unidos, desde que fosse implementado um conjunto de medidas incentivando essa prática", assegura.

Para a diretora da Fiesc, os principais entraves são o aspecto cultural, a tributação incidente na reciclagem do plástico, a falta de linhas de financiamento e a ausência de uma legislação
7 ambiental mais rigorosa. "Há um contra-senso ecológico que força a clandestinidade no Brasil, onde para fabricar garrafa PET virgem paga-se IPI de 10% e para a reciclagem 12%", critica. Ana afirma que essa tributação decorre do interesse governamental em incentivar a indústria química.

Outro problema apontado é que, ao contrário da indústria do alumínio, que é concentrada, o predomínio das pequenas empresas na transformação do plástico dificulta que sejam criadas
8 grandes empresas para reprocessar o lixo. Para Flores, o sucesso brasileiro na reciclagem do alumínio, (o índice é de 65%, um dos mais altos do mundo) decorre da existência de poucas grandes empresas capitalizadas. "As pequenas empresas não têm acesso às linhas de crédito, e isso dificulta a abertura de novas recicladoras", diz Flores.

Mas, há quem aponte outros desafios a superar. "Embora seja um mercado que deve crescer muito, a reciclagem de plástico não é tão simples como normalmente aparece na televisão. O volume mínimo para que a atividade seja economicamente viável, atendendo a todas as
9 exigências legais, é de 100 toneladas mês", diz Ronaldo Cerri, sócio da Moinhos Rone, de São Paulo, que fabrica equipamentos utilizados na moagem do plástico, uma das primeiras etapas da reciclagem. Além disso, explica, a coleta do plástico é mais complicada porque, ao contrário das latas de alumínio - que podem ser amassadas, o volume físico é maior. "Hoje entre 70% e 80% dos moinhos que vendemos são para reciclagem de resíduos industriais", informa. (Elmar Meurer, de Joinville)

ANEXO B - DOCUMENTO2

Cuidados com o Lixo

Endereço na *Web*: http://www.poupetempo.com.br/ambiente/lixo.htm

Todos os seres vivos quando morrem apodrecem: plantas e animais se decompõem e são destruídos por larvas, bactérias e fungos e reabsorvidos pela terra, pela água, pelo ar. É o ciclo da natureza: morte, decomposição, nova vida e crescimento.

Tudo o que é fabricado pelo homem acaba virando lixo. Muito desse lixo não se decompõe facilmente, como a matéria orgânica e passa a ser um problema. Plásticos, latas e vidros demoram muitos anos para se decompor e poluem o meio-ambiente. Por isso, a importância da reciclagem do lixo fabricado pelo ser humano.

O lixo é formado por resíduos sólidos não biodegradáveis e que demoram para se decompor. Restos de alimentos, folhas e frutas são chamados lixo orgânico.

Existem também, além do lixo domiciliar, o lixo industrial, o de vias públicas e o hospitalar, que necessitam de tratamentos especiais, pois oferece perigo à saúde das pessoas.

Devido ao aumento da população das grandes cidades e com o aumento do consumo de produtos, a quantidade de lixo também tem aumentado.

O acúmulo de lixo é um dos principais problemas nas grandes cidades. Muitos materiais que vão para o lixo não podem ser desperdiçados, podendo ser reaproveitados e reutilizados.

Material orgânico:

Tudo o que é resto de comida, de animais, de plantas e frutas é considerado lixo, propriamente dito. Ou seja, você deve acondicioná-los num único recipiente. Essa material é recolhido pela prefeitura e levado para aterros sanitários onde vão sofrer a decomposição natural.

Material reciclável:

É praticamente tudo o que é fabricado pelo homem: material plástico, latas de alumínio e ferro, garrafas de refrigerante de vidro e PET, caixas de papel e papelão, jornais, revistas, livros, aparas de papel, etc.

Se você mora em casa, reúna-se com sua família e com seus funcionários para estabelecer um método de separação desse material.

Dependendo do seu volume diário de lixo, escolha 4 recipientes coloridos para acondicioná-los, azul para papel, vermelho para plástico, verde para vidro e amarelo para metal, ou nomeie cada um deles, conforme sua classificação.

Se você mora em condomínio, faça esse mesmo trabalho reunindo os moradores, estabelecendo regras e instruindo os empregados.

Observação: o lixo orgânico deve estar separado daquilo que é reciclável.

Exemplos:

Providencie uma caixa resistente ou sacolas e fixe nelas um papel com a identificação do tipo de lixo: vidro e nela vá acumulando as garrafas.

Retire anéis e rótulos, e lave as garrafas para não acumular insetos.

Na outra caixa vá juntando o lixo papel, aparas, embalagens de papelão, as perdas da impressora, jornais e revistas velhas, etc.

Latas de conserva são de ferro, e as de refrigerante são de alumínio. Elas devem ser acumuladas limpas, sem rótulo e em caixas separadas. As de alumínio podem ser amassadas como uma sanfoninha o que economizará espaço.

Quando as caixas estiverem cheias, elas devem ser encaminhadas para entidades que trabalham com material reciclável ou simplesmente recolhida pela empresa de sua cidade, responsável pela coleta seletiva. Consulte a prefeitura local.

A destinação do material para reciclagem pode ser feita de várias formas. Uma família mais pobre pode utilizar esse material vendendo para cooperativas e empresas especializadas e conseguir um dinheiro extra.
Os condomínios de melhor padrão econômico podem utilizar o resultado da separação do lixo para reciclagem em benefício de seus funcionários, propiciando a eles um ganho extra na ajuda da triagem desse material.
Uma outra forma é, simplesmente, entregar todo o material para as prefeituras que já possuem o método de coleta seletiva.
Ajude a melhorar o meio-ambiente. É simples: pense antes de comprar. Metade do que nós compramos é lixo. São embalagens que quase sempre não servem para nada e vão direto para o lixo.
Evite embalagens plásticas: elas são pouco recicláveis, enquanto o vidro é totalmente reciclável e muito mais útil no seu reaproveitamento.
Algumas informações sobre materiais produzidos pelo homem:

TEMPO DE DECOMPOSIÇÃO DE ALGUNS MATERIAIS

Lenço de papel	3 meses
Palito de fósforo	6 meses
Caroço de maçã	6 a 12 meses
Ponta de cigarro	1 a 2 anos
Chiclete	5 anos
Lata de aço	10 anos
Garrafa de plástico	100 anos
Garrafa de vidro	Mais de 1.000 anos
Lata de alumínio	Não se corrói nunca

Plástico rígido: Leve, resistente e prático é o material que compõe cerca de 60% das embalagens plásticas, como garrafas de refrigerantes, recipientes para produtos de limpeza e higiene e potes de alimentos, é também matéria-prima básica de bombonas, fibras têxteis, tubos e conexões calçados, eletrodomésticos, além de baldes utensílios domésticos e outros produtos. Ele pode ser reprocessado, gerando novos artefatos plásticos e energia.
Papel ondulado: é usado em caixas para transporte de produtos para fábricas, depósitos, escritórios e residências. Normalmente chamado de papelão, este material tem uma camada intermediária de papel entre suas partes exteriores, disposta em ondulações, na forma de uma sanfona. O material é de fácil coleta em grandes volumes comerciais, sendo facilmente identificadas quando misturadas com outros tipos de papel, por isso seu susto de processamento é relativamente baixo.
Embalagens longa vida: são compostas de várias camadas de material: dupléx, polietileno e alumínio. As embalagens cartonadas precisam ser lavadas após o consumo porque os restos de alimentos contidos nelas dificultam o reprocessamento do material. Para aproveitar melhor o espaço, as embalagens podem ser amassadas.
O papel existente nas embalagens cartonadas pode ser compostado para a produção de húmus utilizado em hortas e jardins.
Pneus: a borracha e sua reciclagem é capaz de devolver ao processo de produção insumo regenerado por menos da metade do custo da borracha natural ou sintética, além disso, economiza energia e poupa petróleo usado como matéria-prima virgem e até melhora as propriedades de materiais feitos com borracha.
Latas de alumínio: além de reduzir o lixo que vai para os aterros a reciclagem desse material proporciona significativo ganho energético. Para reciclar uma tonelada de latas gasta-se 5% da

energia necessária para produzir a mesma quantidade de alumínio pelo processo primário. Isto significa que cada latinha reciclada equivale ao consumo de um aparelho de TV durante 3 horas. A reciclagem evita a extração da bauxita, o mineral beneficiado para a fabricação da alumina, que é transformada em liga de alumínio.

Vidro: a metade dos recipientes de vidro é fabricados no País é retornável. Além disso, o material é de fácil reciclagem: pode voltar a produção de novas embalagens substituindo o produto virgem sem perda da qualidade.

Pet (polietileno tereftalato): as garrafas recicladas são transformadas em cordas e fios de costura, carpetes, bandejas de frutas e até mesmo novas garrafas. Sua reciclagem, além de desviar lixo plástico dos aterros utiliza apenas 30% da energia necessária para a produção da resina virgem, e tem a vantagem de poder ser reciclado várias vezes sem prejudicar a qualidade do produto final.

Latas de aço: Quando reciclado, o aço volta ao mercado em forma de automóveis, ferramentas, vigas para construção civil, arames, vergalhões, utensílios domésticos e inclusive novas latas.

Plástico filme: é uma película plástica normalmente usada como sacolas de supermercados, sacos de lixo, embalagens de leite, lonas agrícolas e proteção de alimentos na geladeira ou microondas. Cerca de 44% é papel e 4% é folha de alumínio.

Ajude a melhorar o meio-ambiente

- Reaproveite sobras e não jogue fora o que puder aproveitar.
- Doe roupas que possam ser reformadas ou consertadas.
- Doe livros para bibliotecas ou instituições beneficentes.
- Use produtos biodegradáveis ou recicláveis.
- Deixe o óleo usado do motor no posto para ser reciclado.
- Leve pneus sem uso para os borracheiros.
- Evite jogar lixo na rua. Jogue o lixo na lixeira.
- Embale o lixo corretamente, sempre que possível encaminhe plásticos, vidros e papel para a reciclagem.

ANEXO C - EXTRAÇÃO MANUAL DE SN DOS DOCUMENTOS

DOCUMENTO1

Linha	Sintagma Nominal	Nível
1	Plásticos	1
1	Reciclagem de plásticos	2
1	Indústria de reciclagem de plásticos	3
2	Lixo	1
2	Lixo doméstico	1
2	Potencial do lixo doméstico	2
3	Reciclagem	1
3	Indústria de reciclagem	2
3	Plástico	1
3	Setor Plástico	1
3	Setor Plástico de Santa Catarina	2
7	Plástico	1
7	Material Plástico	1
7	Indústria de Material Plástico	2
7	Sindicato da Indústria de Material Plástico	3
7	Sindicato da Indústria de Material Plástico no Estado de Santa Catarina	4
8	Resíduos	1
8	Reaproveitamento de resíduos	2
9	Lixo	1
9	Lixo doméstico	1
9	Reciclagem do lixo doméstico	2
10	Embalagens	1
10	Garrafas	1
10	Embalagens e garrafas	2
10	Plástico	1
10	Plástico pós-consumo	1
13	Reciclagem	1
13	Reciclagem de material	2
13	Reciclagem de material pós-consumo	2
13	Sacos	1
13	Embalagens	1
13	Garrafas	1
13	Sacos, embalagens e garrafas	2
15	Resíduos	1

15	Resíduos industriais	1
15	Resíduos industriais como matéria-prima	2
18	Reciclagem	1
18	Reciclagem de plásticos	2
18	Empresas catarinenses de reciclagem de plásticos	3
19	Lixo	1
19	Coleta de lixo	2
22	Plástico	1
22	Transformação de plástico	2
22	As empresas de transformação de plástico	3
27	Resíduos	1
32	Plástico	1
34	Plásticos	1
34	Processamento de plásticos	2
34	Empresas de processamento de plásticos	3
34	Terceirizada de empresas de processamento de plásticos	4
34	Os resíduos	1
36	Reciclada	1
36	Matéria-prima reciclada	1
38	Reciclagem	1
38	Reciclagem de plástico	2
38	Reciclagem de plástico pós-consumo	2
38	Segmento de reciclagem de plástico pós-consumo	3
39	Coleta	1
39	Coleta junto a escolas do município	2
39	Um sistema de coleta junto a escolas do município	3
40	Plásticos	1
40	Materiais plásticos	1
40	Materiais plásticos como sacos, garrafas e tampinhas	2
41	PET	1
43	Lixo	1
43	Lixo orgânico	1
43	A contaminação do plástico	2
43	A contaminação do plástico pelo lixo orgânico	3
44	Lixo	1

44	Lixo doméstico	1
44	Reciclagem de lixo doméstico	2
44	Indústria da reciclagem do lixo doméstico	3
44	O crescimento da indústria da reciclagem do lixo doméstico	4
44	Lixo	1
44	Lixo orgânico	1
44	A simples separação do lixo orgânico	2
44	A simples separação do lixo orgânico do seco	3
49	A garrafa	1
50	Lixo	1
51	Reciclagem	1
51	Reciclagem do plástico	2
51	A indústria da reciclagem do plástico	3
51	A indústria da reciclagem do plástico no Brasil	4
52	Reaproveitamento	1
52	Reaproveitamento do PET	2
53	A reciclagem	1
55	Plástico	1
56	Reciclado	1
58	Reciclagem	1
58	Reciclagem do plástico	2
58	Tributação incidente na reciclagem do plástico	3
61	Garrafa	1
61	Garrafa PET	1
61	Reciclagem	1
63	Indústria Química	1
64	Alumínio	1
64	Indústria do alumínio	2
65	Plástico	1
65	Transformação do plástico	2
65	Empresas na transformação do plástico	3
65	Predomínio das pequenas empresas na transformação do plástico	4
66	Lixo	1
69	Recicladoras	1
69	A abertura de novas recicladoras	2

71	Reciclagem	1
71	Reciclagem de Plástico	2
74	Plástico	1
74	Moagem do plástico	2
74	Reciclagem	1
74	Primeiras etapas da reciclagem	2
75	Coleta	1
75	A coleta do plástico	2
77	Reciclagem	1
77	Reciclagem de resíduos	2
77	Reciclagem de resíduos industriais	3

DOCUMENTO2

Linha	Sintagma Nominal	Nível
1	Lixo	1
1	Cuidados com o lixo	2
5	Lixo	1
5	Lixo	1
6	Matéria Orgânica	1
6	Plásticos, latas e vidros	2
7	Lixo	1
7	Reciclagem do lixo	2
7	A importância da reciclagem do lixo	3
9	O lixo	1
9	Resíduos	1
9	Resíduos sólidos	1
9	Resíduos sólidos não-biodegradáveis	1
9	Restos de alimentos, folhas e frutas	2
10	Lixo	1
10	Lixo orgânico	1
11	Lixo	1
11	Lixo domiciliar, lixo industrial, o de vias públicas e o hospitalar	3
13	Lixo	1
13	A quantidade de lixo	2

15	Lixo	1
15	O acúmulo de lixo	2
15	O lixo	1
17	Material orgânico	1
18	Restos de comida, de animais, de plantas e frutas	4
18	Lixo	1
20	Aterro sanitário	1
20	A decomposição	1
20	A decomposição natural	1
21	Reciclável	1
21	Material reciclável	1
22	Material plástico, latas de alumínio e ferro, garrafas de refrigerante de vidro e PET, caixas de papel e papelão, jornais, revistas, livros, aparas de papel	4
24	Lixo	1
24	Volume diário de lixo	2
31	O lixo	1
31	O lixo orgânico	1
31	Reciclável	1
33	Lixo	1
33	Tipo de lixo	2
33	vidro	1
34	As garrafas	1
35	As garrafas	1
36	O lixo	1
36	O lixo papel, aparas, embalagens de papelão, as perdas da impressora, jornais e revistas velhas	3
38	Ferro	1
38	Alumínio	1
42	Reciclável	1
42	Material reciclável	1
42	Coleta	1
42	Coleta seletiva	1
44	Reciclagem	1
44	Material para reciclagem	2
44	A destinação do material para reciclagem	3

47	Lixo	1
47	Separação do lixo	2
47	Separação do lixo para reciclagem	3
47	O resultado da separação do lixo para reciclagem	4
50	Coleta	1
50	Coleta seletiva	1
50	O método de coleta seletiva	2
53	Lixo	1
53	Embalagens	1
53	O lixo	1
54	Embalagens plásticas	1
54	Pouco recicláveis	1
54	O vidro	1
57	Decomposição	1
57	Decomposição de alguns materiais	2
57	Tempo de decomposição de alguns materiais	3
67	Plástico	1
67	Plástico rígido	1
67	Embalagens plásticas	1
67	Embalagens plásticas, como garrafas de refrigerantes, recipientes para produtos de limpeza e higiene e potes de alimentos.	4
72	Papel ondulado	1
74	Coleta	1
74	Coleta em grandes volumes comerciais	2
74	Fácil coleta em grandes volumes comerciais	2
76	Processamento	1
76	Custo de processamento	2
77	Embalagens	1
77	Embalagens longa vida	1
79	reprocessamento	1
79	Reprocessamento do material	2
82	Hortas e jardins	2
83	Pneus	1
83	Reciclagem	1
83	A borracha e sua reciclagem	2

84	Borracha	1
84	Borracha natural ou sintética	1
84	O custo da borracha natural ou sintética	2
86	Latas de alumínio	1
86	O lixo	1
86	A reciclagem	1
88	Reciclada	1
88	Latinha reciclada	1
89	A reciclagem	1
91	vidro	1
91	Vidro	1
91	Recipiente de vidro	2
91	Fácil reciclagem	1
94	PET	1
94	Recicladas	1
94	Garrafas recicladas	1
95	Reciclagem	1
95	Lixo	1
95	Lixo plástico	1
95	Lixo plástico dos aterros	2
97	Reciclado	1
97	Reciclado várias vezes	1
98	Latas de aço	2
98	Automóveis, ferramentas, vigas para construção civil, arames, vergalhões, utensílios domésticos e inclusive novas latas	3
100	Plástico filme	1
100	Sacolas de supermercado, sacos de lixo, embalagens de leite, lonas agrícolas e proteção de alimentos na geladeira ou microondas	3
102	Folha de alumínio	2
107	Recicláveis	1
107	Produtos biodegradáveis ou recicláveis	1
110	Lixo	1
110	Lixo na rua	2
110	O lixo	1
110	O lixo na lixeira	2

111	O lixo	
111	Reciclagem	1
111	Plásticos, vidros e papel para reciclagem	3

ANEXO D - FERRAMENTA1 DE *TOQUENIZAÇÃO* E ETIQUETAGEM

Endereço na *Web*: hermes.sourceforge.net/hermesweb.html

Texto processado: (Parágrafo 6 do Documento1)

A indústria da reciclagem do plástico no Brasil tem crescido bastante em função do reaproveitamento do PET, que é usado no segmento de monofilamentos, em artigos como vassouras e na indústria têxtil. Conforme Ana Flores, a reciclagem gera 250 mil empregos no País, dos quais 70% são informais. Porém, a maior parte do potencial de mercado ainda está sendo desperdiçado, avalia. "Cerca de 15% do total de plástico que é industrializado no País é reciclado. Em dez anos poderíamos chegar a 60%, como nos Estados Unidos, desde que fosse implementado um conjunto de medidas incentivando essa prática", assegura.

Resultado:

A_ART
indústria_N
da_PREP+ART
reciclagem_N
do_PREP+ART
plástico_N
no_PREP+ART
Brasil_NP
tem_VTD
crescido_ADJ
bastante_ADV
em_PREP
função_N
do_PREP+ART
reaproveitamento_N
do_PREP+ART
PET_N
',_'
que_PR
é_VLIG
usado_ADJ
no_PREP+ART
segmento_N
de_PREP
monofilamentos_N
',_'
em_PREP
artigos_N

como_CONJSUB
vassouras_ADJ
e_CONJCOORD
na_PREP+ART
indústria_N
têxtil_NP
._.
Conforme_CONJSUB
Ana_NP
Flores_N
,_,
a_ART
reciclagem_N
gera_N
250_NC
mil_NC
empregos_N
no_PREP+ART
País_N
,_,
dos_PREP+ART
quais_PR
70_NC
são_VLIG
informais_ADJ
._.
Porém_VTD
,_,
a_ART
maior_ADJ
parte_N
do_PREP+ART
potencial_N
de_PREP
mercado_N
ainda_ADV
está_VLIG
sendo_VLIG
desperdiçado_VTD
,_,
avalia_N
._.
Cerca_N
de_PREP
15_NC
do_PREP+ART
total_ADJ
de_PREP
plástico_N
que_PR
é_VLIG

industrializado_VTD
no_PREP+ART
País_N
é_VLIG
reciclado_VTD
._.
Em_PREP
dez_NC
anos_N
poderíamos_VTD
chegar_VTI
a_ART
60_NC
,_,
como_CONJSUB
nos_PREP+ART
Estados_NP
Unidos_NP
,_,
desde_PREP
que_PR
fosse_VLIG
implementado_ADJ
um_ART
conjunto_N
de_PREP
medidas_N
incentivando_VTD
essa_PD
prática_N
,_,
assegura_VTD
._.

ANEXO D - FERRAMENTA2 DE *TOQUENIZAÇÃO* E ETIQUETAGEM

Endereço na *Web*: http://lael.pucsp.br/corpora/etiquetagem/

A	ARTD	
indústria	N	
da	CPR	
reciclagem		N
do	CPR	
plástico	N	
no	CPR	
Brasil	N	
tem	V	
crescido	PART	
bastante	ADV	
em	PRP	
função	N	
do	CPR	
reaproveitamento		N
do	CPR	
PET	N	
,	PT	
que	PRN	
é	V	
usado	PART	
no	CPR	
segmento		N
de	PRP	
monofilamentos		N
,	PT	
em	PRP	
artigos	N	
como	ADV	
vassouras		N
e	CJ	
na	CPR	
indústria	N	
têxtil	ADJ	
.	PT	
Conforme		ADJ
Ana	N	
Flores	N	
,	PT	
a	ARTD	
reciclagem		N
gera	V	
250	NUM	
mil	N	

empregos N
no CPR
País N
, PT
dos CPR
quais PRN
70 NUM
% PT
são V
informais ADJ
. PT
Porém CJ
, PT
a ARTD
maior ADJ
parte N
do CPR
potencial N
de PRP
mercado N
ainda ADV
está V
sendo V
desperdiçado PART
, PT
avalia V
. PT
“Cerca PRP
de PRP
15 NUM
% PT
do CPR
total N
de PRP
plástico N
que PRN
é V
industrializado PART
no CPR
País N
é V
reciclado PART
. PT
Em PRP
dez NUM
anos N
poderíamos V
chegar V
a ARTD
60 NUM
% PT

,	PT	
como	ADV	
nos	CPR	
Estados	N	
Unidos	N	
,	PT	
desde	PRP	
que	PRN	
fosse	V	
implementado		PART
um	ARTI	
conjunto	N	
de	PRP	
medidas	N	
incentivando		V
essa	PRN	
prática”	N	
,	PT	
assegura	V	
.	PT	

ANEXO E - PROCESSO DE NOMINALIZAÇÃO

Palavra Original	Classe	Substantivo Abstrato	Substantivo Concreto
tem	Verbo	Ɛ	Ɛ
crescido	Verbo no particípio	crescimento	Ɛ
bastante	Advérbio	Ɛ	Ɛ
é	Verbo	Ɛ	Ɛ
usado	Verbo no particípio	uso	usador
têxtil	Adjetivo	Ɛ	tecido
informal	Adjetivo	informalidade	Ɛ
maior	Adjetivo	maioridade	Ɛ
potencial	Adjetivo	potencialidade	Ɛ
ainda	Advérbio	Ɛ	Ɛ
está	Verbo	Ɛ	Ɛ
sendo	Verbo	Ɛ	Ɛ
desperdiçado	Verbo no particípio	desperdício	desperdiçador
total	Adjetivo	totalidade	totalizador
industrializado	Verbo no particípio	industrial	indústria
reciclado	Verbo no particípio	Ɛ	reciclagem
poderíamos	Verbo	Ɛ	Ɛ
chegar	Verbo	Ɛ	chegada
fosse	Verbo	Ɛ	Ɛ
implementado	Verbo no particípio	implemento	implementador
incentivando	Verbo	Incentivo	incentivador

Ɛ = ausência de nominalização