Cláudio Gottschalg-Duque

SiRILiCO

Uma Proposta para um *Si*stema de *R*ecuperação de *I*nformação baseado em Teorias da *Li*ngüística *C*omputacional e *O*ntologia.

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção de título de Doutor em Ciência da Informação.

Linha de Pesquisa: Informação Gerencial e Tecnológica

Orientador: Prof. Dr. Marlene de Oliveira

Belo Horizonte
Escola de Ciência da Informação - UFMG
2005

Ao meu pai (*meu passado*);
minha mãe, meu irmão
e minha esposa (*meu presente*);
e à minha filha (*meu futuro*)!

“Língua Portuguesa

Última flor do Lácio,
inculta e bela,
És, a um tempo,
esplendor e sepultura:
Ouro nativo, que na
ganga impura
A bruta mina entre os
cascalhos vela...”
Olavo Bilac

“Há homens que lutam um dia, e são bons:
Há outros que lutam um ano, e são melhores:
Há aqueles que lutam muitos anos, e são muito bons:
Porém há os que lutam toda a vida
Estes são os imprescindíveis”
Bertold Brecht

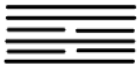
“Wenn ich dieses Wort
ausspreche und dabei
jenes denke, wie ist es
dann überhaupt
möglich, dass du
erkennst, dass ich
jenes denke.”
Kratylos Dialog bei
Platon.

Agradecimentos

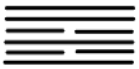
Ao CNPq e a CAPES,
por me proverem financeiramente
em determinados momentos
do meu Doutorado.

Ao Dr. Donizeti, a Carolina, a Cléia, a Denise Queiroz,
enfim, aos amigos e colegas,
que, de alguma maneira,
contribuíram enormemente
para o fechamento deste trabalho
(vocês sabem quem são!).

A minha orientadora, Prof. Dr. Marlene de Oliveira

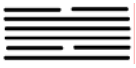


Ao meu orientador na Alemanha,
Prof. Dr. Henning Lobin
(Alles Gute!).



Aos amigos da Uni-Gießen e de Gießen
(bis bald…).

À minha família,
que sempre me ensinou
a valorizar a Cultura
e a Educação.



À Denise,
companheira imprescindível!

Resumo

Este trabalho apresenta estudos para a administração de documentos eletrônicos através de um embasamento cognitivista. Propõe-se uma indexação de textos eletrônicos, disponibilizados em língua portuguesa, por meio da aplicação de teorias de lingüística computacional e utilização de ontologia. A técnica empregada para a indexação é baseada principalmente na teoria de Análise Proposicional proposta por Frederiksen (1975). É baseada na extração de etiquetas sintáticas das palavras que compõem os documentos para a geração de etiquetas semânticas dessas palavras, para então gerar uma ontologia leve automaticamente.

Ao longo deste trabalho são sugeridas várias contribuições, que visam otimizar o desempenho de Sistemas de Recuperação de Informação, por meio da utilização de técnicas que permitam contextualizar as palavras dos textos a serem indexados. Tais contribuições incluem desde a otimização de analisadores sintáticos, até a geração automática de ontologias leves.

Inicialmente um corpus, que é uma pequena coleção de documentos eletrônicos sobre Ciência da Informação, escritos em língua portuguesa e disponibilizados na Web, foi criada. Esta coleção foi utilizada para testar o protótipo construído. O protótipo, nomeado *SiRILiCO* (Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia), foi utilizado em um experimento-piloto e posteriormente em outro experimento, com o intuito de verificar e validar a hipótese de que é possível desenvolver e implementar um Sistema de Recuperação de Informação totalmente baseado em teorias lingüísticas, teorias de lingüística computacional e ontologia.

Os resultados de precisão e revocação dos experimentos realizados com o *SiRILiCO* são comparados com os resultados obtidos com a utilização de um modelo vetorial. A análise dos resultados sugere que não só é viável a hipótese defendida como também é muito promissora.

Abstract

This work presents studies for the administration of electronic documents using a cognitive approach. We propose an automatic index generation of eletronic texts write in Brazilian Portuguese using linguistic theories, theories of computacional linguistics and ontology. The technique used to create the index is based mainly on the theory of Proposicional Analysis proposed by Frederiksen (1975) and it is based on the extraction of syntactic labels of the words that compose the documents for the generation of semantic labels of those words, for then to generate a lightweight ontology automatically.

We suggest, during this work, several contribuitions to improve the Information Retrieval System's performance, using several techniques that allow context words of indexing texts. Such contributions include optimize syntactic parsers, as well as the automatic generation of lightweight ontologies.

Initially a corpus, a small collection of electronic documents about Information Science, written in Brazilian Portuguese and available in the Web, was created. This collection was used to test the prototype. The prototype, nominated *SiRILiCO* (Information Retrieval System based on Computacional LinguisticTheories and Ontology), was used in a first experiment and later in an experiment to verify and to validate the hypothesis that is possible to develop and to implement an Information Retrieval System totally based on linguistic theories, theories of computacional linguistics and ontology.

The *SiRILiCO's* experiments results of precision and recall are compared with the results obtained with the use of a vectorial model. The analysis of the results suggests that not only it is a possible hypothesis as well as it is very promising.

Sumário

Lista de Figuras

Lista de Quadros e Tabelas

1 Introdução

O acelerado desenvolvimento cientifico e tecnológico e também de outros tipos de conhecimentos demanda métodos mais adequados na organização, tratamento e recuperação da informação. Essa necessidade adquire maiores contornos quando dirigida para serviços da Internet. São vários os serviços disponíveis na rede, contudo, a World Wide Web pode ser considerado mais importante. A Rede Mundial de Computadores (ou simplesmente Internet) é considerada um repositório de informações de uma grandeza incomensurável. Há estimativas de que ela cresça exponencialmente, dobrando de tamanho a cada seis meses. Em 1999 existiam mais de 800 milhões de páginas HTML (LAWRENCE & GILES, 1999), com mais de um bilhão de conexões, hiper-vínculos, que as agregam, contendo mais de sete Terabytes de informação (CHAKRABARTI. *et al.*, 1998; CHAKRABARTI, *et al.*, 1999; AIRES & SANTOS, 2002; BAEZA-YATES & CASTILLO, 2004; BALMIN et al., 2004; COSTA & FRASCONI, 2004). Atualmente, os valores referentes à Web impressionam qualquer estudioso das áreas da Ciência da Informação, da Ciência da Computação e até mesmo da Filosofia (LÈVY, 1995; CAPURRO, 2003). Trata-se de um depósito de conhecimento com um volume jamais sonhado anteriormente pela humanidade. Estima-se que 95% das páginas de documentos disponibilizados na Web hoje sejam de relativo acesso, à chamada “Web Livre” (O’NEILL et al., 2003), isto somente na parte denominada de ‘Web superficial’ (“surface web”), pois na ‘Web profunda’ (“deep Web”) calcula-se que existam mais de 550 bilhões de documentos incluindo-se Intranets e Bancos de Dados corporativos cujo acesso é restrito (BRIGHTPLANET, 2000). Somente o Google, hoje a mais eficiente máquina de busca em atividade, indexa mais de três bilhões de páginas Web (GOOGLE, 2005). Essa rica coleção do saber humano (a Web) está disponibilizada em mais de 59.100.880 Web sites (WEB SERVER SURVEY, Fevereiro, 2005) e, teoricamente, acessível a qualquer cidadão do mundo (INCLUSÃO DIGITAL, 2005). Esta é a realização da primeira parte da Revolução da Informação, o acesso instantâneo do indivíduo à informação (LÉVY, 1995; NEGROPONTE, 1995; DERTOUZOS, 1997; TAKAHASHI, 2000). Porém, tamanha coleção gera enormes problemas (WITTEN, MOFFAT & BELL, 1999; LYMAN and VARIAN, 2000). Esta informação que está acessível a todos está sendo gerada e manipulada em vários idiomas e estilos e representa a expressão de várias culturas, ideologias, crenças, etc. Como a informação vem sendo organizada? Como tratá-la e recuperá-la?

Com essa diversidade, quantidade e tipos de informação, é necessário criar-se novas metodologias e técnicas de organização e recuperação. O avanço tecnológico permite que hoje em dia existam sistemas de informação mais modernos que utilizam-se da linguagem natural na recuperação da informação, aproximando cada vez mais os usuários das informações que eles necessitam.

1.2. A Recuperação da Informação Hoje

A Recuperação de Informação (RI), de uma maneira geral, é um campo do conhecimento humano que se ocupa de dois conceitos (MEADOW, 1992):

Como representar a informação;

Como interpretar a estrutura que representa a informação.

A RI envolve processo de seleção. É um processo de comunicação que pode ser entendido como a interface entre o autor/produtor de informação e o leitor/usuário de informação. Para Blair (1990) o problema central da RI está em como representar os documentos (a fonte de informação) para poder recuperá-los.

Atualmente, a RI é feita de maneira mais rápida no meio digital. Suas metodologias baseiam-se principalmente em abordagens quantitativas, fundamentadas em estatística e matemática (SALTON & McGILL, 1983). Estas abordagens, quando empregadas na Web, não permitem identificar e extrair a semântica do conteúdo de suas páginas (ERDMANN et al., 2001; DAVIES et al., 2003). Os modelos quantitativos (ROBERTSON, 1977) não conseguem solucionar problemas relacionados com o texto, como por exemplo, sinonímias (para um usuário comum temos “Web” e “Internet”, por exemplo). Esses modelos normalmente indexam os documentos capturados na Web por meio da freqüência de ocorrências de uma palavra, ou seja, o clássico modelo vetorial com suas otimizações (PERSIN, 1994; PERSIN et al., 1996). Com o crescimento do volume de informações disponibilizadas em meio digital (LYMAN and VARIAN, 2000), assim como a importância cada vez maior da Web, outras abordagens que visam melhorar a qualidade do processo de Recuperação da Informação por meio da extração de conteúdos semânticos (SALTON, 1973; CROFT et al., 1991) podem contribuir para otimizar a qualidade de resposta dos Sistemas de Recuperação de Informação (SRI). É proposta desse estudo colaborar nessa direção.

1.3 Porque Utilizar Processamento de Linguagem Natural na Recuperação de Informação

O conhecimento sobre o idioma é necessário para discriminar as diversas expressões inerentes à língua que podem e são utilizadas pelos autores. Como por exemplo, “o sistema de informação” e “a informação do sistema”. As seqüências de palavras são sintagmas nominais, que são importantes descritores para os Sistemas de Recuperação de Informação (KURAMOTO, 1999; MIORELLI, 2001). Observações simples como estas podem justificar o emprego de Processamento de Linguagem Natural (PLN), objetivando melhorias em um Sistema de Recuperação de Informação. Embora o PLN seja empregado em muitos SRI, o nível de aplicação ainda é muito pequeno. Normalmente os SRIs não utilizam PLN, mas técnicas diretas, como a extração de sentenças ou de radicais de palavras combinadas com técnicas estatísticas.

A contribuição do PLN para a Recuperação de Informação (RI), nesses sistemas, limita-se a identificar frases que possam ser sintagmas nominais, ou que tenham algum conteúdo semântico, por meio de técnicas relativamente simples. A razão para a ausência de técnicas de PLN mais sofisticadas nesses sistemas, reside nas dificuldades da aplicação de Lingüística Computacional ao processamento de textos. Tais dificuldades já foram identificadas e tratadas por muitos autores, que citam a ineficiência, a cobertura limitada e os custos elevados como as principais dificuldades para a construção de léxicos e bases de conhecimento de domínios específicos (STRZALKOWSKI et al., 1998). Porém, a principal questão é que os Sistemas de Recuperação de Informação que utilizam Processamento de Linguagem Natural não apresentam vantagem em relação aos outros sistemas tradicionais, que utilizam apenas abordagens qualitativas (SMEATON, 1999; STRZALKOWSKI et al., 1998).

SMEATON (1997) afirma que o “bater de cabeças” entre a Recuperação de Informação e o Processamento de Linguagem Natural deve-se ao fato de que RI e PLN são disciplinas inerentemente diferentes, o que conduz a resultados bastante imprecisos (RI) e precisos (PLN). Baseando-se nesta visão, supõe-se que o Processamento de Linguagem Natural não pode efetivamente contribuir para a Recuperação da Informação. A princípio a premissa de Smeaton pode ser considerada verdadeira, mas a suposta contribuição relativamente pobre do PLN para o RI sugere que os sistemas de RI atuais não necessitam de

um Processamento de Linguagem Natural realmente avançado. A maioria dos sistemas de RI que empregam PLN o utilizam apenas para a criação de índice da extração de substantivos de frases. Isto não pode ser considerado uma tarefa avançada de PLN. O objetivo de se utilizar PLN em um SRI é enriquecer a representação dos textos utilizando-se da construção de modelos que permitam às pessoas escreverem programas de computador capazes de desempenhar atividades envolvendo linguagem natural de maneira mais "inteligente" (ZHOU & ZHANG, 2003). Portanto, não é surpresa a afirmativa de que as técnicas de PLN utilizadas atualmente não contribuem significativamente para a melhoria do processo de indexação. Afinal, os índices da maioria dos sistemas de RI podem ser facilmente produzidos, com ou sem auxílio de PLN, visto que requerem apenas a geração de listas invertidas (BAEZA-YATES & RIBEIRO-NETO, 1999).

1.4. Objetivo e Justificativa

Esta pesquisa tem por objetivo propor o desenvolvimento de um SRI automatizado, que utiliza teorias da Lingüística Computacional e Ontologia. Presume-se que um SRI elaborado desta forma seja efetivamente mais eficiente que os sistemas atuais, no quesito qualidade de resposta, uma vez que a geração de índices a partir de conceitos estruturados (uma ontologia), como será discutida ao longo deste trabalho, é permitida, empregando-se PLN. A utilização de ontologias e o desenvolvimento das mesmas também são fatores preponderantes para o investimento em abordagens que se propõem a serem mais eficazes que as atuais (FAURE & NEDELLEC, 1998; TODIRASCU, 2001).

A princípio, um questionamento referente ao desenvolvimento de um SRI utilizando abordagens que envolvam Lingüística Computacional é o fato de terem custo computacional elevado. Entretanto, devido ao aumento exponencial da informação disponível em meio digital (WOLFRAM, 2000), nota-se que é necessário o desenvolvimento e aperfeiçoamento de tais sistemas, pois os sistemas de abordagem estatística, mesmo com o auxílio do "julgamento humano" (KLEINBERG, 1998; PAGE et al., 1998), aparentemente não mais apresentam respostas satisfatórias às necessidades dos usuários dos sistemas (MENG et al., 1999; LAWRENCE 2000; LIU & CHIN, 2001).

1.4.1 Objetivo Geral

Desenvolvimento de um Sistema de Recuperação de Informação, para o tratamento de documentos em língua portuguesa do Brasil, utilizando técnicas de Lingüística Computacional e Ontologia.

1.4.2 Objetivos Específicos

- Criação de um modelo padrão de Ontologia.
- Uso das categorias de Frederiksen aliado ao modelo padrão de Ontologia.
- Definição dos procedimentos de análise sintática e análise semântica para determinação do modelo final.
- Geração de um índice representando os conceitos extraídos da coleção.

1.5. Organização da Tese

Este trabalho está estruturado em capítulos. A presente seção, **Capítulo 1 - Introdução**, apresenta o problema que motivou este estudo, bem como o objetivo e a justificativa para tratá-lo, através das abordagens aqui defendidas.

No **Capítulo 2 - Referencial Teórico, Recuperação de Informação, Processamento de Linguagem Natural e Ontologia** aborda-se o referencial teórico, articulam-se as idéias da Recuperação de Informação, do Processamento de Linguagem Natural e de Ontologia. Os conceitos básicos são apresentados e as técnicas de Processamento de Linguagem Natural e de Ontologia são descritas.

No **Capítulo 3 - Metodologia** descreve-se a metodologia empregada para o desenvolvimento do trabalho. O modelo de Sistema de Recuperação de Informação desenvolvido é explicitado, cada módulo que o compõe é pormenorizado.

Os resultados obtidos com a utilização de um Sistema de Recuperação de Informação baseado no Modelo Vetorial e na utilização de um Sistema de Recuperação de informação baseado na abordagem defendida neste estudo, o Modelo *SiRILiCO*, são apresentados em forma de tabelas e gráficos no **capítulo 4 - Resultados**.

No **capítulo 5 - Discussão e Considerações Finais** discutem-se as propostas e os resultados obtidos na utilização de ambos os modelos. Além disso, as contribuições advindas desse estudo, bem como as prováveis conseqüências do mesmo, são apresentadas.

2 Referencial Teórico, Recuperação de Informação, Processamento de Linguagem Natural e Ontologia

2.1. Introdução

O referencial teórico desta pesquisa foi construído com base em autores das temáticas: Recuperação da Informação, Processamento de Linguagem Natural e Ontologia. Tais autores são oriundos da Ciência da Computação, da Ciência da Informação e da Lingüística. Esse procedimento foi necessário considerando-se a inerente interdisciplinaridade do tipo de pesquisa desenvolvido. Assim, procurou-se explicitar as abordagens e os conceitos de Recuperação de Informação, Processamento de Linguagem Natural e Ontologia. Embora essas três temáticas da ciência sejam amplas, heterogêneas e fragmentadas, sob a ótica da investigação cientifica, apresentamos uma sucinta introdução que evidencia a importância de cada um desses ramos de conhecimento para o estudo em questão.

2.1.1. Recuperação da Informação: Gênese e Desenvolvimento

A Recuperação da Informação é uma temática investigada tanto pela Ciência da Informação quanto pela Ciência da Computação. A Recuperação da Informação tornou-se tema de estudo da Ciência da Informação desde a sua origem, quando essa área apontou como sua finalidade a tarefa de tornar acessível um acervo crescente de registros de conhecimentos.

Em 1945, Vanevar Bush, um respeitado cientista do MIT (Massachussets Institute of Technology – USA), publica seu famoso artigo “As We May Think”, onde identifica e define o problema de tornar acessível o acervo crescente de conhecimentos e propõe uma solução. Descreve, então, o Memex, um dispositivo teórico que seria o precursor dos computadores atuais. No mesmo ano, Shannon & Weaver publicam a Teoria Matemática da Comunicação e, em 1949, Weaver escreve sobre a Informação. A Ciência da Informação usufrui dessas teorias, notadamente para começar a estudar e entender seu objeto, a Informação.

Outra contribuição oriunda das teorias de Shanon e Weaver, referem-se aos conceitos relacionados às medidas de desempenho de um Sistema de Recuperação:

- Relevância;
- Redundância;
- Precisão;
- Revocação;
- Taxa de Incerteza;

Existem outras medidas de avaliação de Sistemas de Recuperação, sendo que as mais disseminadas, sob a ótica da Ciência da Computação, são as medidas de **Precisão**, que é a fração de documentos da coleção que já foram examinados e que são relevantes para uma busca específica, e **Revogação**, que é a fração de documentos da coleção, dentre os que já foram examinados para uma busca específica, e que são relevantes (FRAKES & BAEZA-YATES, 1992; GEY, 1992; FERNEDA, 2003).

Agregado aos estudos da informação em 1951, Mooers cria o termo Recuperação da Informação e define os problemas a serem abordados por esta nova disciplina. Segundo esse autor:

> *...a recuperação da informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação* (MOOERS, 1951).

Surgiram então os estudos de Recuperação da Informação. Um marco desses estudos foi a realização dos primeiros testes de desempenho de um Sistema de Recuperação de Informação. O Sistema de Recuperação de Informação avaliado era denominado “Uniterm” (Unitermo) e foi criado por Mortimer Taube. Testes foram desenvolvidos por organismos dos Estados Unidos (Armed Services Technical Information Agency – ASTIA) e do Reino Unido (College of Aeronautics in Granfield) e foram denominados “Os testes de Granfield”. O sistema previa a representação do documento por termos únicos, retirados do título ou resumo do documento. Este procedimento era inovador na indexação por assunto. Os testes de Granfield foram importantes para a história da Recuperação da Informação, uma vez que estabeleceram embasamento teórico para o desenvolvimento da disciplina (CENDÓN, 2005).

Outro marco importante deve ser creditado ao projeto SMART, que realizou uma série de experimentos sobre a Recuperação da Informação. Estes experimentos foram feitos por Gerard Salton, um cientista que dedicou grande parte de sua vida a este projeto, permitindo o desenvolvimento e o aprimoramento de diversas técnicas computacionais e resultando em um modelo de Recuperação de Informação automatizado (FERNEDA, 2003).

Lesk (1995), em seu artigo intitulado “The Seven Ages of Information Retrieval”, inspirando-se em Shakespeare, propõe a seguinte cronologia para a Recuperação de Informação: Infância (1945-1955); Idade Escolar (anos 60); Maioridade (anos 70); Maturidade (anos 80); Crise da Meia-Idade (anos 90); Realização (anos 2000) e Aposentadoria (2010). O artigo retrata a história da Recuperação de Informação ao longo dessas sete fases, referenciando-se às previsões de Bush (1945), estabelecendo um paralelo entre as mesmas com as diferentes etapas que compõem a vida humana. O quadro 2.1 apresenta as três grandes seções, que compõem a Ciência da Informação, de acordo com a visão de Summers et al. (1999): Ciência da Informação propriamente dita, Gerenciamento da Informação e Tecnologia da Informação.

Quadro 2.1.

Critério para a Ciência da Informação
Seção 1 (Área Núcleo) Ciência da Informação
A teoria e prática de criar, aferir, acessar e validar, organizar, armazenar, transmitir, recuperar e disseminar informação.
Informação: suas características, provedores e usuários
Fontes de informação
Armazenamento e recuperação de informação
Análise da informação
Teoria da Ciência da Informação
Seção 2 Gerenciamento da Informação
O gerenciamento do total dos recursos de informação das organizações.
Planejamento
Comunicações
Gerenciamento das Informações e Sistemas de Controle
Gerenciamento de recursos humanos
Gerenciamento financeiro
Promoção, economia e marketing
Fatores políticos, éticos, sociais e legais
Seção 3 Tecnologia da Informação
Tecnologia que pode ser usada na Ciência da Informação ou Gerenciamento da Informação.
Sistemas de Computadores: hardware, software
Telecomunicações
Aplicações de tecnologia da informação
Ambiente

Critério para a Ciência da Informação de acordo com o *Institute of Information Science.* (Adaptado de SUMMERS et al., 1999).

Na área núcleo e na seção 3 do quadro 2.1 observa-se que a Recuperação da Informação é objeto de estudo da Ciência da Informação.

2.2. Entendimento de Recuperação de Informação

A Recuperação de Informação (RIJSBERGER, 1979; PÔSSAS et al., 2002) é entendida como o clássico problema da recuperação efetiva e eficiente de documentos pertinentes extraídos de uma grande coleção (que nos dias de hoje pode ser entendida como um armazém de informação ou uma base de dados digital) de acordo com uma necessidade de informação específica (ROWLEY, 1996; BAEZA-YATES & RIBEIRO-NETO, 1999 WITTEN et al., 1999). Nas áreas de Biblioteconomia e Ciência da Informação, o conceito "Recuperação de Informação" tem sido utilizado para designar a busca de literatura (LANCASTER & WARNER, 1993). Para a recuperação de informação utilizam-se ferramentas que facilitam essa tarefa. São os sistemas e redes de recuperação de informação (Figura 2.1) (LANCASTER, 1979; MEADOW, 1992) que, atualmente, são automatizados e atividades essencialmente computacionais (CENDÓN, 2005). Neste trabalho, focaliza-se a informação textual, apesar da existência e ampla divulgação de outros formatos de informação digital (como imagens, por exemplo).

Na Figura 2.1 tem-se o cenário de um Sistema de Recuperação de Informação. O sistema é cíclico, a informação geralmente flui no sentido horário do diagrama. Primeiramente, a comunidade de usuários ou potenciais usuários do sistema produzem informação sobre o mundo e esta informação é, por sua vez, utilizada para afetar o mundo (MEADOW, 1992).

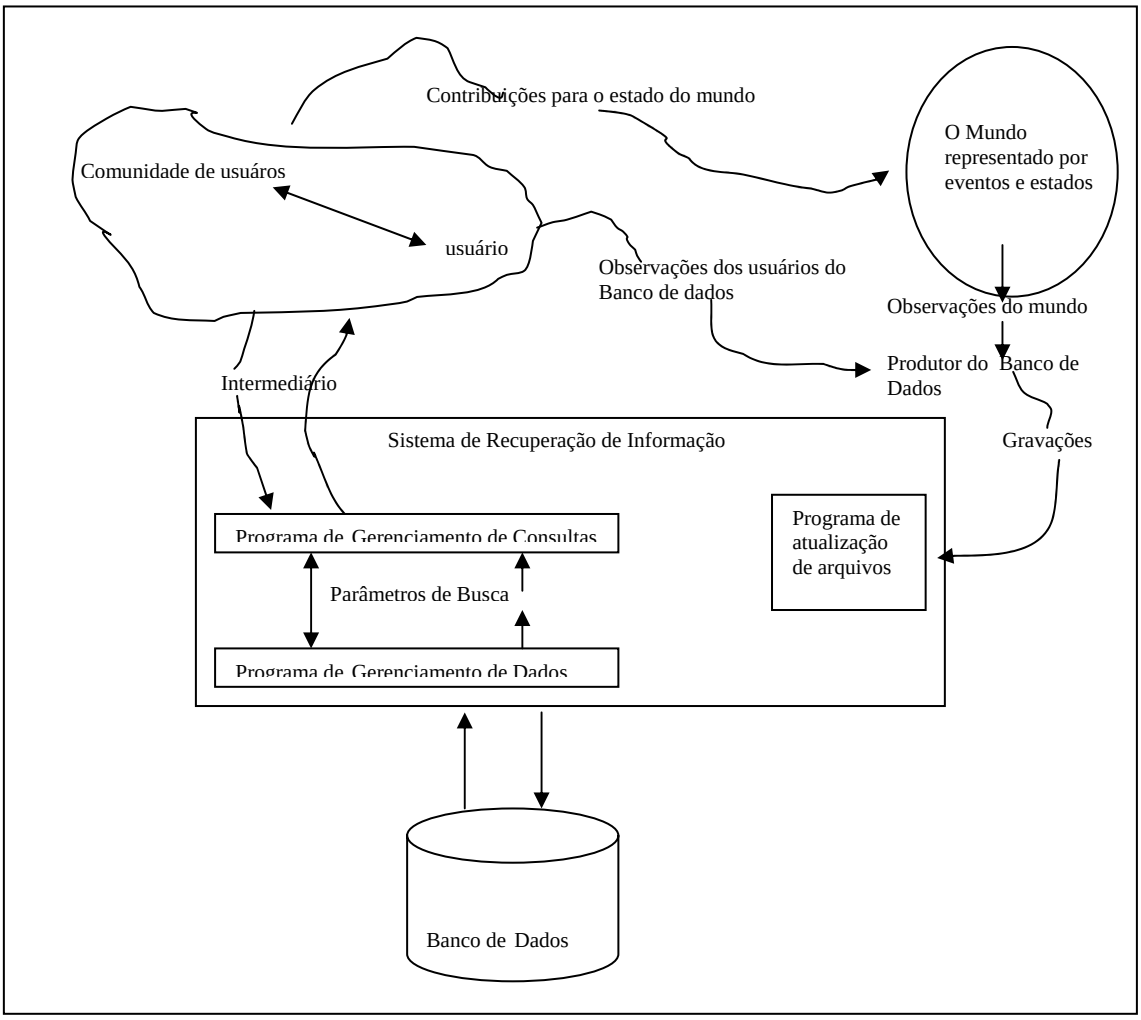


Figura 2.1. Visão Geral do cenário de um Sistema de Recuperação de Informação, adaptado de MEADOW 1992.

Neste estudo, será enfatizada a Recuperação da Informação. A área de Recuperação de Informação (RI) (RIJSBERGEN, 1979; WITTEN, et al. 1999; GARFIELD, 2001) cresceu muito em importância, particularmente devido ao aumento da disponibilidade de informação em formato digital, tais como: bancos de dados (HONKELA et al., 1996), bibliotecas digitais (NÜRNBERG et al., 1995; ARMS, 2000), publicações eletrônicas (ROSENFELD & MORVILLE; 2002), documentos digitalizados (ARAUJO & GUIMARÃES, 2000) e a Web (BERNERS-LEE et al., 2001). O projeto de RI aqui discutido aponta para o desenvolvimento de novas técnicas de indexação, baseadas em técnicas utilizadas em diferentes áreas, objetivando otimizar a qualidade da informação recuperada.

A Recuperação de Informação (GUDIVADA et al., 1997) consiste de três processos: **Coleta** (LAENDER et al., 2002); **Indexação** (LAWRENCE et al., 1999; CLEVELAND & CLEVELAND, 2000; LAHTINEN, 2000) e **Ordenação** (LOSADA & BARREIRO, 2000; GLOVER et al., 2002). Enquanto o processo de coleta em meio digital está basicamente resolvido, com a utilização de robôs virtuais (ZIVIANI et al., 1999), os dois últimos não estão.

A indexação é um processo de categorização. Consiste em nomear palavras-chave de um documento (as palavras-chave são a representação do documento). Para isso pode-se utilizar um vocabulário, derivado de uma linguagem própria, um vocabulário controlado ou uma linguagem documental (GUINCHAT e MENOU, 1994). A indexação é um dos principais processos cognitivos desenvolvidos em Bibliotecas e Unidades de Informação (OLIVEIRA e ARAUJO, 2005).

O processo de ordenação consiste em disponibilizar os documentos de acordo com uma graduação que condiz com as representações que satisfaçam às necessidades do usuário. A ordenação se dá por relevância (SPERBER & WILSON, 1986; GREISDORF, 2000). Há uma graduação entre os documentos que atendem às necessidades informacionais do sujeito/usuário. A relevância de um determinado documento depende da inferência[1] do classificador. Independente do processo ser manual ou automático, e sim da política adotada antes da geração da ordenação (ALLEN, 1996). Ellis (1996) apresenta o julgamento de relevância de um documento, por meio do exemplo de algumas propostas e testes realizados com as mesmas, como uma tarefa complexa e subjetiva. Portanto, tal qual é apresentada, é inviável do ponto de vista científico.

Ranganathan (1933) idealizou um sistema de classificação em facetas. As facetas podem ser entendidas como categorias práticas, que são preenchidas com termos pertinentes à classificação, e tal preenchimento dá-se de maneira correta. Uma classificação facetada é um grupo de categorias mutuamente exclusivas e conjuntamente exaustivas, sendo que cada categoria isola uma perspectiva única (a faceta) do item classificado. Combinadas, as facetas promovem a descrição completa de todo e qualquer objeto (DENTON, 2003). O sistema baseia-se em conceitos básicos, que permitem que todo e qualquer conhecimento seja estruturado. Este princípio, o princípio da análise facetada, consiste em dividir os assuntos em seus vários componentes. O princípio da síntese consiste em combinar os componentes

[1] Inferência é um processo cognitivo por meio do qual afirma-se a verdade de uma proposição baseando-se em outras reconhecidamente verdadeiras e coligadas a mesma.

objetivando a adequada classificação de um documento. Ranganathan (1985) criou uma ordem de citação de assuntos utilizando cinco categorias distintas:

- Personalidade/Entidade (Personality/Entity);
- Matéria (Matter);
- Energia (Energy);
- Espaço (Space);
- Tempo (Time);

Com suas teorias Ranganathan propunha a indexação de todo e qualquer documento por meio da categorização por assunto[2], salientando que existe um número ilimitado de assuntos específicos que podem ser utilizados para esta classificação. Para Ranganathan, o princípio de relevância objetiva garantir a autenticidade das facetas, ou seja, que as facetas definidas sejam a proposta, o assunto e o escopo do tema a ser tratado.

Um SRI, apesar da evolução dos sistemas de hardware, não pode incorporar todo o conteúdo da informação de uma coleção de documentos, porque isto compromete a eficiência do sistema, que trata de processar a representação do conteúdo informacional da coleção. Indexar normalmente é a determinação da representação do documento. A conversão da consulta do usuário para uma representação também é possível, permitindo assim o processo de ordenação dos documentos contidos na coleção em função da consulta do usuário. Assim, a forma e a maneira de representar a informação para a devida disponibilização é fundamental para todo o SRI.

[2] Assunto, de acordo com Ranganathan, é um corpo de idéias organizadas ou sistematizadas.

2.2.1. Recuperação de Informação na Web

Como acessar a informação necessária no menor tempo possível e de maneira eficiente? As técnicas clássicas de Recuperação de Informação, quando aplicadas na Web, são ineficientes e insipientes, isso devido ao tamanho e heterogeneidade da "coleção" disponível (HUANG, 2000).

Atualmente, mudou-se o enfoque do problema. A questão não é mais saber se determinada informação encontra-se disponível na Web, mas saber a localização exata dessa informação. As máquinas de busca (BRIN & PAGE, 1998; RAKHSHAN et al., 2003) (ALTAVISTA, TODOBR, por exemplo) e os diretórios (CADÊ, YAHOO, por exemplo) (BAEZA-YATES & RIBEIRO-NETO, 1999; CENDÓN, 2001) geralmente indexam e recuperam as páginas Web (normalmente em HTML, porém já existem muitas páginas em outros formatos, como XML, por exemplo) (ALMEIDA, 2002), baseando-se somente no texto e desprezando os vínculos que foram construídos pelos autores das páginas.

Entretanto, o Google (GOOGLE, 2005) utiliza-se do julgamento humano (através das escolhas dos seus usuários) para reordenar as páginas apresentadas em uma determinada consulta. Como funcionam tais "depósitos de informação?" Os diretórios são construídos através de julgamento humano. Há profissionais, especialistas de várias áreas do conhecimento humano, conjuntamente com especialistas da tecnologia da informação, que navegam pela Web coletando e indexando as páginas Web; isto restringe extremamente o tamanho dos diretórios e o tempo de atualização dos mesmos.

Por outro lado, as máquinas de busca coletam e indexam as páginas disponíveis na Web através da utilização de robôs coletores (PINKERTON, 1994; KOSTER, 1995; CHO et al., 1998; ZIVIANI et al., 1999). Esses robôs são programas que percorrem a estrutura de hipertexto da Web recuperando páginas HTML. Após a coleta, as páginas são armazenadas e indexadas. O armazenamento e a indexação se dão de várias maneiras (WITTEN et al., 1999; FRAKES & BAEZA-YATES, 1992; BAEZA-YATES & RIBEIRO-NETO, 1999). Normalmente utiliza-se de uma estrutura de arquivo invertido.

Como tipos de estruturas de dados para a construção de arquivos invertidos podem ser citados:

1 **Arranjo Ordenado.** Obtém-se o arranjo ordenado através da leitura do texto, com a identificação das palavras-chave (todas as palavras distintas do texto) que são armazenadas numa estrutura contendo pelo menos a palavra e o endereço (posição) da mesma no arquivo texto, para que se possa recuperá-la. Obtido esse arranjo, deve-se fazer a ordenação lexicográfica do mesmo. A seguir, numa terceira etapa, as duplicações das palavras-chave são removidas e incorporadas a uma lista de ocorrências desta chave (a palavra).

Vantagens: este método apresenta facilidade de implementação e boa velocidade na recuperação dos dados.

Desvantagens: este método apresenta dificuldade para a inclusão de uma nova chave (palavra) e alto custo de espaço durante o processo de montagem da lista.

Tabela 2.1 Arranjo Ordenado.

Número	Termo	(Documento; Palavras)
1	Claro	(1; 3), (3; 10), (4; 5)
2	Informação	(3; 5), (4; 2), (5; 7)
3	Ordenação	(2; 4), (5; 11), (9; 3). (11; 15)
4	Relevância	(5; 7), (7; 4), (11; 4). (44; 30)
5	Estudo	(1; 4), (3; 15)

2 **Árvores B.** Constrói-se uma estrutura em árvores B (com as palavras-chave).

Vantagens: este método apresenta facilidade para inclusão de uma nova chave e boa velocidade de busca, principalmente se o arquivo invertido é mantido em uma memória secundária.

Desvantagens: este método requer maior espaço para a estrutura, se comparado com o método de arranjo ordenado.

3 **Árvores Trie.** Monta-se uma estrutura em árvore Trie através da decomposição digital das palavras. Destaca-se nessa categoria a utilização de Árvores Patricia, uma árvore cuja construção é baseada na representação binária dos caracteres ASCII que compõem o texto.

Vantagens: este método apresenta facilidade para efetuar buscas aproximadas de expressões regulares.

Desvantagens: este método utiliza um tipo de árvore de difícil geração.

4 **Estruturas com Hashing** (método de transformação de chave). Por meio de uma função, transforma-se a chave de pesquisa em um endereço de tabela.

Vantagens: este método permite a redução do tamanho do arquivo de índice e apresenta rápida recuperação.

Desvantagens: este método é difícil de implementar, pois pode apresentar colisões entre as chaves.

Como modelos de indexação de documentos eletrônicos podemos citar:

1 O modelo Booleano (SALTON & McGILL, 1983). Primeiro método utilizado e até hoje muito difundido, não ordena os documentos. Simplesmente apresenta como resposta, uma lista com os documentos que contém o termo da consulta. Para operações com os termos, pode-se utilizar técnicas de *stemming* (raiz gramatical), que consiste na redução das palavras às suas raízes. As operações com os documentos envolve *parsing*, que consiste na "quebra" do documento em seus vários elementos constituintes. Esse modelo é utilizado em computadores com arquitetura Von Neuman, que consistem em máquinas que tem apenas um processador, que executam instruções sequencialmente e que são destinadas a uso geral.

2 O modelo Booleano Estendido (SALTON et al., 1983). Como o nome indica, uma extensão do modelo booleano. No modelo tradicional, perante uma cláusula [*X and* Y], seriam retornados apenas os documentos que satisfizessem ambas as condições. No modelo estendido pode-se atribuir um valor de importância aos conectivos *and* e *or*, tal que, mesmo que um elemento satisfaça apenas a condição *X*, este será também retornado, embora com um valor de ordenação mais baixo que um outro que satisfaça ambas as condições. Esse modelo é mais adequado ao problema das consultas aproximadas.

3 **O modelo Probabilístico** (FUHR, 1989). Baseia-se na idéia de que dado um termo de consulta estima-se as distâncias semânticas entre o termo e os documentos que contém o termo, para através de interação com o usuário determinar a relevância dos documentos. A estrutura desse modelo é baseada em *signatures* (assinaturas), que consistem de padrões de bits que representam os documentos. A operação com os documentos se dá através da utilização de *clusters* (agrupamentos). Agrupam-se os documentos através do grau de similaridade entre os mesmos.

4 **O modelo String Search** (FRAKES & BAEZA-YATES,1992). Simplesmente faz casamento de padrões, ou seja, dada uma "*string*" (uma frase, por exemplo), ele retorna todos os documentos que contém a *string*. A operação com os termos se dá através de Tesaurus (RUGE, 1997) e é otimizada através da eliminação *stopwords* (palavras mais freqüentes e que não apresentam valor semântico, artigos p. e.) utilizando-se de *stoplists* (listas contendo *stopwords*).

5 **O modelo Vetorial.** Ordena os documentos. Para isto realiza uma filtragem (PERSIN, 1994; PERSIN et al., 1996). Portanto, considera-se o tamanho do texto para calcular a freqüência do termo no mesmo.

No quadro 2.2 (adaptado de FRAKES & BAEZA-YATES, 1992) apresentam-se, de maneira sucinta, os métodos mais comuns de indexação e ordenação de documentos eletrônicos (de acordo com os autores para Sistemas de Recuperação de Informação em geral, incluindo documentos Web) e suas características:

Quadro 2.2

Modelo	**Estrutura**	**Operações c/ queries**	**Operações c/ termos**	**Operações c/ documentos**	**Hardware**
Booleano	Arquivos lineares	Feedback	Stem	Parse	VonNeuman
Booleano extendido	Arquivos invertidos	Parse	Peso	Apresentar	Paralelo
Probabilístico	Assinaturas	Booleanas	Thesaurus	Cluster	Específico de RI
String search	Árvores PAT	Cluster	Stoplists	Ordenar	Discos óticos
Vetorial	Grafos Hashing		Truncagem	Sort Atribuir IDs	Discos Magnéticos

Classificação de Sistemas de Recuperação de Informação (adaptado de FRAKES & BAEZA-YATES, 1992, p. 02).

Como citado anteriormente, a melhor maneira de armazenar os documentos Web é através da geração de um arquivo invertido, pois o arquivo invertido permite uma rápida recuperação do(s) documento(s) que contém o(s) termo(s) da busca.

2.3. Lingüística e Lingüística Computacional

A Lingüística é o ramo da Ciência que estuda a linguagem e todas as suas implicações (CHOMSKY, 1995; SPERBER & WILSON, 1995; FARIA, 1998). A Lingüística Computacional é entendida como a utilização de conhecimentos sobre a língua e a comunicação humana, tanto para comunicação com sistemas computacionais como para melhorar a comunicação entre seres humanos (SANTOS, 2001).

2.3.1. Conceitos básicos

Existem vários meios disponíveis para a veiculação de linguagem humana, dentre os quais o texto é o mais utilizado para a veiculação de informação digital. O texto pode ser entendido como uma macro unidade, composta de informações de diversas naturezas, presente na estrutura de uma língua natural (o português do Brasil, por exemplo). Resumidamente e focando a questão da pertinência do texto para a recuperação da informação pode-se assumir que o texto é o lugar, o centro comum que se faz no processo de interação entre autor e leitor (MEDEIROS, 1992).

O texto pode apresentar-se de duas formas: verbal ou escrito. Quando escrito, o texto é, ainda, o meio visual verbal mais utilizado para a veiculação de conhecimentos (GOTTSCHALG-DUQUE, 1998). Ele apresenta outras informações, como por exemplo, o material necessário para a produção de representações mentais na forma de letras e palavras. O leitor percebe visualmente que as letras e as palavras são traços verticais e horizontais, diagonais, circulares, etc; e que tais palavras pertencem à sua língua ou são passíveis de a ela pertencerem (desinências e sufixos das palavras são fortes indícios) (GOTTSCHALG-DUQUE & DILLINGER, 1994). As informações ativadas no acesso lexical permitem ao leitor a construção da estrutura sintática das frases orações e período.

A habilidade de relacionar estas informações, provenientes dos diferentes constituintes[3] do texto, é importante para a leitura (BAUMANN, 1987), além de ser uma parte relevante do processo de compreensão da linguagem. Essa habilidade permite ao leitor processar e armazenar informações dos constituintes advindos do que se está lendo, e, concomitantemente, fomenta os recursos computacionais necessários para o processamento dos constituintes que estão por vir (JUST & CARPENTER, 1992). A coesão textual (HALLIDAY & HASAN, 1976; KOCH, 1989, 1997) é que auxilia essa habilidade do leitor.

Essas informações, que a princípio são processadas apenas por seres humanos, têm hoje extrema importância para o tratamento digital da informação (HIEMSTRA, 2001). Os documentos e as expressões de busca são objetos lingüísticos e a utilização de embasamento lingüístico no tratamento da informação. No caso da recuperação da informação, visa à obtenção de um processamento de texto totalmente automático (CANCEDDA et al., 2003). Para a indexação de documentos (ARAUJO & LUNA, 2002), por exemplo, tem-se utilizado teorias oriundas da Lingüística Textual (ou Lingüística do Texto) (MARCUSCHI, 1983; KOCH & MONTEIRO, 1998; LOBIN, 2003).

A contribuição da Lingüística e da Lingüística Computacional para a Biblioteconomia e para a Ciência da Informação não é recente. Mesmo no português, especificamente no português do Brasil, existem contribuições em pelo menos sete grupos temáticos referentes a estas áreas do saber (MENDONÇA, 2000), como apresentado no quadro 2.3.

Quadro 2.3.

Grupos Temáticos	**Características**
1- Teórico	Abordagem Textual.
2- Quantitativo	Lingüístico e Bibliométrico.
3- Temático	Processamento Intelectual, abordagem semântica, conceitual e terminológica.
4- Aplicativo	Projetos e modelos de indexação automática e linguagem natural.
5- Ensino	Relações curriculares.
6- Tecnológico	Sistemas especialistas e inteligência artificial.
7- Normativo	Lingüística e classificação decimal universal

Interface Lingüística/Ciência da Informação, Organização do Conhecimento. (Adaptado de MENDONÇA, 2000).

[3] Constituintes do texto são os elementos semânticos que o compõem.

Baseado na abordagem apresentada neste quadro, este trabalho pode ser identificado como pertencente ao Grupo 4, que inclui estudos de indexação automática e de linguagem natural. A finalidade do grupo é a de "apresentar análise e avaliação dos sistemas de indexação mediados pelo uso do computador e que se pautam na área da lingüística como área de apoio à sua evolução", que é uma questão contemplada na presente pesquisa.

2.4. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN), pode ser entendido como um ramo da Lingüística[4] que estuda a geração e recepção automática de textos, ou seja, fazer com que máquinas sejam capazes de "ler", "escrever" e "traduzir" textos. Esse anseio científico tem como um dos marcos acadêmicos o texto 'Translation', também conhecido como "Weaver Memorandum", escrito por Warren Weaver em 1949 onde convidava universidades e empresas para desenvolverem projetos de "tradução automática" (MT, Machine Translation).

Na recuperação de informação o PLN já é utilizado como auxílio à otimização de modelos de SRI já consagrados (FALOUTSOS & OARD, 1995; SMEATON, 1997; 1999), porém ainda há muitas possibilidades para a utilização de PLN em RI (DING, 2001; REHM, 2002).

O Processamento de Linguagem Natural (RANCHHOD, 2001) assim como a própria Lingüística e a Ciência da Informação (SCHRADER, 1986: BATES, 1999), abarca várias áreas do conhecimento. Atualmente, devido à sua complexidade, o Processamento de Linguagem Natural subdivide-se em vários tópicos, porém, para efeito deste trabalho, abordamos especificamente os tópicos referentes ao processamento de textos (língua escrita), e ao processamento de textos digitais em língua portuguesa, português do Brasil.

[4] Embora também conte com outras áreas complexas do conhecimento humano, como a psicologia, filosofia, sociologia e comunicação social, por exemplo.

2.4.1. Conceitos básicos

O Processamento de Linguagem Natural surgiu na década de 50 e sua origem é Atribuida a Weaver, por meio do já mencionado “Weaver Memorandum”. Esses primeiros estudos eram embrionários e em sua grande maioria sequer demonstravam alguma fundamentação oriunda da lingüística.

Apesar do descrédito perante a comunidade científica, o Processamento de Linguagem Natural continuou sendo um fértil campo de estudos para pesquisadores impetuosos. Em 1970 Winograd, um estudante de doutorado do MIT, publicou sua Tese que propunha um sistema denominado SHRDLU, que simulava, por meio de uma representação gráfica apresentada no monitor do computador, a movimentação de um braço mecânico sobre a superfície de uma mesa. Esse sistema permitia que o usuário “conversasse” com a máquina utilizando-se de instruções em linguagem natural (inglês), ele demonstrou que fazer pesquisas em Processamento de Linguagem Natural er viável.

O recorte epistemológico dessa pesquisa em Processamento de Linguagem Natural aponta para o texto, material a ser processado para fins de recuperação de informação. Para isso torna-se necessário recorrer a campos de estudos específicos da Lingüística. O processamento automático das “informações lingüísticas” inerentes ao texto permitirá que o computador se torne um instrumento capaz de discernir os fenômenos da língua natural e, conseqüentemente, de criar uma taxonomia de textos de uma maneira mais “humana”, mais natural.

As informações lingüísticas mais pertinentes aos estudos de Processamento de Linguagem Natural são as morfológicas, as sintáticas, as semânticas e as pragmáticas. Tais tipos de informações são descritos a seguir.

2.4.1.1. Análise Morfológica

A análise morfológica (HAGEGÉ, 1997; SANTOS, 2001; PAULO et al. 2002) é aquela em que o texto é fragmentado para a determinação de seus componentes, as palavras e os sinais. As palavras são processadas de acordo com suas partes (raiz, afixos, prefixos e sufixos), e os sinais, como a pontuação, são separados da palavra, podendo ou não ser considerados relevantes. Para a frase apresentada a seguir pode-se fazer a análise morfológica:

1 O processo de disponibilização de um periódico eletrônico na World Wide Web é um empreendimento composto de várias etapas.

2 Separar as palavras e sinais:

O_1 $processo_2$ de_3 $disponibilização_4$ de_5 um_6 $periódico_7$ $eletrônico_8$ na_9 $World_{10}$ $Wide_{11}$ Web_{12} $é_{13}$ um_{14} $empreendimento_{15}$ $composto_{16}$ de_{17} $várias_{18}$ $etapas_{19}$.$_{20}$

3 Desmembrar as palavras:
disponibilizar + ação; disponível; dispor...

Esse processo de segmentação de palavras nas formas que as compõem às vezes é precedido de algum tipo de análise sintática com o intuito de otimizar as interpretações dos afixos (prefixos e sufixos), que podem depender da categoria sintática da palavra. A palavra “processo”, por exemplo, pode ser tanto um substantivo quanto o verbo processar conjugado na primeira pessoa do presente do indicativo. Alguns estudos de morfologia foram feitos com o intuito de desenvolver sistemas, como o Nptool (1992) por exemplo, para a extração automática de palavras representativas do texto, neste caso os sintagmas nominais, visando a geração de índice (VOUTILAINEN, 1992).

2.4.1.2. Análise Sintática

A análise sintática (CLARK & CLARK, 1977; CRAIN & STEEDMAN, 1985; CHOMSKY, 1986; 1995; BICK, 1996; SANTOS, 2001) é aquela em que cada termo da frase, e consequentemente do texto, recebe um nome que exprime a sua função dentro da estrutura oracional, função esta que é decorrente do seu relacionamento com um outro termo. Essa análise sintática necessita dos resultados da análise morfológica, para criar uma descrição estrutural da frase. O processo consiste em converter a lista de palavras que formam a frase em uma estrutura hierárquica, onde cada palavra tem o seu "valor" sintático (a categoria a qual ela pertence) explicitado. No caso da frase de exemplo, o resultado da análise é:

"O" – artigo masculino definido no singular.
"Processo" - substantivo masculino no singular.
"de" – preposição.
"disponibilização" - substantivo feminino no singular.
"de" – preposição.
"um" – artigo masculino indefinido no singular.
"periódico" – adjetivo masculino no singular.
"eletrônico" – adjetivo masculino no singular.
"em" – preposição.
"a" – artigo feminino no singular.
"World Wide Web" – nome próprio feminino no singular.
"é" – verbo ser ou estar na terceira pessoa do presente do indicativo no singular.
"um" – artigo masculino indefinido no singular.
"empreendimento" – substantivo masculino no singular.
"composto" – adjetivo.
"de" – preposição.
"várias" – pronome feminino no plural.
"etapas" – substantivo feminino no plural.

Esta análise foi feita manualmente. A análise apresentada nas figuras 2.2, 2.3, 2.4, 2.5, e 2.6 foi feita automaticamente pelo programa Palavras (BICK, 1996).

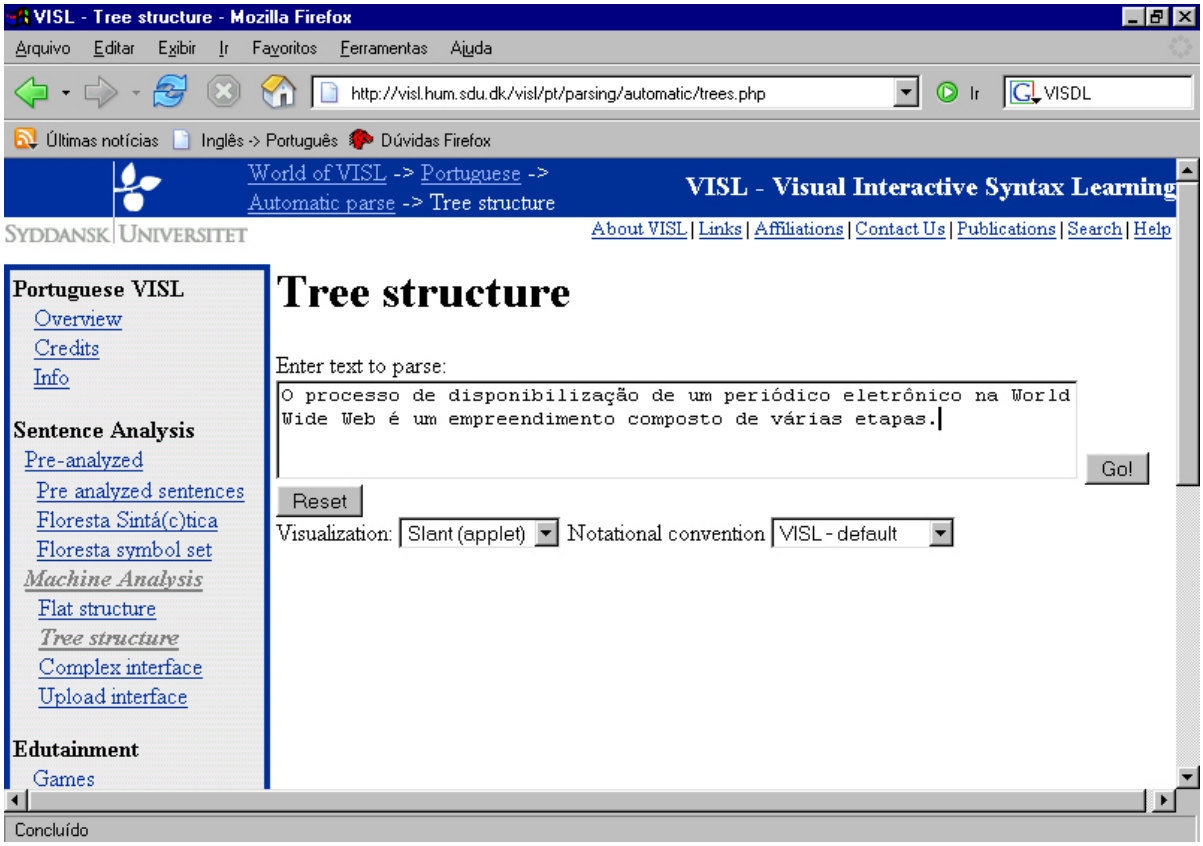


Figura 2.2. Tela principal do analisador sintático Palavras.

Na figura 2.2 apresenta-se a primeira tela do analisador sintático automático do português “Palavras”. Na tela pode-se ler a frase exemplo, “O processo...”, que foi digitada na caixa de diálogo do sistema para ser analisada automaticamente.

Todas as frases de todos os textos da coleção utilizada neste estudo foram analisadas automaticamente pelo programa Palavras.

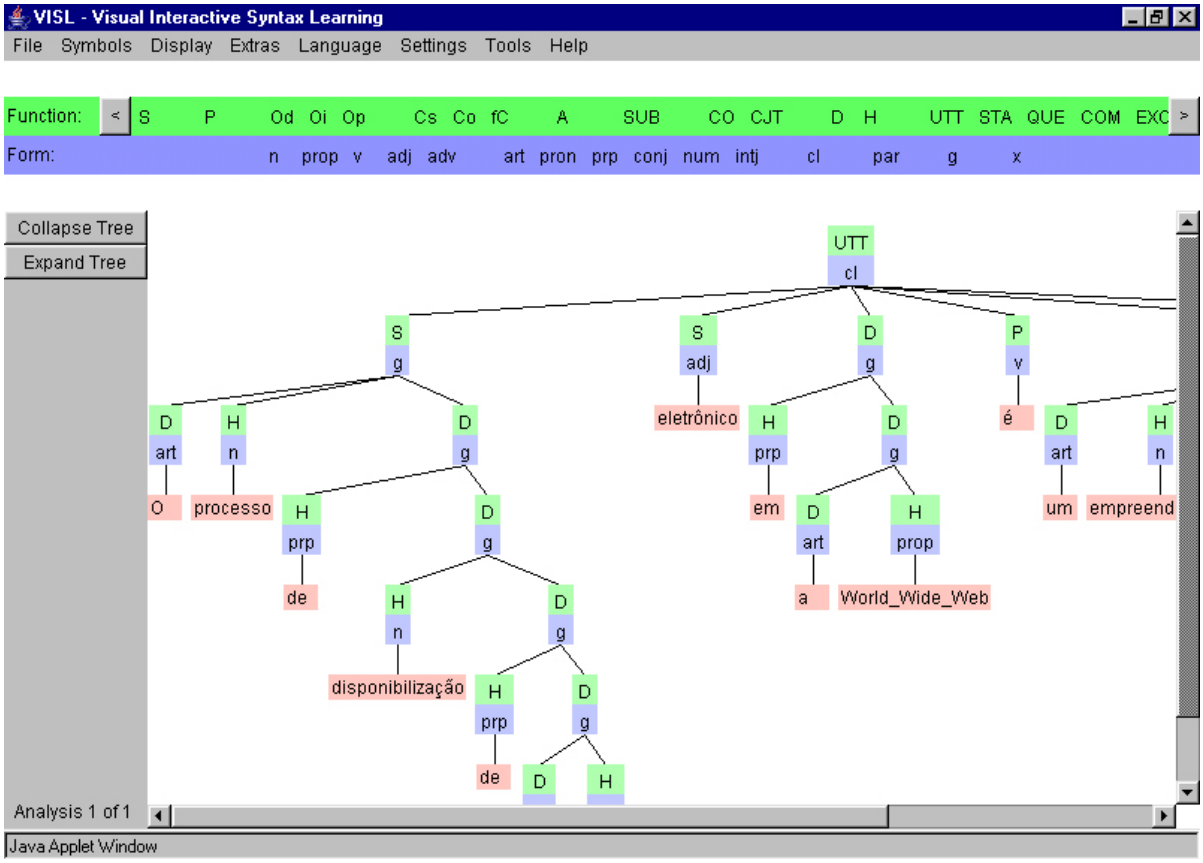


Figura 2.3. Primeira parte da apresentação em forma de árvore da análise sintática exemplo.

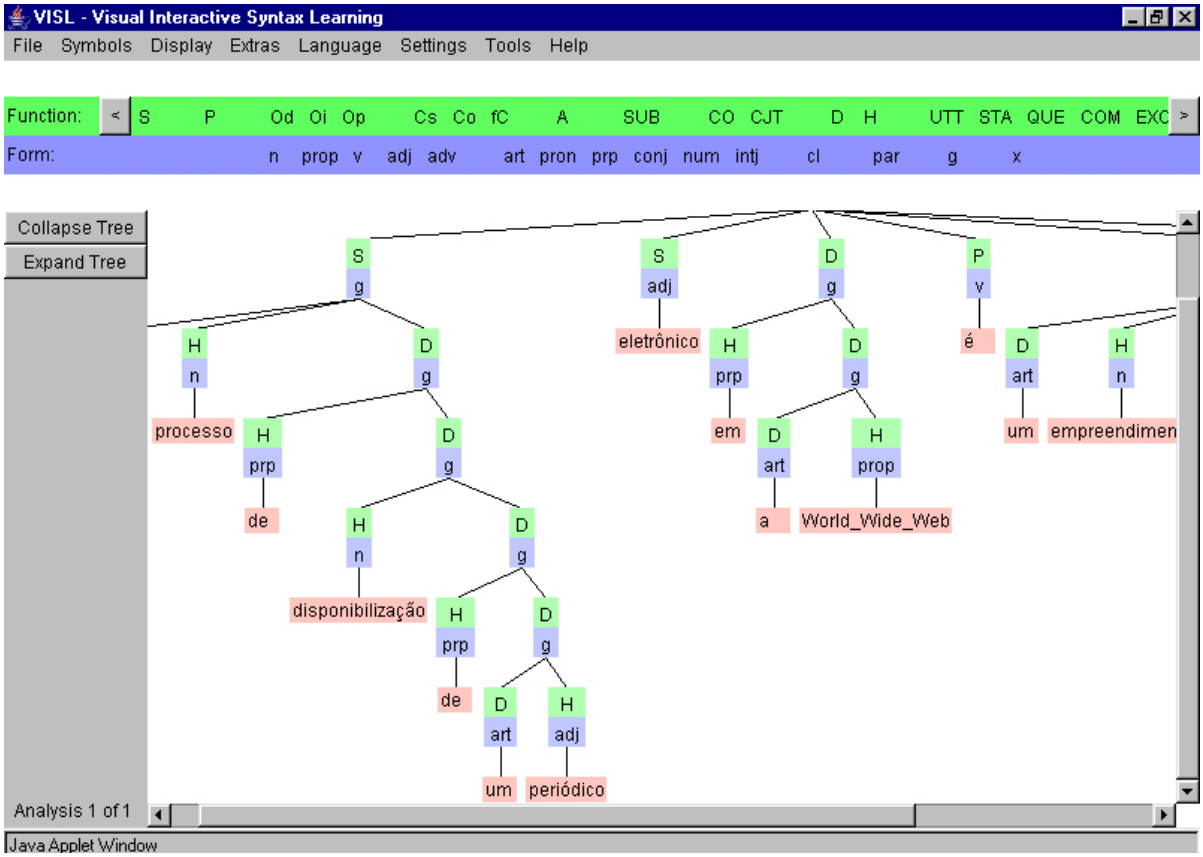


Figura 2.4. Segunda parte da apresentação em forma de árvore da análise sintática exemplo.

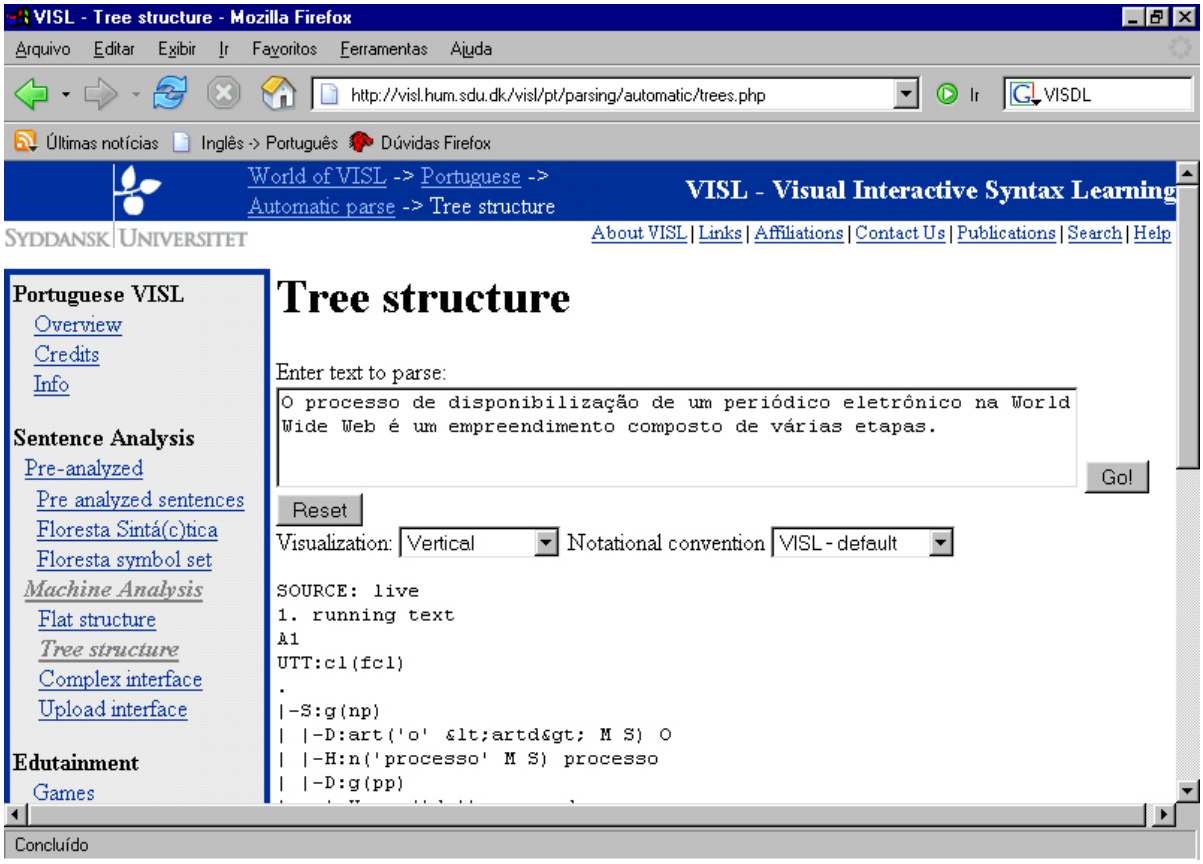


Figura 2.5 Apresentação da análise sintática de maneira vertical (parte I).

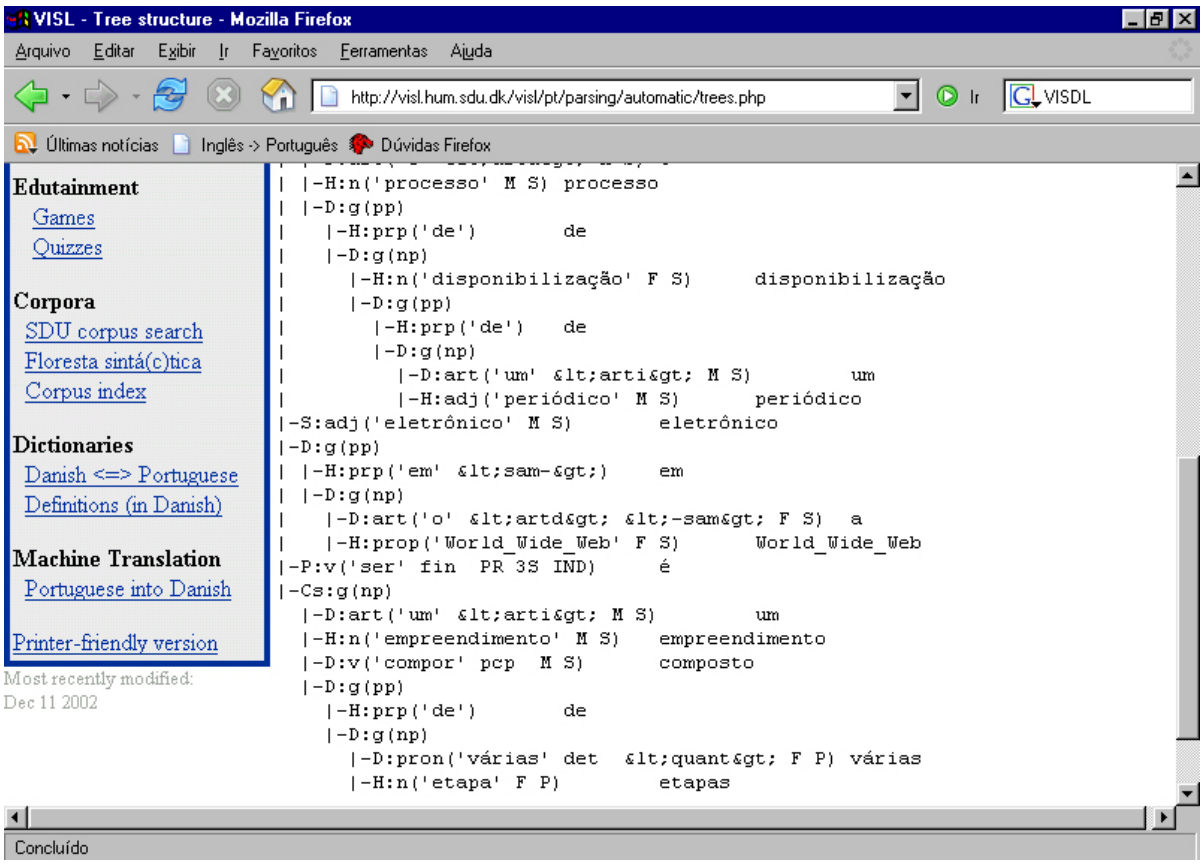


Figura 2.6 Apresentação da análise sintática de maneira vertical (parte II).

Nas figuras 2.3 e 2.4 apresenta-se o resultado da análise sintática da frase exemplo, “O processo...”, através de uma representação em forma de árvore, onde é possível distinguir graficamente os elementos que compõem sintaticamente o texto.

Nas figuras 2.5 e 2.6 apresenta-se o resultado da análise sintática da frase exemplo, “O processo...”, através de uma representação vertical, onde é possível distinguir através das palavras e símbolos os elementos que compõem sintaticamente o texto.

Há erros de análise no processo automático. “Composto”, no exemplo, é adjetivo (que se compôs; que se formou; constituído), mas a análise automática considerou “composto” como sendo particípio (forma nominal) do verbo “compor”.

```
SOURCE: live
1. running text
A1
UTT:cl(fcl)
.
|-S:g(np)
| |-D:art(&#039o&#039 &lt;artd> M S)    O
| |-H:n(&#039processo&#039 M S) processo
| |-D:g(pp)
|   |-H:prp(&#039de&#039)       de
|   |-D:g(np)
|     |-H:n(&#039disponibilização&#039 F S)     disponibilização
|     |-D:g(pp)
|       |-H:prp(&#039de&#039)   de
|       |-D:g(np)
|         |-D:art(&#039um&#039 &lt;arti> M S)   um
|         |-H:adj(&#039periódico&#039 M S)      periódico
|         |-D:adj(&#039eletrônico&#039 M S)     eletrônico
|         |-D:g(pp)
|           |-H:prp(&#039em&#039 &lt;sam->)     em
|           |-D:g(np)
|             |-D:art(&#039o&#039 &lt;-sam> <artd> F S) a
|             |-H:prop(&#039world_wide_web&#039 F S)    world_wide_web
|-P:v(&#039ser&#039 fin  PR 3S IND)     é
|-Cs:g(np)
  |-D:art(&#039um&#039 &lt;arti> M S)   um
  |-H:n(&#039empreendimento&#039 M S)   empreendimento
  |-D:v(&#039compor&#039 pcp  M S)      composto
  |-D:g(pp)
    |-H:prp(&#039de&#039)       de
    |-D:g(np)|
      |-D:pron(&#039várias&#039 det  &lt;quant> F P)    várias
      |-H:n(&#039etapa&#039 F P)        etapas
```

Figura 2.7 A análise sintática de maneira vertical em detalhes.

A figura 2.7 apresenta a estrutura hierárquica da análise sintática da frase “O processo de disponibilização de um periódico eletrônico na World Wide Web é um empreendimento composto de várias etapas” feita automaticamente pelo analisador sintático Palavras com suas respectivas etiquetas (BICK, 1996). Na figura 1 temos que “O” é um artigo (art); “processo” é um nome (substantivo) (n) e é uma palavra masculina que se apresenta no singular (M S);

“de” é uma preposição (prp); “disponibilização” é um nome (substantivo) (n) e é uma palavra feminina que se apresenta no singular (F S); “um” é um artigo (art) masculino e no singular (M S); “periódico” é um adjetivo (adj) masculino no singular (M S); “eletrônico” é um adjetivo (adj) masculino no singular (M S); “em” é uma preposição (prp); “a” é artigo (art) feminino no singular (F S); World Wide Web é um nome próprio (prop) feminino no singular (F S); “é” é o verbo ser no presente (PR) na terceira pessoa do singular (3S) no indicativo (IND); “empreendimento” é um nome (n) masculino no singular (M S); “composto” é um verbo (v) no particípio (pcp) masculino no singular (M S); “várias” é um pronome (pron) feminino no plural (F P); “etapas” é um nome (n) feminino no plural (F P).

2.4.1.3. Análise Semântica

A análise semântica (FILLMORE, 1968; FREDERIKSEN, 1975; 1986; JACKENDOFF, 1990, 1994; GERNSBACHER, 1994) permite a identificação do significado de cada termo (palavra) da frase, isolada e conjuntamente com outros termos. Permite a identificação dos conceitos primitivos do texto, aqueles que mantém a essência do texto. O significado é inerente ao termo e é parte integrante do texto como um todo, ou seja, o significado de “processo” isoladamente nos sugere um verbo ou um substantivo, porém, conforme o exemplo anterior, “O processo de disponibilização..” significa que “processo”, inserido nessa frase, é um substantivo. Para a interpretação semântica, sob o ponto de vista cognitivo, é necessário que o leitor construa proposições.

Proposições são enunciados, para a lógica tradicional de matriz aristotélica. Uma proposição é uma expressão lingüística de uma operação mental (o juízo), composta de sujeito, verbo (sempre redutível ao verbo ser) e atributo, e passível de ser verdadeira ou falsa. As proposições são representações semanticamente completas. Integram as suposições que se localizam nas memórias de longo prazo. Neste estudo, uma proposição é uma unidade constituída de sentido e que é maior que o significado de uma palavra e menor que uma narrativa ou uma teoria. São como unidades básicas do significado, podendo ser uma mera palavra ou um texto inteiro. No exemplo citado temos “processo”, que por si só é uma proposição e “O processo de disponibilização de um periódico eletrônico na World Wide Web é um empreendimento composto de várias etapas”, que é uma proposição mais complexa.

A elaboração das proposições ocorre por meio das estruturas sintáticas do texto. Para a psicologia cognitiva (KINTSCH & VAN DIJK, 1978; FREDERIKSEN *et al.*, 1990), as proposições são constituídas de um predicador e um ou mais argumentos e servem tanto para representação da informação conceitual do texto, quanto como unidades de informação para o raciocínio lógico e a solução de problemas (FREDERIKSEN, 1975). No caso de nosso estudo, utilizamos uma metalinguagem (FREDERIKSEN, 1986) para a identificação das proposições e seus constituintes. Baseando-se nessa abordagem, as proposições são compostas por núcleo ou predicador, por argumentos ou participantes e por atributos da predicação (mais detalhes veja GOTTSCHALG-DUQUE, 1998).

O relevante para este estudo são os sintagmas nominais, que podem ser “Agentes”, aqueles que causam o(s) evento(s); “Objetos afetados”, aqueles que são afetados pela ação; e “Instrumentos”, o que se usa para executar a ação. Na figura 2.8 temos a representação da análise da frase “O processo de disponibilização de um periódico eletrônico na World Wide Web é um empreendimento composto de várias etapas”:

PARA·ESTA·PRIMEIRA·FRASE·TEMOS·OS·SEGUINTES·SNs·EXTRAÍDOS·AUTOMATICAMENTE	PARA·ESTA·PRIMEIRA·FRASE·TEMOS·OS·SEGUINTES·VERBOS·EXTRAÍDOS·AUTOMATICAMENTE
O·processo·de·disponibilização·de·um·periódico·eletrônico·em·a·World·Wide·Web	É·(verbo·SER·ou·ESTAR·sempre·ligado·a·ESTADO)
Um·empreendimento	composto
Várias·etapas¶	

Figura 2.8 Representação da Análise semântica.

A figura 2.8 apresenta, sob a forma de tabela, o resultado da extração automática das proposições e de seus constituintes através da utilização da metalinguagem de Frederiksen (1975). Os verbos e os sintagmas nominais são produtos da análise sintática. A análise semântica automática é baseada nas etiquetas sintáticas

2.4.1.4. Análise Pragmática

A análise pragmática (KINTSCH & van DICK, 1993; DRESNER & DASCAL, 2001) refere-se ao processamento daquilo que foi dito ou escrito em contraste com o que realmente se quis dizer ou escrever. Muitos estudiosos consideram tais análises como sendo extralingüísticas, que não pertencem ao domínio da Lingüística e sim da Psicologia, da Filosofia e da Antropologia. Nos estudos de Processamento de Linguagem Natural a Pragmática está associada à Análise do Discurso (SINGER, 1994), que também é entendida como ambiente extralingüístico. Para realizar-se tal análise é necessário considerar vários fatores subjetivos, tais como o contexto, as condições de produção do discurso e a formação discursiva. Tais questões, embora sejam atualmente relevantes para o desenvolvimento do Processamento de Linguagem Natural, não são contempladas por este estudo.

2.5. Ontologia

Ontologia é um ramo da filosofia que estuda o ser e tudo que se relaciona ao ser (HEIDEGGER, 1925). Neste estudo ontologia é restrita à ótica da Inteligência Artificial. É apenas uma especificação formal de uma conceitualização compartilhada, que é uma visão abstrata e simplificada do universo que se pretende representar (GRUBER, 1993). A ontologia fornece um vocabulário comum de uma área e define, com diferentes níveis de formalismo, o significado dos termos e dos relacionamentos entre os mesmos (GOMEZ-PÉREZ & BENJAMINS, 1999). Ontologias podem ser muito úteis para um Sistema de Recuperação de Informação porque são estruturadas de tal modo (classes instâncias relações), que permitem ir consideravelmente além das possibilidades oferecidas por outros sistemas de classificação (GERDA, 1997, BREWSTER, 2002) como Tesaurus, por exemplo, que é um vocabulário de um ramo do saber que descreve, sem ambigüidade, os conceitos a ele atinentes (FRAKES & BAEZA-YATES, 1992).

As Ontologias, portanto, são conjuntos de asserções, afirmações categóricas, que definem as relações entre conceitos e estabelecem regras lógicas de raciocínio sobre eles. Elas permitirão que sistemas operacionais sejam capazes de processar o significado das informações. Esse avanço fará com que as máquinas sejam capazes de se comunicar com

outras máquinas, e também com seres humanos, de uma maneira mais “natural”, ou seja, através da utilização de representações semânticas, conceitos e seus atributos (LÈVY, 2000). As ontologias já estão sendo utilizadas na Recuperação de Informação. Entretanto normalmente essas ontologias são criadas previamente e manualmente e são utilizadas para expandirem as consultas dos usuários.

2.5.1. Conceitos Básicos

Embora não exista muito sobre a definição de Ontologia (mais detalhes no capítulo 3), as ontologias aplicadas, conceito mais adotado e divulgado pelos pesquisadores da área da Ciência da Computação conhecida como Inteligência Artificial (I.A.), apresentam várias características comuns.

Uma Ontologia (strictu sensu) é composta de classes, relações, regras e instâncias (CORAZZON, 2003). Uma Ontologia é um “catálogo de tipos de coisas”, às quais assume-se existir em um domínio de interesse (SOWA, 1999). Para BORST (1997), uma Ontologia é uma especificação formal e explicita de uma conceitualização compartilhada. As diferenças encontradas entre as abordagens distintas residem principalmente na estrutura, função e aplicação. Entretanto, as ontologias existentes nas pesquisas desenvolvidas pela I.A. apresentam mais afinidades do que discrepâncias, permitindo até o intercâmbio dos dados gerados e editados por um modelo em outro modelo distinto. Para este estudo, um fator preponderante para a utilização de Ontologias na geração de índice da coleção de documentos é o fato de que as mesmas são estruturadas de maneira a permitir um considerável ganho de qualidade, quando empregadas em um sistema de classificação. Elas oferecem maiores possibilidades estruturais (classes; instâncias; parte-todo; pai-filho; etc.) para a classificação de documentos do que as que são oferecidas por outros sistemas, como por exemplo, Thesauri (RUGE, 1997).

2.5.2. Utilização de Ontologias

Atualmente, as ontologias são utilizadas de maneiras variadas e para vários fins (GUARINO, 1997; DING & FOO, 2001). Para a aplicação na Recuperação Automática de Informação a utilização de “Ontologias Leves” parece ser uma opção mais prática, pois, a princípio, elas podem ser automatizadas de modo mais simples.

As “Ontologias Leves” (DIN & ENGELS, 2001) são ontologias simples, “incompletas”, pois são compostas apenas de classes e instâncias, não contendo funções (relações especiais entre as classes) ou outros tipos de primitivas de representação. As principais características de “Ontologias Leves” são:

- Apresentam uma estrutura de árvore rasa.
- Podem ser extraídas diretamente das linguagens naturais.
- Podem ser geradas semi-automaticamente a partir de documentos de um dado domínio.
- Base teórica e metodológica advinda do Processamento de Linguagem Natural (PLN), Aprendizado Automático (*Machine Learning*), Extração de Informação (EI) e da Recuperação da Informação (IR).
- Contêm muito ruído, a ambigüidade das palavras é de difícil tratamento.
- O refinamento das mesmas, independentemente de terem sido obtidas automática ou semi-automaticamente, não é trivial e requer abordagens heurísticas.
- A identificação e aprendizado automático das relações existentes entre seus elementos ainda é um problema de tratamento complexo.
- São usadas para uma determinada tarefa ou para um domínio bem especifico.

O Sub-Módulo de Estrutura de Índice (SMEI) é o índice da coleção propriamente dito. É uma “ontologia leve”, pois todos os conceitos dessa coleção encontram-se neste sub-módulo. É uma lista invertida de **proposições**. Dada uma **proposição** tem-se os textos que a contém. É interessante salientar que, como os usuários do SRI normalmente não fazem consultas utilizando-se de **proposições**, os termos das consultas são expandidos para as proposições que os contém.

4 Resultados e Discussões

A seguir, apresentamos os experimentos e seus resultados. O Experimento-Piloto e o Experimento de Validação, realizados com as coleções compostas dos documentos extraídos da Revista Ciência da Informação. Os resultados apresentados neste capítulo são discutidos no capítulo 5.

4.1. Descrição do Modelo de SRI

Para a criação e utilização dos módulos do protótipo para o Experimento-Piloto empregamos programas de computador que já desenvolvidos e disponibilizados para o uso. O programa Palavras (Bick, 1996) e o programa Protégé (Stanford Medical Informatics, 2005). Além disso, foi criado e desenvolvido um software específico para o analisador semântico, chamado de GeraOnto (Registro de Software INPI nº 00065066, 2004).

O módulo de processamento de linguagem natural (MPLN) é constituído pelo analisador sintático (syntactic parser) do português chamado Palavras (BICK, 1996), que usa regras gramaticais formuladas com base na Constraint Grammar Formalism (GCF) (BICK, 2000; 2000; AFONSO et al., 2002) e pelo analisador semântico GeraOnto. Optamos pelo Palavras, uma vez que o mesmo é considerado um dos melhores analisadores sintáticos para o português e o acesso ao mesmo (via Web, FTP) é gratuito. O analisador semântico (SMOSe) (figura 3.4) foi desenvolvido e implementado especificamente para este projeto e denominado GeraOnto, que gera uma “ontologia leve”. No protótipo desenvolvido, ele encontra-se no

Protégé (2005).

O Protégé é um editor de ontologias desenvolvido na Universidade de Stanford. É um software gratuito, desenvolvido em Java. É *open source*, ou seja, seu código fonte é disponibilizado para eventuais modificações por parte dos usuários. Existe atualmente uma comunidade de mais de 3500 colaboradores que desenvolvem o Protégé. Ele atualmente já dispõe de vários recursos opcionais, tais como: funcionamento em rede, adição de visualizadores gráficos, etc. Para o Módulo Gerador de Ontologia (MGO) empregou-se o editor de ontologias Protégé. O Módulo Gerador de Índice (MGI) foi simulado manualmente, através da verificação da ocorrência do termo da consulta em alguma proposição e identificação dos textos que continham tal proposição. O índice gerado foi armazenado no Protégé.

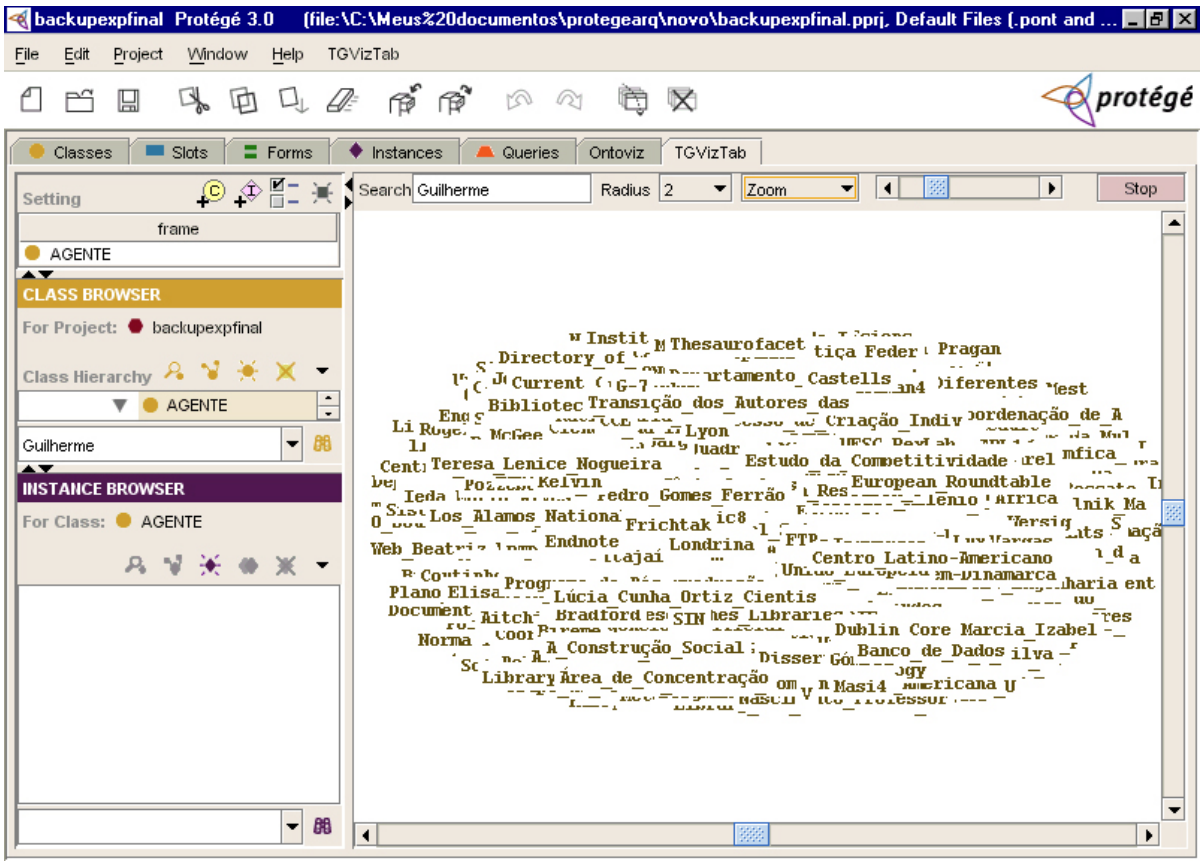


Figura 4.1. Tela do Protégé para "AGENTE".

Na figura 4.1 vê-se a tela do programa Protégé, com a representação gráfica dos

elementos pertencentes à classe AGENTE.

A seguir, é apresentada uma análise passo a passo de cada módulo.

O MPLN

A utilização do MPLN (Módulo de Processamento de Linguagem Natural) em um SRI visa otimizar o processo de indexação, identificando conceitos estruturados encontrados nos textos. Assim, os textos são indexados em função dos conceitos, tal como foi apresentado na descrição do SMRI. Portanto, esse módulo analisa as frases nos documentos objetivando a identificação de conceitos.

O SMA

A atomização do texto. O texto é dividido em partes. O autor, o título e as palavras-chave são enviados para o SMOF. As frases que compõem o texto são enviadas para o SMOSi e processadas sintaticamente. No Experimento-Piloto, esse processamento foi simulado manualmente para comparação com a análise processada pelo “Palavras”.

Como representado na figura 4.2 o autor, o título do artigo e as palavras-chave são

enviados para o SMOF. O texto, com as sentenças discriminadas, etiquetadas, é enviado para o SMOSi.

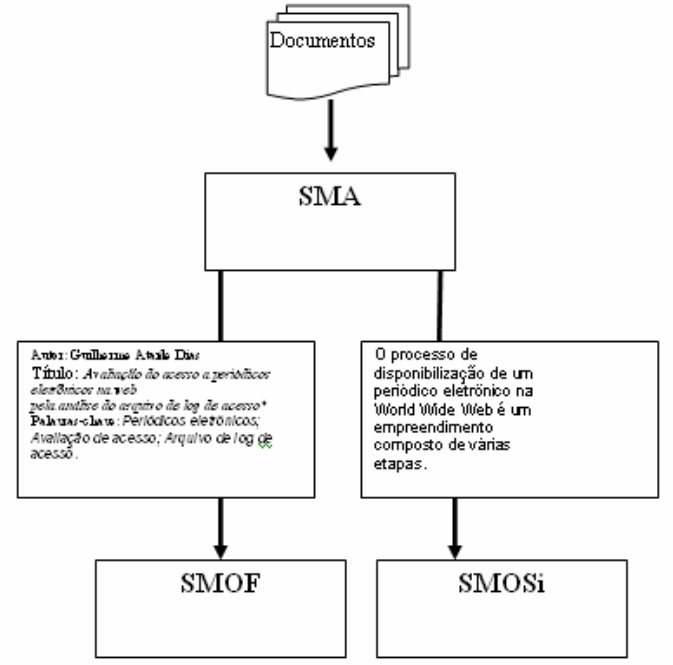


Figura 4.2. SMA e sua saída.

O SMOSi

O SMOSi processa sintaticamente cada frase do texto. Após a etiquetagem sintática, o produto do SMOSi é enviado para o SMOSe no qual se origina a análise semântica. No Experimento-Piloto esta análise sintática foi realizada pelo software PALAVRAS (BICK, 1996).

Na figura 4.3 temos uma frase analisada sintaticamente e devidamente etiquetada.

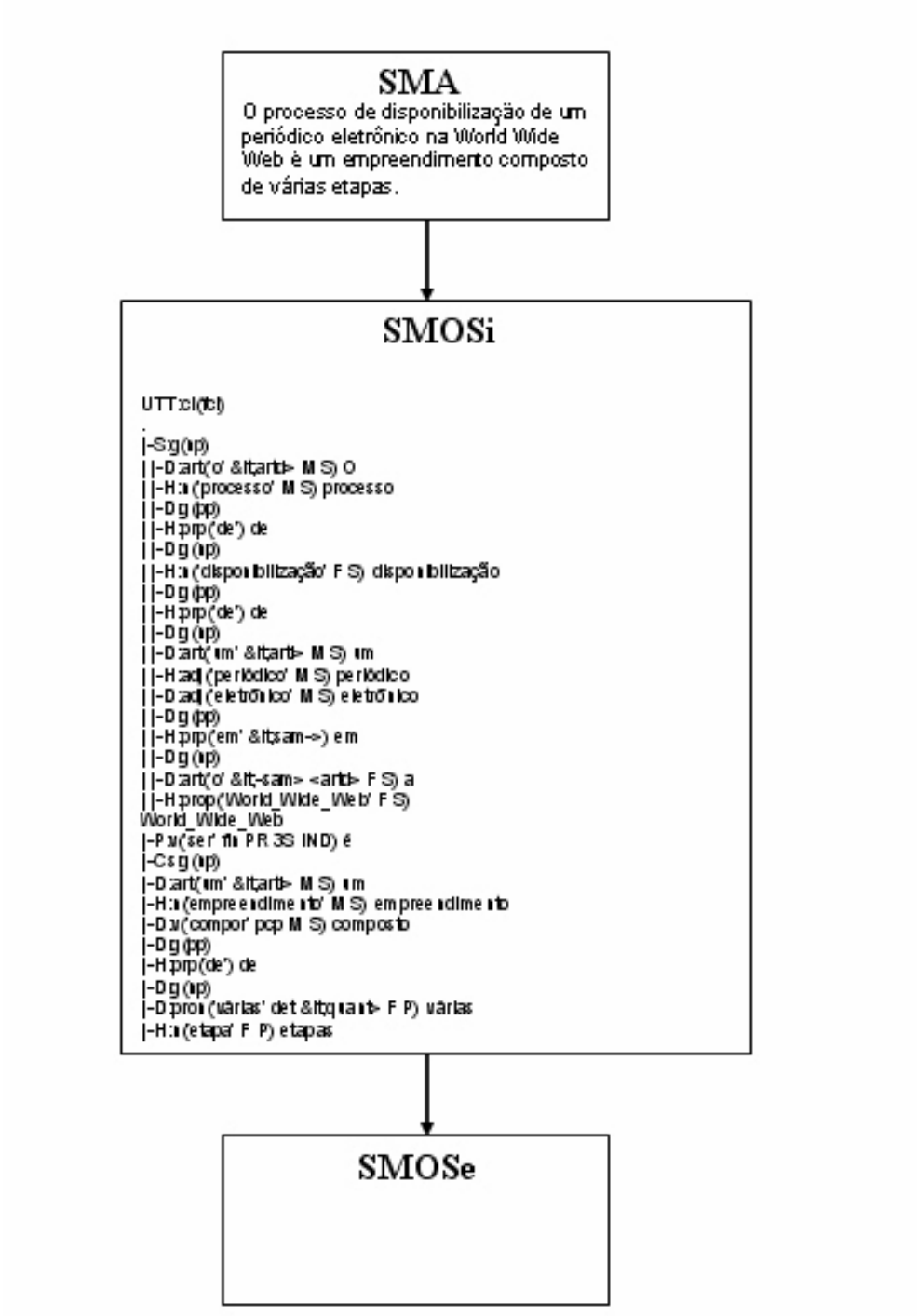


Figura 4.3. SMOSi e sua saída.

O SMOSe

O SMOSe procede à análise semântica de cada frase do texto já processada sintaticamente. Após a etiquetagem sintática e, de acordo com esta, os elementos semânticos são identificados e discriminados. Neste estágio, a identificação automática de todos os elementos semânticos ainda é incipiente.

Como o objetivo principal deste estudo é a otimização do processo de recuperação de informação em uma dada coleção, a identificação do núcleo proposicional e dos termos do texto que preenchem o espaço ocupado por agentes, objetos e instrumentos é totalmente passível de ser automatizada. Neste experimento, essa identificação foi feita manualmente, em substituição ao processo automático.

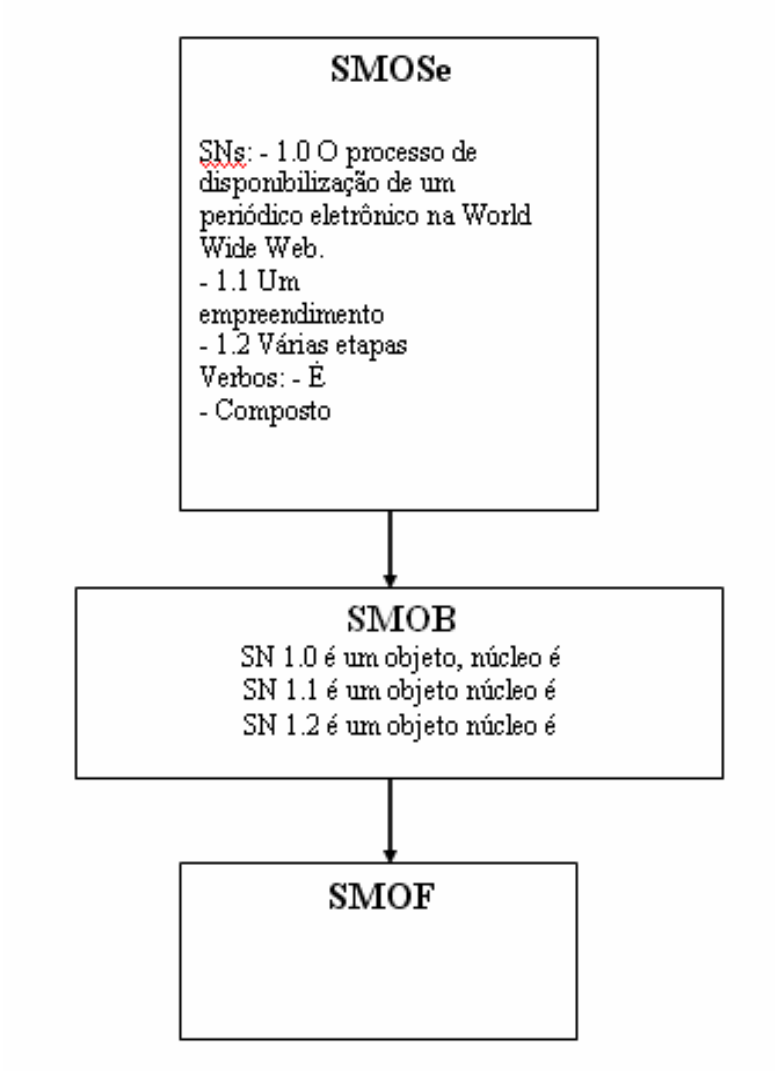


Figura 4.4. SMOSe e sua saída.

Na figura 4.4 temos a frase analisada semanticamente, devidamente etiquetada e enviada para o SMOB para a verificação.

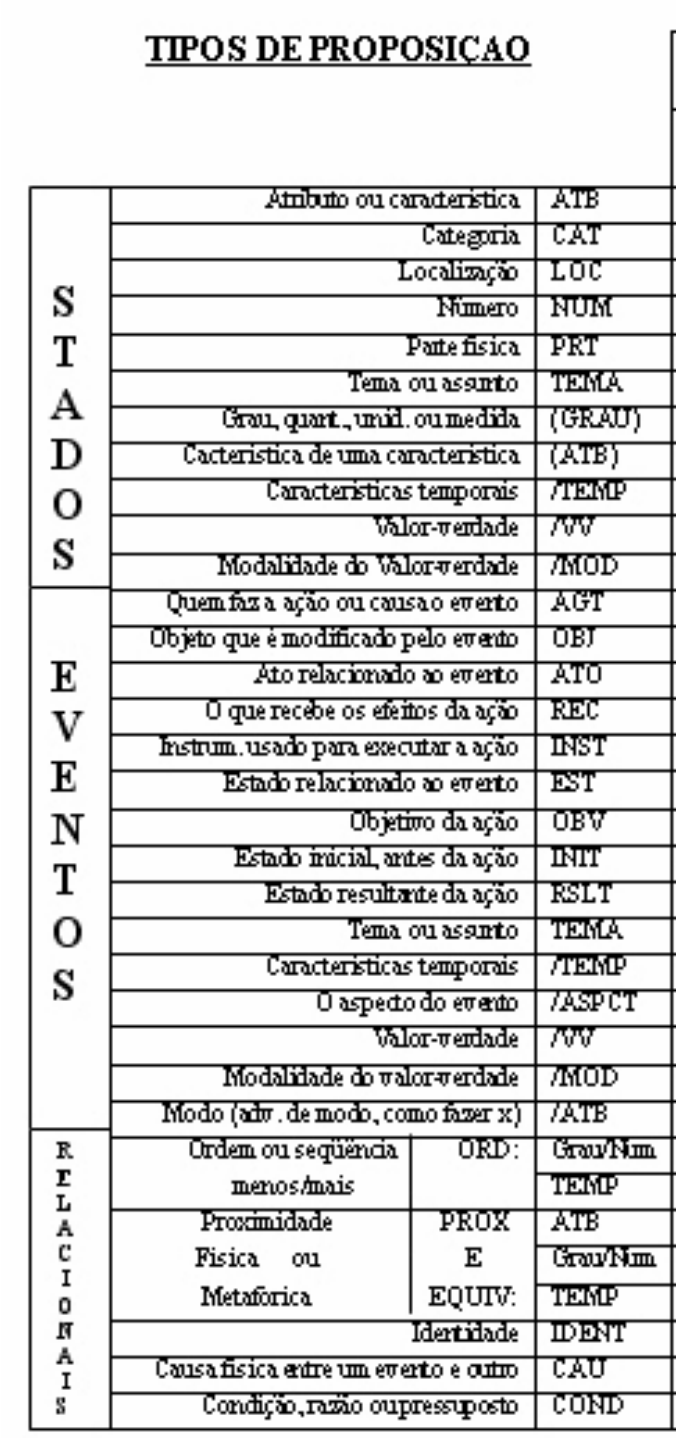
O MGO

O Módulo Gerador de Ontologia, no Experimento-Piloto, é o editor de ontologias Protégé. A ontologia básica, assim como a ontologia gerada, encontra-se nesse módulo. Os conceitos extraídos dos textos da coleção tornam-se então as classes da ontologia gerada pela coleção.

O SMOB

O Sub-Módulo de Ontologia Básica é uma ontologia criada e armazenada no Protégé. Essa ontologia, fundamentada na análise proposicional de Frederiksen (1975). É o padrão referencial para a conversão automática das etiquetas sintáticas em etiquetas semânticas. Baseado em suas classes, é possível identificar, por meio das relações sintáticas entre os termos, as possíveis relações semânticas.

Na figura 4.5 apresentam-se os tipos de proposições utilizados na Ontologia Básica.



TIPOS DE PROPOSICAO

ESTADOS	Atributo ou caracteristica		ATB
	Categoria		CAT
	Localização		LOC
	Número		NUM
	Parte fisica		PRT
	Tema ou assunto		TEMA
	Grau, quant., unid. ou medida		(GRAU)
	Cacteristica de uma caracteristica		(ATB)
	Caracteristicas temporais		/TEMP
	Valor-verdade		/VV
	Modalidade do Valor-verdade		/MOD
EVENTOS	Quem faz a ação ou causa o evento		AGT
	Objeto que é modificado pelo evento		OBJ
	Ato relacionado ao evento		ATO
	O que recebe os efeitos da ação		REC
	Instrum. usado para executar a ação		INST
	Estado relacionado ao evento		EST
	Objetivo da ação		OBV
	Estado inicial, antes da ação		INIT
	Estado resultante da ação		RSLT
	Tema ou assunto		TEMA
	Caracteristicas temporais		/TEMP
	O aspecto do evento		/ASPCT
	Valor-verdade		/VV
	Modalidade do valor-verdade		/MOD
	Modo (adv. de modo, como fazer x)		/ATB
RELACIONAIS	Ordem ou seqüência menos/mais	ORD:	Grau/Num
			TEMP
	Proximidade Fisica ou Metafórica	PROX	ATB
		E	Grau/Num
		EQUIV:	TEMP
	Identidade		IDENT
	Causa fisica entre um evento e outro		CAU
	Condição, razão oupressuposto		COND

Figura 4.5. SMOB, tipos de proposições possíveis (GOTTSCHALG-DUQUE 1998).

O SMOF

O Sub-Módulo de Ontologia Formada é uma ontologia leve, criada automaticamente a partir dos conceitos encontrados nos textos da coleção e armazenada no Protégé. Essa ontologia, obtida através dos conceitos extraídos da análise proposicional dos textos da coleção, serve de base para a geração do índice da coleção. De acordo com as suas classes, é possível identificar quais os conceitos relevantes para a coleção e em que textos eles se encontram. Como exemplo, a seguir a classe "**AGENTE",** composta por todos os autores dos artigos da coleção, pois autor é um agente (quem / o que causou um evento).

Na figura 4.6 apresenta-se o SMOF com a classe **"AGENTE"** e suas subclasses.

Figura 4.6. SMOF, agentes, mais precisamente autores da coleção do Experimento-Piloto.

O MGI

O Módulo Gerador de Índice, no Experimento-Piloto, é o editor de ontologias Protégé. O Sub-Módulo de Regras de Índice foi simulado manualmente (como detalhado no capítulo 5 e nos Anexos). Basicamente, o SMEI é uma lista invertida de conceitos. Para cada conceito há uma lista com os textos nos quais os mesmos conceitos aparecem. Observe-se que, para um conceito, pode existir mais de um termo.

Na figura 4.7 apresenta-se um exemplo de lista invertida de conceitos para indexação dos textos da coleção.

Conceito	**Textos**
Guilherme Ataíde Dias	Texto 1
O processo de disponibilização de um periódico eletrônico na World Wide Web	Texto 1
Ilza Leite Lopes	Texto 4, Texto 10

Figura 4.7. SMEI, AGENTES, exemplo da lista invertida de indexação dos textos através dos conceitos.

4.2. Resultados do Experimento-Piloto

Realizou-se um experimento, o Experimento-Piloto, objetivando o refinamento metodológico do material a ser utilizado no Experimento de Validação e a conseqüente verificação de coerências na utilização de teorias de Lingüística Computacional e Ontologia em um Sistema de Recuperação de Informação. Observou-se uma otimização da resposta do mesmo, e, especificamente, que a análise semântica automática dos textos permite indexar por conceitos. Nesse Experimento-Piloto foram utilizados 41 artigos, produzidos em língua portuguesa publicados na Revista Ciência da Informação. Os artigos foram extraídos da revista 31, números 1, 2 e 3 (2002) e da revista 32, número 1 (2003). Para a realização do Experimento-Piloto e do Experimento de Validação, utilizou-se apenas o título, autor, as

palavras-chave e a introdução dos referidos artigos. Isto não deprecia os experimentos, visto que, geralmente nos documentos científicos, as palavras mais relevantes do texto encontram-se na introdução (LAROCCA NETO et al., 2000; PEREIRA et al., 2002). O Experimento-Piloto e o experimento de validação contaram com 37 sujeitos, que realizaram, as tarefas descritas a seguir:

1. A primeira tarefa consistiu na leitura dos textos por 10 sujeitos e na atribuição de 10 palavras-chave por texto para indexação deles. As palavras-chave foram atribuídas aos mesmos pelos sujeitos e independentemente delas existirem ou não nos mesmos. Esta tarefa teve como objetivos:

   Verificar se os sujeitos realmente leram os textos.

   Contrastar as palavras-chave do(s) autor(es) dos artigos com as palavras-chave dos sujeitos e com as palavras-chave geradas automaticamente pelo SiRILiCO.

   Identificar as possíveis relações semânticas não encontradas automaticamente.

2. A segunda tarefa consistiu em formular consultas à mesma coleção e avaliar a relevância das respostas obtidas num SRI com modelo vetorial e no protótipo (SiRILiCO) (mais detalhes nos capítulos 4 e 5).

Os participantes do experimento são alunos de graduação do Curso de Ciência da Informação (CI) da Pontifícia Universidade Católica de Minas Gerais PUC-Minas, sendo 16 deles alunos do sétimo período e 21 alunos do sexto período. Os fatores que levaram a essa escolha foram: a facilidade para utilização de laboratório de informática e a disponibilidade dos alunos para a realização da tarefa. A média de idade dos sujeitos foi de 22 anos.

Estes colaboradores realizaram as tarefas individualmente, utilizando computadores contendo a coleção selecionada, um SRI com o modelo vetorial e um simulador (protótipo SiRILiCO) do SRI proposto. As tarefas foram realizadas em dias distintos. Cada sujeito foi informado, verbalmente e por escrito, sobre os procedimentos a serem cumpridos. Após o

informe o sujeito era questionado a respeito de possíveis dúvidas com relação à suas obrigações. Sanadas as possíveis dúvidas realizou-se o trabalho:

1. Leitura de cada um dos 10 textos designados e elaboração de 10 conceitos (palavras-chave) por texto. Esses conceitos se prestariam para a indexação dos respectivos textos.

2. Realização de consultas à coleção, utilizando palavras-chave fornecidas para o experimento, e avaliação dos cinco primeiros documentos das respostas dos dois sistemas, de acordo com a relevância do documento para a consulta. Foram utilizados os níveis: Muito Relevante, Relevante, Satisfatório, Irrelevante e Muito Irrelevante.

A avaliação da eficácia de um SRI requer o uso de metodologias objetivas e, principalmente, confiáveis (ELLIS, 1996; SU, 1998) para constatar a relevância (SPERBER & WILSON, 1995) da resposta dos sistemas para uma determinada tarefa e, como estes podem ser otimizados. A avaliação pode se dar sob dois pontos de vista, a princípio não biunívocos: o ponto de vista do sistema e o ponto de vista do usuário (BENNETT et al., 1999, WU & SONNENWALD, 1999; PRATT & FAGAN, 2000). A avaliação de Recuperação de Informação Textual ainda é controversa (MEADOW, 1992; ELLIS, 1996; GREISDORF, 2000) e a realização de experimentos como este, utilizando sujeitos nativos falantes de Português do Brasil e coleções em Português do Brasil, contribuem para a definição e validação de critérios avaliativos adotados na Recuperação de Informação de documentos nesse idioma (SANTOS, 2004).

Utilizamos como métricas avaliativas a precisão e a revocação, sendo que:

- **Precisão** (*precision*): avalia se os documentos recuperados da coleção são todos relevantes.
- **Revocação** (*recall*): avalia se todos os documentos relevantes da coleção foram recuperados pelo sistema.

Temos que:

- **N**: é o conjunto de documentos da coleção que são relevantes para determinadas consultas, sendo que esta relevância é determinada por especialistas.
- **R**: é o conjunto de documentos da coleção recuperados pelo sistema.

A precisão e dada por:

$$\frac{|\ N \cap R\ |}{|\ R\ |}$$

A revocação é dada por:

$$\frac{|\ N \cap R\ |}{|\ N\ |}$$

Foram objetivos da realização desse Experimento-Piloto:

1 Avaliar os materiais e a análise a ser utilizada no experimento;

2 Refinar a abordagem teórica necessária para a efetiva indexação dos documentos da coleção.

Nesta seção apresentamos os resultados produzidos no Experimento-Piloto. Esse Experimento-Piloto contou com 48 artigos extraídos da Revista Ciência da Informação referentes aos números 1, 2 e 3 do Volume 31 (2002) e o número 1 do Volume 32 (2003). A coleção foi composta de 22.633 palavras sendo a mais freqüente o termo “da”. O primeiro termo com valor semântico, “Informação”, apareceu na vigésima quinta posição, o segundo termo com valor semântico, “Ciência” apareceu logo a seguir (vigésima sexta posição). Foram identificados e extraídos automaticamente 1307 Sintagmas Nominais, a partir dos quais, foram geradas 1307 Classes da Ontologia.

Consulta	Vetorial Precisão	Vetorial Revocação	**SiRILiCO Precisão**	**SiRILiCO Revocação**
1	0	0	**0**	**0**
2	0,66	1	**1**	**1**
3	0,88	1	**1**	**1**
4	0,14	0,8	**0,6**	**1**
5	0,14	1	**0,2**	**1**
6	1	1	**1**	**1**
7	0	0	**0**	**0**
8	1	1	**1**	**1**
9	0	0	**0**	**0**
10	1	1	**1**	**1**
11	1	1	**1**	**1**
12	1	0,6	**1**	**0,85**
13	1	0,13	**1**	**1**
14	0	0	**0**	**0**
15	0	0	**0**	**0**
16	0	0	**0**	**0**
17	0,63	1	**0,71**	**1**
18	0	0	**0**	**0**
19	0,75	0,69	**0,75**	**0,75**
20	1	1	**1**	**1**
21	1	0,14	**1**	**0,29**
22	0,42	0,55	**1**	**0,89**
23	1	0,57	**1**	**0,86**
24	0	0	**0**	**0**
25	1	0,33	**1**	**0,33**

Tabela 4.1 Precisão e Revocação do Modelo Vetorial e do Modelo SiRILiCO para as consultas realizadas no Experimento-Piloto.

Na tabela 4.1 apresentou-se os resultados obtidos para as consultas utilizando o Modelo Vetorial e o Modelo *SiRILiCO*. A utilização de Sintagmas Nominais para a geração de Classes em uma Ontologia, que servem como indexadores, otimiza o desempenho do Sistema de Recuperação de Informação em relação ao sistema comum, que gera apenas uma lista invertida baseada apenas em termos. Como a coleção é pequena, alguns termos das consultas apresentam apenas um documento relevante. Isso justifica, em parte, a presença dos valores “1” e “0” encontrados nas respostas.

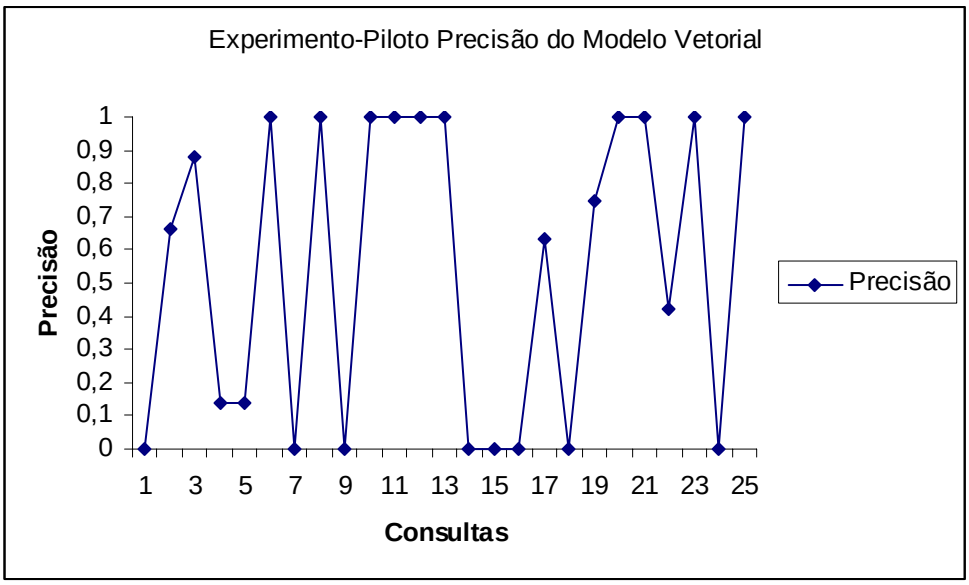


Figura 4.8.: Resultado de Precisão do Experimento-Piloto utilizando o Modelo Vetorial

Na Figura 4.8 tem-se os resultados de Precisão obtidos nas consultas. Como citado anteriormente, para a Recuperação da Informação, Precisão é o número total de documentos relevantes retornados dividido pelo numero total de documentos retornados. O Sistema de Recuperação de Informação utilizado nestas consultas é o sistema chamado “tradicional”, pois utiliza-se do modelo vetorial para indexar os documentos a partir dos termos que compõem os textos. Este Sistema de Recuperação de Informação, doravante chamado de Modelo Vetorial, não despreza as “stop-words” (termos muito freqüentes nos textos e considerados vazios de conteúdo semântico, sem valor para indexação, como artigos por exemplo). O Modelo Vetorial apresentou precisão média de 0,55 para as consultas realizadas no Experimento-Piloto (veja tabela 4.1).

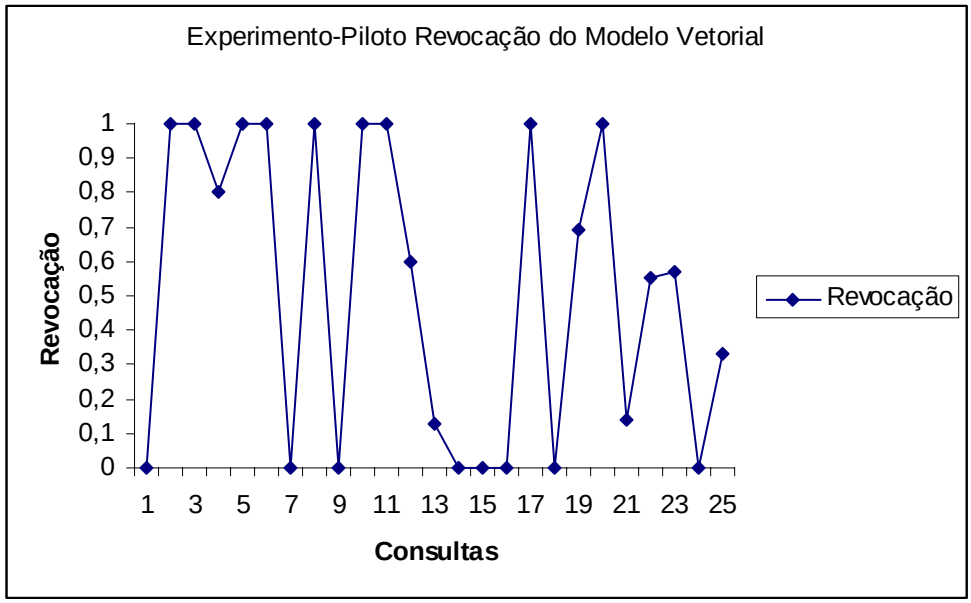


Figura 4.9.: Resultado de Revocação do Experimento-Piloto utilizando o Modelo Vetorial

Na Figura 4.9 estão representados os resultados de Revocação obtidos nas consultas. Como mencionado anteriormente, temos que a Revocação é o número total de documentos relevantes retornados dividido pelo número total de documentos relevantes para aquela consulta. O Modelo Vetorial, apresentou revocação média de 0,51.

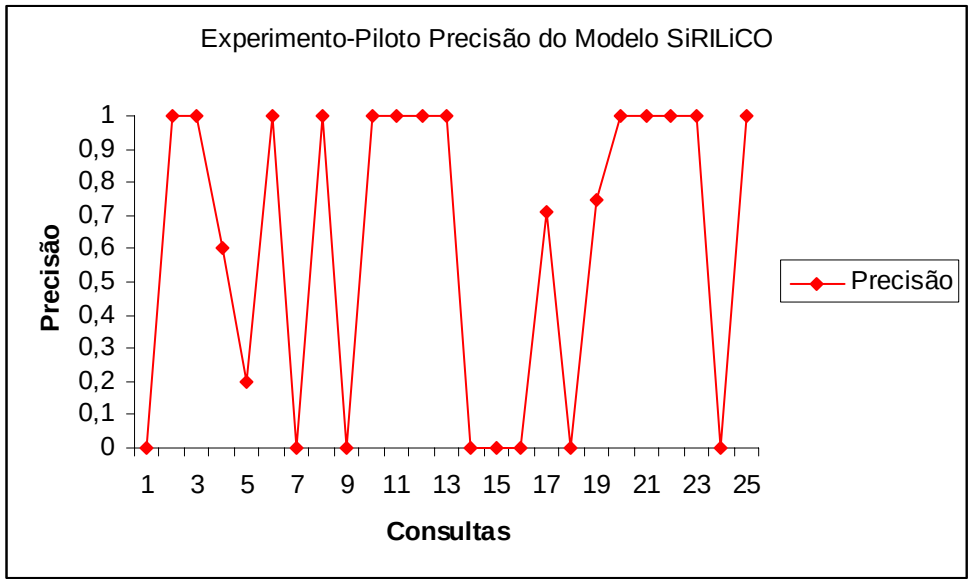


Figura 4.10.: Resultado de Precisão do Experimento-Piloto utilizando o Modelo SiRILiCO

Na Figura 4.10 apresentam-se os resultados de Precisão obtidos nas consultas realizadas com o Sistema de Recuperação de Informação que utiliza abordagens de análise sintática e semântica dos textos dos documentos da coleção a ser indexada. Esse modelo apresentou precisão média de 0,61 para as consultas realizadas no Experimento-Piloto.

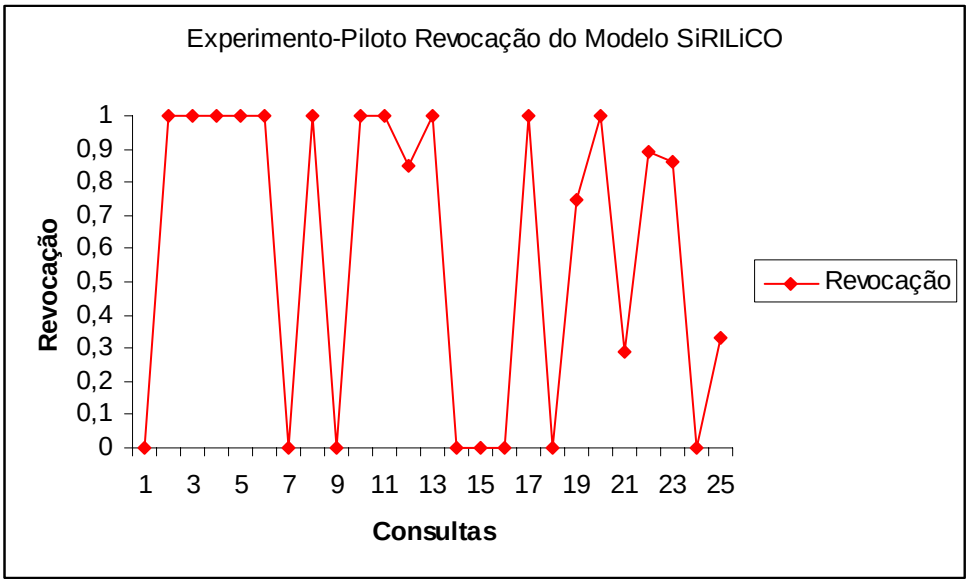


Figura 4.11.: Resultado de Revocação do Experimento-Piloto utilizando o Modelo SiRILiCO.

A Figura 4.11 estão representados os resultados de Revocação obtidos nas consultas do Modelo *SiRILiCO*. A revocação média foi de 0,60.

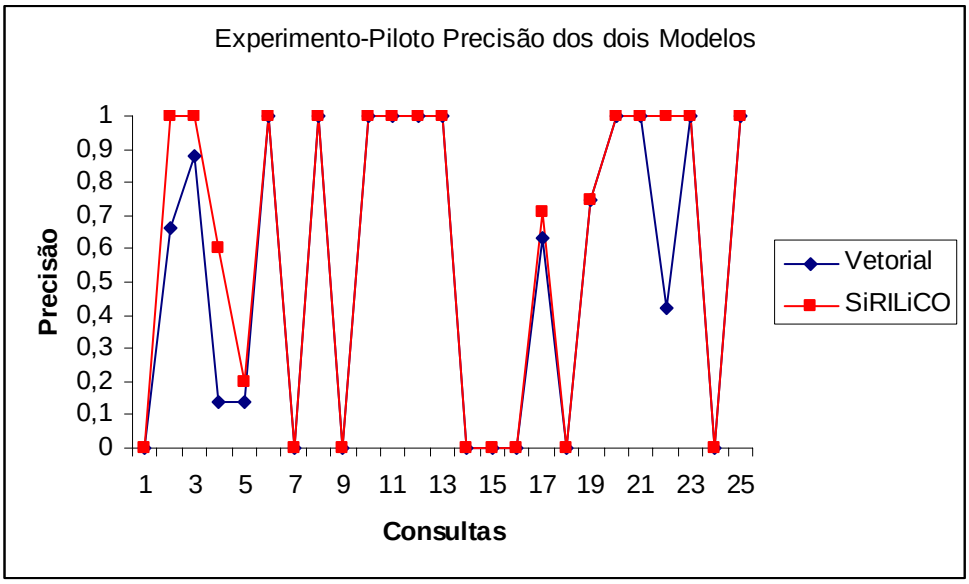


Figura 4.12.: Resultado de Precisão do Experimento-Piloto comparando os dois Modelos.

A Figura 4.12 observam-se os resultados de Precisão obtidos tanto para o Modelo Vetorial quanto para o Modelo SiRILiCO. Os resultados de Precisão apresentados pelo Modelo SiRILiCO foram em média 9,84 % (nove, oitenta e quatro por cento) superiores aos resultados apresentados pelo Modelo Vetorial nas consultas realizadas no Experimento-Piloto.

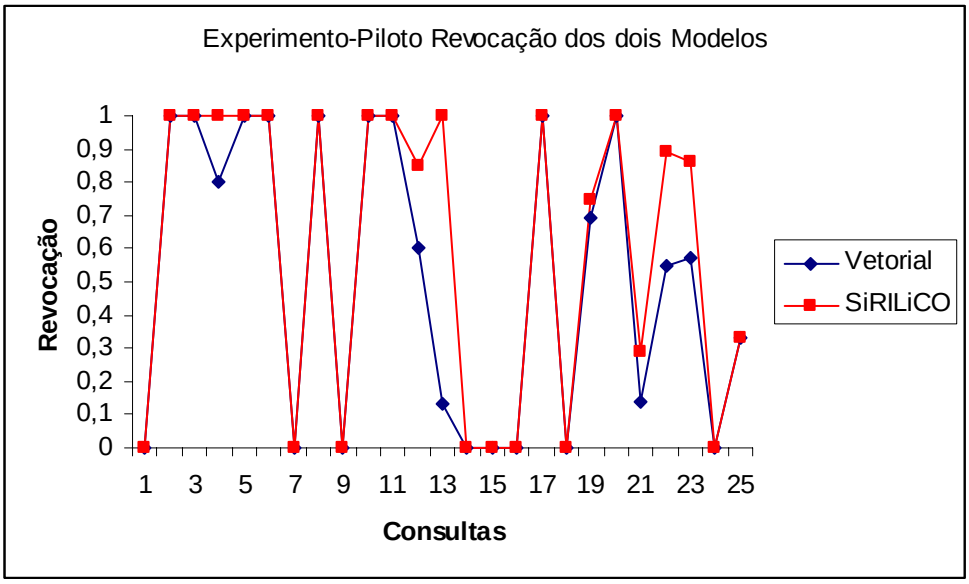


Figura 4.13.: Resultado de Revocação do Experimento-Piloto comparando os dois Modelos.

Na Figura 4.13 estão os resultados de Revocação obtidos nas consultas realizadas com os dois modelos, o Modelo Vetorial e o Modelo SiRILiCO. Os resultados de Revocação apresentados pelo Modelo SiRILiCO foram em média 14,96 % (catorze, noventa e seis por cento) superiores aos resultados apresentados pelo Modelo Vetorial nas consultas realizadas no Experimento-Piloto.

4.3. Resultados do Experimento de Validação

Nesta seção apresentamos os resultados produzidos no Experimento de Validação. Esse Experimento contou com 221 textos da Revista Ciência da Informação referentes aos 3 números do Volume 24 (1996), ao Volume 31 (2002) e o número 1 do Volume 32 (2003). No total são 104206 palavras. Foram identificados e extraídos automaticamente 4619 Sintagmas Nominais. A partir desses Sintagmas, foram geradas 4619 Sub-Classes da Classe OBJETO-AFETADO da Ontologia do Experimento de Validação. Foram identificados e extraídos, automaticamente, 2542 Sintagmas Verbais. A partir desses Sintagmas, foram geradas 2542 Sub-Classes da Classe Predicador da Ontologia do Experimento de Validação. Também foram identificados e extraídos, automaticamente, 1206 Nomes Próprios. A partir desses Nomes Próprios, foram geradas 1206 Sub-Classes da Classe AGENTE da Ontologia do Experimento de Validação.

Para a realização do Experimento de Validação utilizou-se dos mesmos sujeitos e critérios adotados no Experimento-Piloto. Os mesmos programas de computador utilizados na criação e utilização dos módulos do protótipo do Experimento-Piloto também foram empregados no Experimento de Validação. Além dos textos utilizados no Experimento-Piloto, utilizou-se mais 180, totalizando 221 textos extraídos da Revista da Ciência da Informação. Ou seja, todos os artigos em português, sequencialmente, da Revista 24, número 01, 1995 até a revista 32 número 01, 2003. O fato desta escolha deve-se ao fato de que o número 01 da revista 32 de 2003 ter sido o ultimo número a ser editado na data de realização do experimento deste estudo. Os 37 alunos que colaboraram com o Experimento-Piloto também colaboraram com o segundo experimento.

Consulta	Vetorial Precisão	Vetorial Revocação	**SiRILiCO Precisão**	**SiRILiCO Revocação**
1	0,33	0,5	**1**	**0,66**
2	0,55	0,9	**0,79**	**1**
3	0,34	0,85	**0,65**	**1**
4	1	0,25	**1**	**0,38**
5	0,8	0,66	**1**	**1**
6	1	1	**1**	**1**
7	1	0,25	**1**	**0,25**
8	0,5	0,7	**1**	**0,9**
9	1	1	**1**	**1**
10	0	0	**0**	**0**

Tabela 4.2 Precisão e Revocação do Modelo Vetorial e do Modelo SiRILiCO para as consultas realizadas no Experimento de Validação.

Na tabela 4.2 apresentam-se os resultados obtidos para as consultas utilizando-se o Modelo Vetorial e o Modelo SiRILiCO. Reforça-se a sugestão da tabela 4.1 de que a utilização de Sintagmas Nominais para a geração de Classes em uma Ontologia, que servem como indexadores, otimiza o desempenho do Sistema de Recuperação de Informação através da valorização do conceito, do valor semântico do termo.

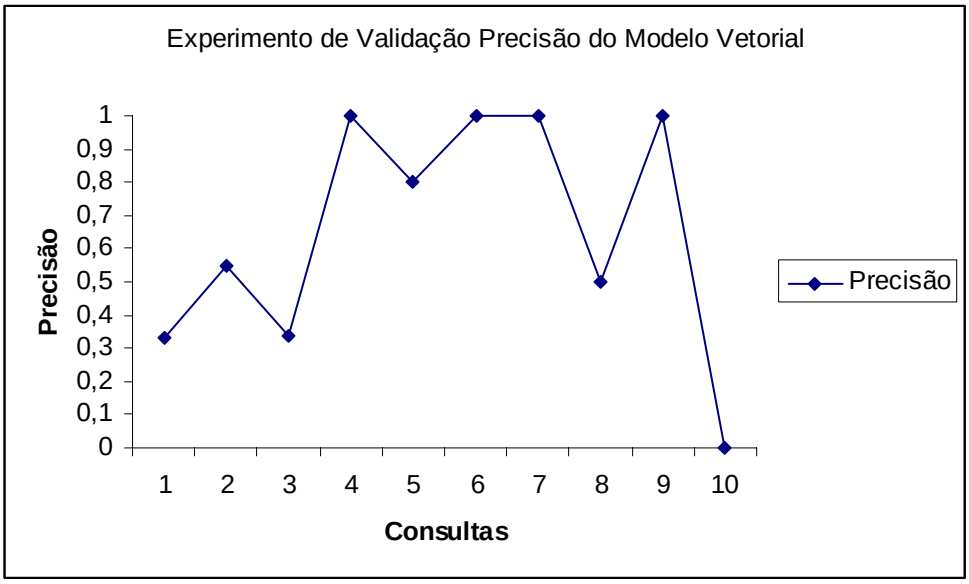


Figura 4.14.: Resultado de Precisão do Experimento de Validação utilizando o Modelo Vetorial.

Na Figura 4.14 estão os resultados de Precisão obtidos nas consultas feitas com o Modelo Vetorial na coleção do Experimento de Validação. O Modelo Vetorial apresentou precisão média de 0,65 para as consultas realizadas no Experimento de Validação (veja tabela 4.2).

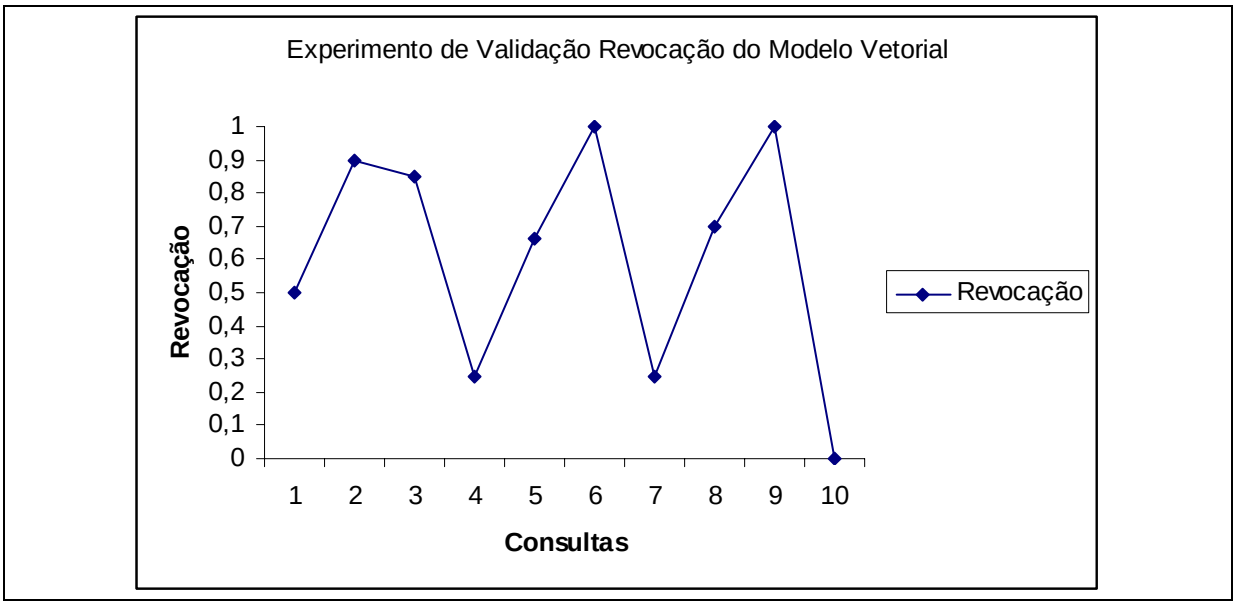


Figura 4.15.: Resultado de Revocação do Experimento de Validação utilizando o Modelo Vetorial

Na Figura 4.15 estão os resultados de Revocação obtidos nas consultas feitas com o Modelo Vetorial na coleção do Experimento de Validação. O Modelo Vetorial apresentou precisão média de 0,61 para as consultas realizadas (veja tabela 4.2).

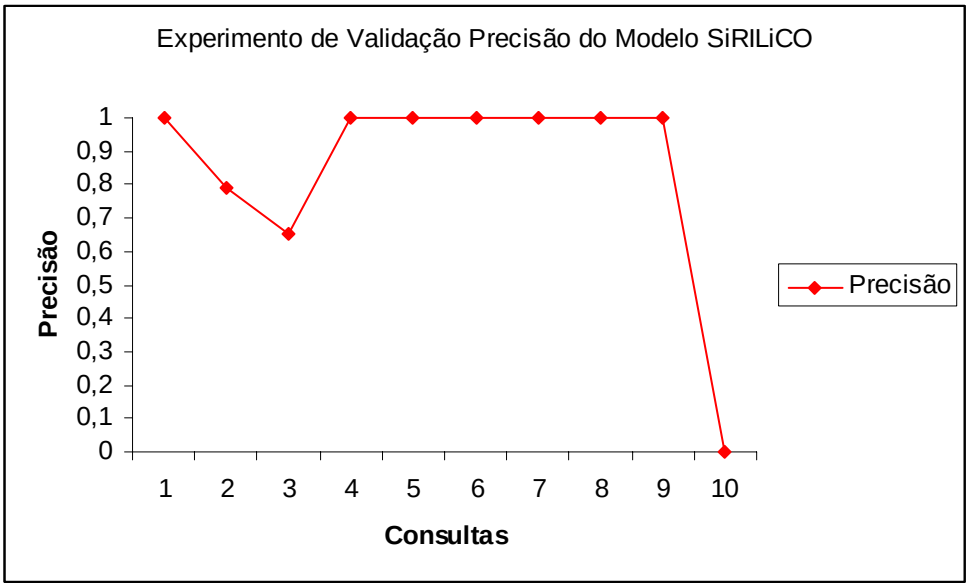


Figura 4.16.: Resultado de Precisão do Experimento de Validação utilizando o Modelo SiRILiCO

Na Figura 4.16 são apresentados os resultados de Precisão obtidos nas consultas feitas com o Modelo SiRILiCO na coleção do Experimento de Validação. O Modelo SiRILiCO apresentou precisão média de 0,84 para as consultas realizadas (veja tabela 4.2).

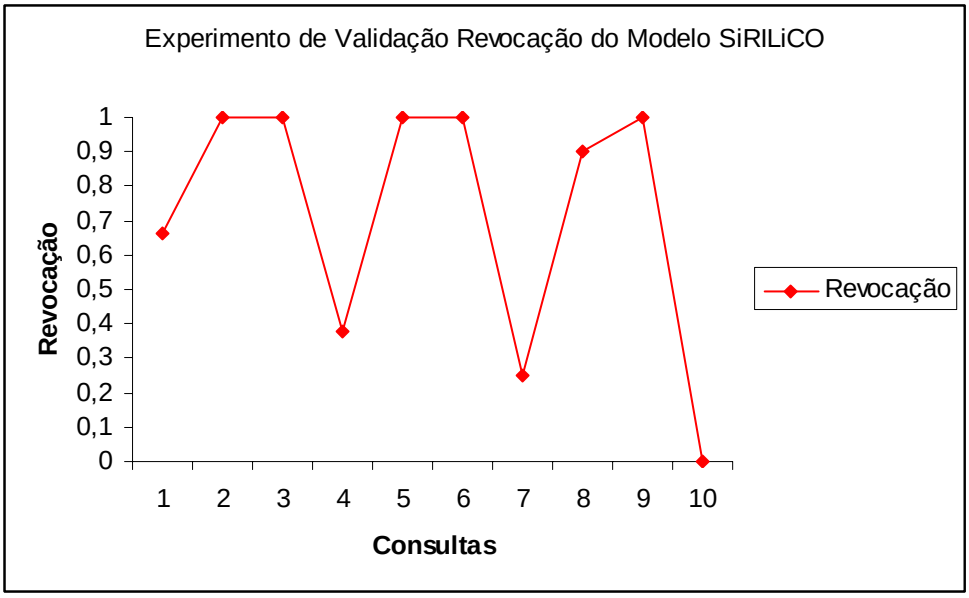


Figura 4.17.: Resultado de Revocação do Experimento de Validação utilizando o Modelo SiRILiCO.

Na Figura 4.17 estão os resultados de Revocação obtidos nas consultas feitas com o Modelo SiRILiCO na coleção do Experimento de Validação. O Modelo Vetorial apresentou precisão média de 0,72 para as consultas realizadas no mesmo Experimento (veja tabela 4.2).

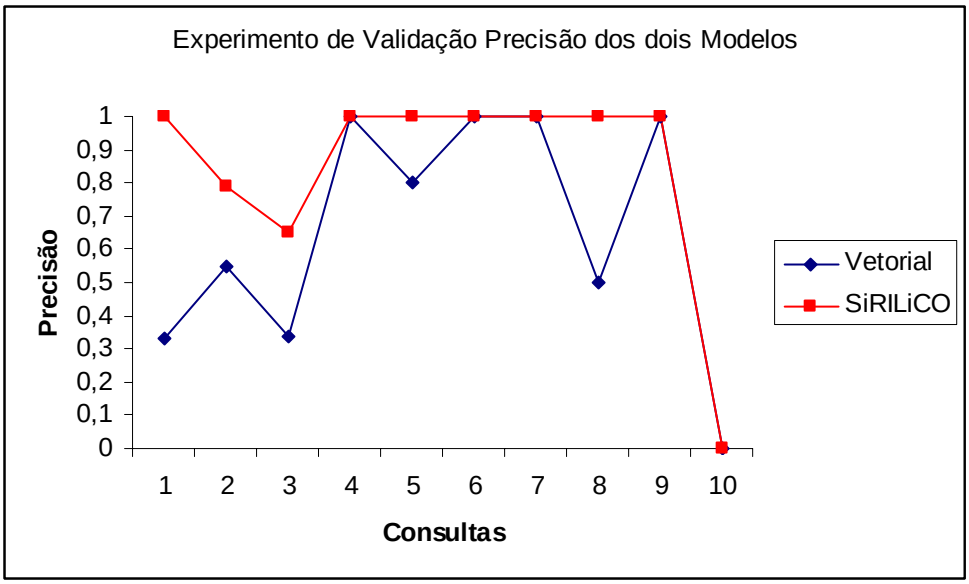


Figura 4.18.: Resultado de Precisão do Experimento de Validação comparando os dois Modelos.

Na Figura 4.18 estão os resultados de Precisão obtidos nas consultas realizadas com os dois modelos, o Modelo Vetorial e o Modelo SiRILiCO. Os resultados de Precisão apresentados pelo Modelo SiRILiCO foram em média 22,75 % (vinte e dois, setenta e cinco por cento) superiores aos resultados apresentados pelo Modelo Vetorial nas consultas realizadas no Experimento.

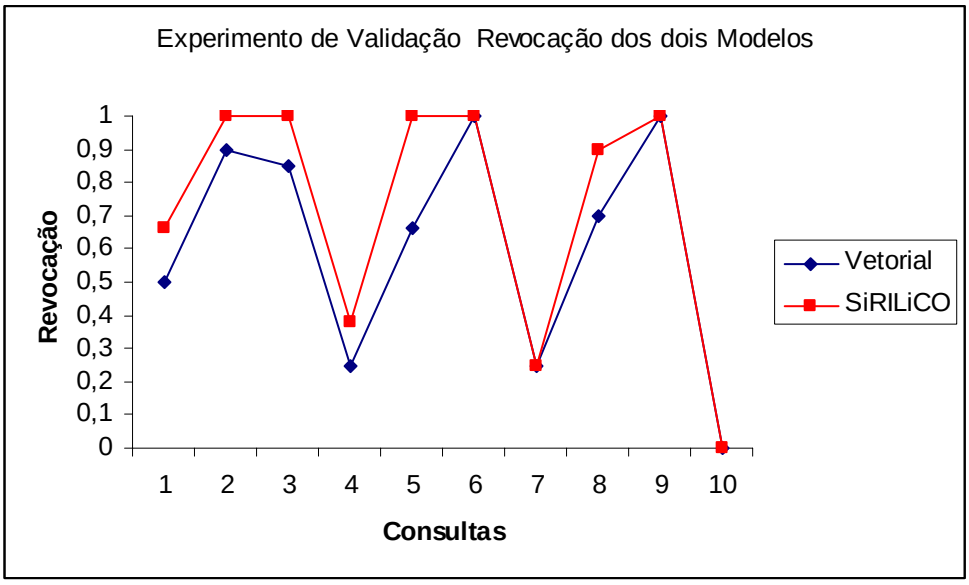


Figura 4.19.: Resultado de Revocação do Experimento de Validação comparando os dois Modelos.

Na Figura 4.19 estão os resultados de Revocação obtidos nas consultas realizadas com os dois modelos, o Modelo Vetorial e o Modelo SiRILiCO. Os resultados de Revocação apresentados pelo Modelo SiRILiCO foram em média 15 % (quinze por cento) superiores aos resultados apresentados pelo Modelo Vetorial nas consultas realizadas no Experimento.

A precisão média obtida pelo Modelo Vetorial no Experimento-Piloto foi de 0,55, enquanto que a precisão média obtida pelo Modelo *SiRILiCO* foi de 0,61. Esse pequeno ganho (0,06) deve-se principalmente ao fato de que a coleção é pequena e muito homogênea. Por exemplo, para a consulta número 04, cujo termo é "Ciência da Informação" (Anexo 7), o Modelo Vetorial, baseado apenas na existência do termo, recuperou vários documentos em cujo cabeçalho constava o termo, tendo assim uma precisão de 0,18. Para a mesma consulta o Modelo *SiRILiCO,* utilizando-se de informações semânticas para indexar, apresentou uma precisão de 0,60. Ou seja, apenas os documentos que continham "Ciência da Informação" como proposição ou parte de uma proposição é que foram retornados pelo Modelo *SiRILiCO.* Com a utilização dessa simples regra o Modelo *SiRILiCO* apresentou uma resposta a consulta 3 vezes melhor que a resposta do Modelo Vetorial, o ganho qualitativo do Modelo *SiRILiCO* sobre o Modelo Vetorial

A revocação média obtida pelo Modelo Vetorial no Experimento- Piloto foi de 0,51, enquanto que a revocação média obtida pelo Modelo *SiRILiCO* foi de 0,60. Houve um pequeno ganho médio (0,09) para o Modelo *SiRILiCO.* As consultas "Tecnologia da Informação" e "Filosofia e Prática" foram as que apresentaram as maiores diferenças entre as respostas de revocação entre os modelos, isso provavelmente porque o Modelo Vetorial desconsidera o termo "Tecnologia de Informação", além do mais, no caso do Modelo *SiRILiCO,* "Tecnologia da Informação", por exemplo, aparece contextualizado, como na seguinte proposição: "As novas Tecnologias de Informação e a capacitação dos recursos humanos lotados na burocracia estatal"

A precisão média obtida pelo Modelo Vetorial no Experimento foi de 0,65, enquanto que a precisão média obtida pelo Modelo *SiRILiCO* foi de 0,84. Houve um ganho de 0,19 a favor do Modelo *SiRILiCO* , nota-se que com o aumento da coleção o Modelo Vetorial, que preza pela quantidade, aumenta o ruído nas respostas. Isto é, provavelmente, devido ao fato de que com o aumento da coleção aumenta o número de termos, sem necessariamente aumentar o conteúdo semântico da coleção.

A revocação média obtida pelo Modelo Vetorial no Experimento de Validação foi de 0,61, enquanto que a revocação média obtida pelo Modelo *SiRILiCO* foi de 0,72. Houve um ganho médio (0,11) para o Modelo *SiRILiCO.* A revocação é favorecida pelas abordagens qualitativas (Modelo *SiRILiCO).* Os SRIs qualitativos retornam um número menor de textos para as consultas, porém textos significativos, que contém "semanticamente" as mesmas, ou seja, os termos estão contextualizados nos textos retornados pelo sistema, não são apenas palavras soltas mas sim palavras que estão contidas em sintagmas nominais, que estão contidos em proposições.

4.4. Questões limitadoras do trabalho

Durante o desenvolvimento dos protótipos e da realização dos experimentos surgiram alguns problemas de ordem técnica. Esses problemas não comprometem os pressupostos elaborados, defendidos e testados nesse trabalho, mas merecem especial atenção para a efetiva implementação de um Sistema de Recuperação de Informação baseado na Lingüística Computacional e Ontologia. Os problemas podem ser resumidos a um só, ruído (informação errônea). O protótipo foi desenvolvido com a contribuição de programas já existentes e disponíveis gratuitamente. O volume de trabalho exigido para a elaboração e implementação de um SRI demanda um período de tempo que vai alem dos 04 anos de pesquisa de doutorado. Por esta razão foi preciso considerar os ruídos e tratá-los ao longo do experimento.

A análise automática de textos, como já explanado ao longo desse estudo, ainda é falha, a análise sintática ainda apresenta algumas ambigüidades difíceis de serem tratadas automaticamente e a análise semântica ainda está começando a apresentar seus primeiros resultados promissores. Um exemplo desse ruído: "ser lingüístico" deveria ser um nome, na análise automática temos que "ser" é verbo e "lingüístico" é um adjetivo. Como o analisador sintático apresentou a análise erroneamente, o analisador semântico dará continuidade ao erro, então "ser" será um PREDICADOR quando o esperado seria "ser lingüístico" ser um OBJETO-AFETADO.

Estima-se que aproximadamente 5 a 10% do número de classes geradas automaticamente na ontologia apresentem propagação do ruído gerado na análise sintática (Anexo 7). Calcula-se também que o número de classes geradas é inferior ao número real, devido às falhas do protótipo. Esses problemas não afetaram o experimento-piloto nem o experimento. A verificação dos resultados deu-se manualmente e foi repetida várias vezes, em

conformidade com os algoritmos apresentados no capítulo 3. Para preservar o caráter cientifico dos experimentos, optou-se por utilizar apenas 10 termos nas consultas realizadas no Experimento de Validação, facilitando assim o controle sobre o mesmo.

Os resultados apresentados pelos dois modelos apresentam de um modo geral, vantagem em favor do Modelo *SiRILiCO*.

5. Discussão, Conclusão e Considerações Finais

5.1. Discussão

Este trabalho desenvolveu-se com base na abordagem de que a Lingüística Computacional (OKSEFJELL & SANTOS, 1998; NUNES et al., 1999; MITKOV, 2003) e a Ontologia (CORAZZON, 2003) podem oferecer grandes contribuições para as áreas de disseminação e recuperação de informação, especialmente para o tratamento automático de textos. Este estudo descreveu a estruturação de um Sistema de Recuperação de Informação aplicado a uma coleção de textos em língua portuguesa que versam sobre a Ciência da Informação. Para o auxílio da especificação de consultas utilizou-se de conceitos extraídos dos textos da própria coleção, que, depois de identificados e extraídos automaticamente foram disponibilizados em uma estrutura, em uma ontologia leve. Os resultados experimentais obtidos, apesar da precariedade do protótipo desenvolvido em vista da potencialidade do modelo aqui proposto, foram mais que satisfatórios e sugerem que os usuários de SRI podem beneficiar-se enormemente de uma estrutura hierárquica desenvolvida a partir de análise lingüística de textos. A coleção criada e analisada para este estudo, 221 textos analisados em sua totalidade sintaticamente e em partes semanticamente, será disponibilizada e utilizada em trabalhos futuros relacionados à Lingüística (avaliação de análise sintática automática, extração automática de papéis semânticos, tradução automática, etc.) e à recuperação de informação (avaliação de SRIs, desenvolvimento de interfaces, expansão automática de consultas, etc).

5.1.1. Conclusão

O Modelo *SiRILiCO*, apesar dos problemas de ruído apresentados pelo protótipo, apresentou resultados superiores aos resultados apresentados pelo Modelo Vetorial para a coleção em questão. A idéia de utilizar conhecimento de ciências cognitivas para indexar uma coleção de documentos eletrônicos através de frases com conteúdo semântico (proposições), mostrou-se promissora.

Embora o Modelo *SiRILiCO* tenha sido testado através de uma coleção pequena e homogênea, constatou-se empiricamente a qualidade desse modelo. A idéia inicialmente proposta, desenvolver um SRI para lidar com todo e qualquer tipo de coleção, independente de homogeneidade/heterogeneidade dos documentos que compõem a coleção, da língua em que os documentos foram escritos e do tamanho da coleção, não foi refutada. A simplicidade e aparente robustez (capacidade de processar qualquer entrada) do Modelo *SiRILiCO* são qualidades que permitem a otimização e reformulação das hipóteses defendidas ao longo deste trabalho. A continuidade deste trabalho deve ser uma boa opção para as pesquisas que visam à Web Semântica.

O desenvolvimento e implementação de um produto, um software (GeraOnto) que permite a criação automática de uma "ontologia leve" em português a partir de uma pequena coleção de textos contribui para o desenvolvimento da área de Tratamento e Recuperação de Informação.

A disponibilização dessa pequena coleção para experimentos, cujos textos foram analisados sintática e semanticamente de maneira automática, sendo que a análise semântica ainda encontra-se muito incipiente, é uma contribuição para os estudos de análise automática de textos, lingüística textual e análise do discurso.

Os modelos de Sistema de Recuperação de Informação mais utilizados (quantitativos) baseiam-se no Modelo Vetorial de Salton (PERSIN, 1994) e utilizam-se de casamento de padrão entre as palavras-chave presentes na consulta do usuário e presentes nos documentos que compõem a coleção indexada pelo sistema.

Um SRI tradicional (nos moldes citados anteriormente) foi utilizado para contrastar os resultados obtidos com o modelo *SiRILiCO*. O Modelo *SiRILiCO* prezou (apesar dos ruídos apresentados pelo protótipo) pela qualidade da indexação através de conceitos. A indexação deu-se por proposições (Sintagmas Nominais e Sintagmas Verbais) e não meramente por freqüência de termos.

Nenhum dos dois modelos aqui apresentados (Modelo Vetorial e *SiRILiCO*) realizou tratamento de normalização das variações lingüísticas (técnica de *stemming* ou de busca em Tesaurus) ou de *stopwords*, pois essas técnicas podem descontextualizar um termo específico ou até mesmo dificultar o julgamento de relevância de um documento (RILOFF, 1995). Portanto, ambos os modelos apresentavam-se destituídos de recursos auxiliares (como os citados acima) que pudessem otimizar as respostas às consultas fornecidas pelos sistemas aos usuários.

5.2 Considerações Finais

5.2.1. Principais Contribuições desta Pesquisa

Foi possível demonstrar que é viável a criação de uma ontologia leve, automaticamente única e exclusivamente a partir de análises sintáticas e semânticas dos textos da coleção da qual se quer uma ontologia. A criação e análise de uma coleção de documentos em português do Brasil toda analisada sintaticamente e parcialmente analisada semanticamente também é uma contribuição, visto que não existem muitas coleções desse tipo disponibilizadas para experimentos.

Este estudo apresentou como contribuição para a Ciência da Informação a possibilidade de desenvolver um modelo de um sistema de recuperação de informação e potencialmente, implementá-lo utilizando-se de teorias de Lingüística e Lingüística Aplicada. Isto abre um leque de possibilidades de estudos, como por exemplo: geração automática de índice; recuperação automática de informação e utilização de ontologias para a filtragem da busca do usuário da informação, dentre outros. Para o Processamento de Linguagem Natural, as possibilidades de pesquisa passam pelo processamento e análise automática de textos, fomentando, através da utilização dos módulos que compõem o protótipo deste estudo, até a possibilidade de “compreensão” automática de texto. A adaptação do modelo de análise proposicional do texto (FREDERIKSEN, 1975), serve para que se desenvolva mais subsídios para a efetivação da Web Semântica, tal qual proposta. Além disto, este estudo fornece subsídios metodológicos para pesquisadores interessados não somente nas questões pautadas na Ciência da Informação, mas também nas questões relacionadas com Lingüística Aplicada, Lingüística Computacional e geração, manutenção, tratamento de Ontologias.

Este estudo apresenta como contribuições empíricas, seus dados sistemáticos sobre textos em português do Brasil, analisados sintaticamente, e, em parte, semanticamente. Isto pode também servir como embasamento e parâmetro para estudos em outras línguas, principalmente porque não existem muitos estudos a respeito de análise semântica automática ou mesmo a respeito da geração automática de uma ontologia leve utilizando-se de produtos oriundos de analisadores automáticos de sintaxe e semântica. O processamento automático específico de características semânticas de textos em português também é uma colaboração deste estudo.

5.3 Outros Problemas de Pesquisa

O presente estudo, principalmente devido ao tema abordado ser de interesse multidisciplinar, deixa muitas questões específicas a serem discutidas. Apesar dos inúmeros benefícios que as novas tecnologias podem propiciar ao ser humano em se tratando de aquisição de informação (GRÉGOIRE et al., 1996; KINTSCH et al., 1993; NEGROPONTE, 1995; GATES et al., 1995; LÈVY, 1995), como recuperá-las, processá-las automaticamente de uma maneira adequada? Como produzir Sistemas de Recuperação de Informação mais eficientes? Qual a real possibilidade que a Lingüística do Texto fornece ao desenvolvimento de processadores automáticos de texto como mineradores de texto, por exemplo? Quais estratégias cognitivas utilizadas pelos seres humanos na leitura podem ser emuladas pelas máquinas para auxiliar o tratamento automático de textos?

Pesquisas como este estudo permitem, cada vez mais, a compreensão do processo de indexação automática e de identificação e extração de conhecimentos através de técnicas baseadas em abordagens lingüísticas. Fornecem subsídios teóricos mais precisos que irão auxiliar o “profissional da informação” e o “profissional do conhecimento” a desenvolverem de uma maneira otimizada os seus trabalhos.

5.4. Trabalhos Futuros

Como trabalhos futuros sugerem-se temas para o desenvolvimento de pesquisas nas áreas de Recuperação de Informação, através de novos experimentos envolvendo a coleção ora utilizada. Em particular trabalhos que envolvam a expansão de consultas e o desenvolvimento de interfaces gráficas específicas para SRIs. Esta abordagem teórica, explicitada com o Modelo *SiRILiCO,* pode contribuir para as possíveis soluções do problema

de visualização das consultas e das respostas, através da exibição de proposições e não de meros termos ou de meros trechos desconexos dos textos indexados.

Sugere-se também pesquisas que permitam a combinação dessa abordagem com outras. Pode-se utilizar essa abordagem como filtro de um SRI baseado no Modelo Vetorial, ou vice-versa. Para a área de Lingüística Computacional é possível desenvolver pesquisas que otimizem os analisadores sintáticos e semânticos, através do desenvolvimento de *Chunk Parsers*, por exemplo, que são analisadores sintáticos que processam apenas partes específicas do texto, visando principalmente à diminuição do ruído inerente ao sistema.

Pode-se também utilizar essa abordagem para desenvolver pesquisas na área de geração automática de textos, e/ou mineração de textos. A sumarização automática de textos e o desenvolvimento de material didático-pedagógico para a Educação a Distância (*eLearning*) também são áreas que podem ser auxiliadas pelas contribuições acadêmicas expostas neste trabalho.

Outra questão a ser tratada é a expansão das consultas formuladas pelos usuários. Esta é, sob a ótica da Recuperação da Informação, uma aplicação típica para ontologias.

REFERÊNCIAS BIBLIOGRÁFICAS

AFONSO, S.; BICK, E.; HABER, R. & SANTOS, D. Floresta sintá(c)tica: a treebank for Portuguese, Proceedings of LREC'2002, 2002.

AIRES, R. & SANTOS, D. 2002. Measuring the Web in Portuguese. Euroweb 2002 Conference. The Web and the GRID: from e-science to e-business. St Anne's College Oxford, UK, December, 17 and 18th 2002.

ALLEN, B. L. *Information tasks: toward a user-centered approach to information systems*. San Diego: Academic Press, 1996.

ALMEIDA, M. B. Uma Introdução ao XML, sua utilização na Internet e alguns conceitos complementares. Revista Ciência da Informação Vol. 31, N. 2, 2002.

ALTAVISTA. www.altavista.com

ARAÚJO, A. de A. & GUIMARÃES, S.J.F. Recuperação de informação visual com base no conteúdo em imagens e vídeos digitais, Edição Especial em Computação Gráfica e Processamento de Imagens, Revista de Informática Teórica e Aplicada - RITA, UFRGS, Porto Alegre-RS, Brazil, vol. 7, no. 2, ISSN no. 0103 4308, 2000.

ARAUJO, G. M. L. & LUNA, M. J. de M. A contribuição da lingüística textual no processo de indexação. Seminário Nacional de Bibliotecas Universitárias, XII SNBU 2002, Recife, 2002.

ARMS, W. Y. Digital Libraries (Digital Libraries and Electronic Publishing). The Mit Press, Cambridge, MA, London, England, 2000.

BAEZA-YATES, R. & CASTILLO, C. Crawling the Infinite Web: Five Levels are Enough. Proceedings of the third Workshop on Web Graphs. Springer, Rome, Italy. 2004.

BAEZA-YATES, R. & RIBEIRO-NETO, B. Modern Information Retrieval. Addison-Wesley Pub Co; 1st edition (May 1999).

BALMIN, A., HRISTIDIS, V. and PAPAKONSTANTINOU, Y. ObjectRank: Authority-based keyword search in databases. VLDB, 2004.

BATES, M. J. 1999. The Invisible Substrate of Information Science. Journal of The American Society for Information Science. 50(12) 1043-1050.

BAUMANN, J. F. Anaphora in Basal Reader Selections: How Frequently do They Occur? *Journal of Reading Behavior*, Vol. XIX, Nº 2, 1987.

BENNETT, J.; TONG, X.; EVANS, D. A. CLARIT TREC-8 Experiments in Searching Web Data. 1999.
http://trec.nist.gov/pubs/trec8/papers/index.track.html.

BERNERS-LEE, T.; HENDLER, J. & LASSILA, O. The semantic web. Scientific American, May 2001.

BICK, E. Automatic Parsing of Portuguese. In García, Laura Sánchez (ed.), Anais / II Encontro para o Processamento Computacional de Português Escrito e Falado. Curitiba: CEFET-PR. 1996.

_____ The Parsing System Palavras: automatic grammatical analysis of Portuguese in a constraint grammar framework. Arhus University Press. (PhD Dissertation), 2000.

_____ Portuguese Syntax (Teaching Manual), epositorio/Bick_Portuguese_Syntax3.doc and http://visl.sdu.dk/visl/pt , 2000-1.

BLAIR, D. C. Language and Representation in Information Retrieval. Elsevier Science Publishers, 1990.

BORST, W. N. Construction of Engineering Ontologies. PhD Thesis. 1997. Disponivel em Http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf

BREWSTER, C. Techniques for Automated Taxonomy Building: Towards Ontologies for Knowledge Management . In Proceedings CLUK Research Colloquium, Leeds, UK. 2002.

BRIGHTPLANET, 2000. White Paper. www.brightplanet.com.

BRIN, S.& PAGE, L. "The Anatomy of a Large-Scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, April 1998. http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

BUSH, V. As We May Think. Atlantic Monthly, Vol. 176 n. 1, pag. 101-108, 1945.

CADE, 2004. www.cade.com.br.

CANCEDDA, N., GAUSSIER E., GOUTTE C. and RENDERS J.-M. Word-Sequence kernels. In Journal of Machine Learning Research, Special Issue on Machine Learning Methods for Text. 2003

CAPURRO, R. Lässt sich Wissen managen? Eine informationswissenschaftliche Perspektive. 2003. http://www.capurro.de/wissensmanagement.html

CENDÓN, B. Ferramentas de Busca na Web. Revista Ciência da Informação, Vol. 30, n. 1, pág. 39-49, 2001.

CENDÓN, B. Sistemas e redes de informação. ECI-UFMG, Texto Didático. 2005.

CHAKRABARTI, S.; DOM, B.; RAGHAVAN, P.; RAJAGOPALAN, S.; GIBSON, D.; KLEINBERG, J. 1998. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. Proc. 7th International WWW, 1998.

CHAKRABARTI, S.; DOM, B.; RAGHAVAN, P.; RAJAGOPALAN, S.; GIBSON, D.; KLEINBERG, J. Hypersearching the Web. Scientific American, 1999.

CHAN, L. M.; RICHMOND, P.; SVENONIUS, E. (org.). THE NEED for a faceted classification as the basis of all methods of information retrieval. In: Theory of subject analysis; a sourcebook. Littleton, Co.: Libraries Unlimited, 1985.

CHO, J.; GARCIA-MOLINA, H.; LAWRENCE, P. Efficient Crawling Through URL Ordering. The Seventh International WWW Conference, Brisbane, Australia, 1998.

CHOMSKY, N. *Knowledge of language: it's nature, origin, and use*. New York: Praeger Publishers, 1986.

_____ *The Minimalist Program*. The MIT Press. Cambridge, Massachusetts, London England, 1995.

CLARK, H. H. & CLARK, E. V. *Psychology and Language*. New York: Hardcourt Brace Jovanovich, 1977.

CLEVELAND, D. B. & CLEVELAND, A.D. Introduction to Indexing and Abstracting. Libraries Unlimited, 3 edição, 2000.

CORAZZON, R. 2003. Descriptive and Formal Ontology.
http://www.formalontology.it/index.htm

COSTA, F. & FRASCONI, P. Distributed Community Crawling. ACM, WWW2004. 2004.

CRAIN, S. & STEEDMAN, M. On not being led up the path: the use of context in the psychological syntax processor. In: DOWTY, D.R., KARTUNNEN, L. ZWICKY, A.M.(Ed.). *Natural language parsing*. Cambridge: Cambridge University, 1985. p.320-358.

CROFT, W. B., TURTLE, H. R.; LEWIS, D. D. The use of phrases and structured queries in information retrieval. Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval. 1991.

DAHLBERG, I. Begriffs- und Definitionstheorie in ihrem Zusammenhang. I: DUTZ. K. D. (HRSG.) Studien zur Klassifikation, Systematik und Terminologie. Theorie und Praxis. Studium Sprachwissenshaft, Beiheft 5, Arbeiten zur Klassifikation. Institut für Allgemeine Sprachwissenshaft der Wesfählischen Wilhelms Universität, Münster. 1984

DAVIES J.; FENSEL, D. and van HARMELEN, F. Towards The Semantic Web. Ontology-Driven Knowledge Management. John Wiley an Sons Ltd. 2003.

DENTON, W. How to Make a Faceted Classification and Put It On the Web. November 2003. http://www.miskatonic.org/library/facet-web-howto.html

DERTOUZOS, M. O Que Será – Como o Novo Mundo da Informação Transformará Nossas Vidas. Companhia das Letras. 1997.

DING, Y. IR and AI: The role of ontology. In *Proc. 4th International Conference of Asian Digital Libraries*, Dec 10-12, Bangalore, India. 2001.

DING, Y. & ENGELS, R. IR and AI: Using Co-occurrence Theory to Generate Lightweight Ontologies. DEXA Workshop, pp 961-965. 2001.

DING, Y. & FOO, S. Ontology Research and Development: Part 1 – A Review of Ontology Generation. Journal of Information Science 28(2). 2002.

DRESNER E. and DASCAL M. Semantics, pragmatics, and the digital information age. Studies in Communication Sciences 1(2): 1-22. 2001.

ELLIS, D. *Progress and Problems in Information Retrieval*. London: Library Association Publishing, 1996.

ERDMANN, M.; MAEDCHE, A.; SCHNURR, H. –P.; STAAB, S. From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools. In: ETAI Journal – Section on Semantic Web (Linköping Electronic Articles in Computer and Information Science), 6(2001). 2001.

FARIA, I. H.; PEDRO, E. R.; DUARTE, I.; GOUVEIA, C. A.M. (orgs.), Introdução à Linguística Geral e Portuguesa, Lisboa: Caminho. 1998

FALOUTSOS, C. & OARD, D. A Survey of Information Retrieval and Filtering Methods. Technical Report CS-TR-3514. Dep. Of Computer Science, Un. Of Maryland. 1995.

FAURE, D. & NEDELLEC, C. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In. LREC workshop on adapting lexical and corpus resources to sublanguages and applications. Granada, Spain, 1998.

FERNEDA, E. Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. Tese de Doutorado. Escola de Comunicação e Artes. Universidade de São Paulo. São Paulo, 2003.

FILLMORE, C. "The case for case". In Universals in Linguistic Theory, edited by E. Bach and R. Harms. New York: Holt, Rinehart, and Winston. 1968.

FRAKES, W. & BAEZA-YATES, R. Information Retrieval: Data Structure & Algorithms. Prentice Hall, 1992.

FREDERIKSEN, C. Representing Logical and Semantic Structure of Knowledge Aquired from Discourse. Cognitive Psychology 7, pp 371-458, 1975.

_____ Cognitive models and discourse analysis. In: Cooper, C. , Greenbaum, S. (Eds.). Studying writing: Lingustic approaches. Beverly Hills, CA: Sage, p.227-267, 1986.

FREDERIKSEN, C. JUST, M. & CARPENTER, P. A Capacity Theory of Compreension: Individual Diferences in Working Memory. *Psychology Review* 99 Nº 1, pp 122-149 1990.

FREDERIKEN, C., BRACEWELL, B. A. & RENAUD A. Psycognitive Representation and Processing of Discourse: Function and Dysfunction. *In*: Joanette Y.& Brownell H. (org.). *Discourse Hability and Brain Demade: Theoretical and Empirical Perspective*. N. Y. Springer Verlag. (1990).

FUHR, N. "Models for Retrieval with Probabilistic Indexing." Information Processing and Management, 25(1), 55-72. 1989.

GATES, B., MYHRVOLD, N. & RINEARSON, P. *The Road Ahead*. London, Viking. 1995.

GARFIELD, E. A Retrospective and Prospective View of Information Retrieval and Artificial Intelligence in the 21st Century. Journal of The American Society for Information Science and Technology. 52(1), 2001.

GERDA R., `Automatic detection of thesaurus relations for information retrieval applications', in Foundations of Computer Science: Potential - Theory - Cognition, pp. 499-506, (1997).

GERNSBACHER, M. A. (ed.) *Handbook or Psycholinguistics*. San Diego, New York: Academic Press, 1994.

GEY, F. Models in Information Retrieval. Folders of Tutorial Presented at at the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR), 1992.

GLOVER, E. J.; TSIOUTSIOULIKLIS, K.; LAWRENCE, S.; PENNOCK, D.M.; FLAKE, G.W. Using Web Structure for Classifying and Describing Web Pages. Proceedings of WWW-02, International Conference on the World Wide Web. 2002.

GOMEZ-PÉREZ, A. & BENJAMINS, V.R. Overview of Knowledge sharing and reuse components: Ontologies and problem-solving methods. In: International Joint Conference on Artificial Intelligence (IJCAI-99), Workshop on Ontologies and Problem-Solving Methods (KRR5), V.R. Benjamins, et al., Editors. Stockolm,1999.

GOOGLE, www.google.com. Última visita em fevereiro de 2005.

GOTTSCHALG-DUQUE, C. A Leitura em Ambiente Multimidia: A Produção de Inferências por parte do Leitor a partir da Compreensão de Hipertextos (Master Thesis). Programa de Pós-Graduação em Estudos Lingüísticos da FALE-UFMG. 16/11/1998. 1998.

GOTTSCHALG-DUQUE, C. E DILLINGEr, M. Características do Verbo realmente determinam as Relações Gramaticais? Palavras Inventadas X Palavras Normais. Resumos da II Semana de Iniciação Cientifica da UFMG. UFMG, 1994.

GREFENSTETTE, G. Explorations in Automatic Thesaurus Discovery. USA: Kluwer Academic Publishers, 1994.

_____ Evaluations Technique for automatic semantic extraction: comparing syntatic and window based approaches. In: BOGURAEV Branimir, PUSTEJOVSKY, James (Eds.). Corpus processing for lexical acquisition. MIT Press, 1995.

GRÉGOIRE, R. inc.; BRACEWELL, Robert; and LAFERRIÈRE, Thérèse.
SCHOOLNET/RESCOL. The Contribuition of New Technologies to Learning and Teaching in Elementary and Secondary Schools.Technical Report McGill University. August 1st, 1996.

GREISDORF, H. "Relevance: An Interdisciplinary and Information Science Perspective," Informing Science (3:2), 2000, pp. 67-71.
http://citeseer.ist.psu.edu/greisdorf00relevance.html

GRUBER; T. R. A Translation approach to portable ontology specifications. Knowledge Acquisition, 1993.

GUARINO, N., Understanding, Building, and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development",by van Heijst, Schreiber, and Wielinga. International Journal of Human and Computer Studies, 1997. 46: p. 293-310.

GUDIVADA, V. N.; RAGHAVAN, V. V.; GROSBY, W. I.; KASANAGOTTU, R. Information Retrieval on the World Wide Web. IEEE Internet Computer, 1997.

GUINCHAT, C. & MENOU, M. Introdução geral às ciências e técnicas da informação e documentação. 2 ed. Brasília: MCT/CNPq/IBICT, 1994.

HAGEGÉ C. SMORPH: um analisador/gerador morfológico para o português. Workshop sobre taggers para o português. Lisboa, Portugal. 1997.

HALLIDAY, M. A. & HASAN, R. *Cohesion in English*. London: Longman 1976.

HEIDEGGER, M. Logik. Die Frage nach der Wahrheit. Semesterkurs 1925/26, Marburg, 1925.

HIEMSTRA, D. Using Language Models for Information Retrieval. Phd Thesis University of Twente, Enschede, 2001.

HONKELA et al. Exploration of Full-Text Databases with Self-Organizing Maps. Proceedings of the ICNN96, International Conference on Neural Networks. 1996.

HOUDE, S. & HILL, C. What do Prototypes Prototype? In: HELANDER, M. LANDAVER, T. K. And PRABHU, P.(eds.) Handbook of Human-Computer Interaction. Second Edition, Elsevier Science B.V. 1997.

HOTHO, A.; STAAB, S. & MAEDCHE, A. Ontology-based text clustering. In Proceedings of the IJCAI- 2001 Workshop "Text Learning: Beyond Supervision", August, Seattle, USA. 2001.

HUANG, L. A Survey on Web Information Retrieval Technologies , Technical Report. Feb 2000. http://www.ecsl.cs.sunysb.edu/tech_reports.html

INCLUSÃO DIGITAL, http://www.idbrasil.gov.br/ última visita em 23de junho de 2005.

JACKENDOFF, R. *Semantic Structures*. Cambridge: MIT Press, 1990.

_____. Consciousness and the Computational Mind. Bradford Book. The MIT Press, Cambridge, Massachusetts, London, England, 1994.

JUST, A. M. & CARPENTER, P. A. A Capacity theory of Comprehension: Individual Differences in Working Memory. *Psychology Review*, 99 (1), 1992, pp 122-149.

KINTSCH, W. & van DIJK, T. A. Toward a model of text comprehension and production. *Psychological Review,* 85 (5), 1978. pp 363-394.

_____ *Strategies of Discourse Compreension*. N. Y. Academic Press. 1993.

KLEINBERG, J. M. Authoritative Sources in a Hyperlinked Environment. Proceedings of ACM-SIAM, 1998.

KOCH, I. G. V. A coesão textual. ed. 1, Contexto, 1989.
_____. O Texto e a Construção dos Sentidos. Contexto, 1997.
KOCH, I. G. V.; MONTEIRO, K.; Tópicos em Lingüística Textual e Análise da Conversação. ed. 1, EDUFURN, Vol. 1, 1998

KOSTER, M., "A Standart for Robot Exclusion", 1995. http://info.webcrawler.com/mak/projects/robots/robots.html

KURAMOTO, H. Proposition d'un système de recherche d'information assistée par ordinateur: avec application au portugais.. Thèse (Doctorat en Sciences de l'information et de la communication) - Université Lumière–Lyon 2, Lyon, França, 1999.

LAENDER, A. H. F., RIBEIRO-NETO, B., DA SILVA, A. S. and TEIXEIRA, J S. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record*, 2002.

LAHTINEN, T. Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods. Academic Dissertation. University of Helsinki. 2000.

LANCASTER, F. W. Information Retrieval Systems: Characteristics, testing and evaluation. 2nd. New York: Wiley Interscience, 1979.

LANCASTER, F. W. and WARNER, A. J. Information Retrieval Today. Arlington: Information Resources Press, 1993.

LAROCCA NETO, J.; SANTOS, A. D.; KAESTNER, A. A.; FREITAS, A. A. Generating Text Summaries through the Relative Importance of Topics. In M.C. Monard And J.S. Sichman (eds.), Lecture Notes in Artificial Intelligence, No. 1952, pp 300-309.Spring-Verlag. 2000.

LAWRENCE S. "Context in Web Search", IEEE Data Engineering Bulletin Vol 23 (3) pp. 25-32, 2000.

LAWRENCE, S.; GILES, C. L. and BOLLOCKER, K. Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Volume 32, Number 6, pp. 67-71, 1999.

LESK, M. The Seven Ages of Information Retrieval. Conference for the 50th anniversary of "As We May Think", 12-14 October, 1995. MIT Press, 1995.

LÈVY, P. As Tecnologias da Inteligência: O Futuro do Pensamento na Era da Informática. Rio de Janeiro: Ed 34. 1995.

LÈVY, P. A Inteligência Coletiva. Editora Loyola. 2000

LIU, B. & CHIN, C. W. Searching People on the Web According to Their Interest. poster of The 11th International. 2001.

LOBIN, H. Textauszeichnung und Dokumentgrammatiken. In: Texttechnologie.LOBIN, H & LEMNITZER, L. (ed.). Stauffenburg Verlag, 2003.

_____. Komplexität und Einfachheit in der Evolution von Dokumentgrammatiken. Zeitschrift für Literaturwissenschaft und Linguistik. Pp.106-122, September 2003.

LOSADA, D. E. & BARREIRO, A. Efficient algorithms for ranking documents represented as DNF formulas. In Proc. ACM SIGIR-2000 Workshop on Mathematical and Formal Methods in Information Retrieval. Pgs 16-24. Athens, Greece, July 2000.

LYMAN, P. and VARIAN, H. R. How Much Information. Technical Report. UCLA, Berkley. http://www.sims.berkley.edu/how-much-info/ (2000).

MAEDCHE, A. & STAAB, S. Mining Ontologies from Text. In: Dieng, R. & Corby, O. (Eds). EKAW-2000 - 12th International Conference on Knowledge Engineering and Knowledge Management. October 2-6, 2000, Juan-les-Pins, France. LNAI, Springer. 2000.

MARCUSCHI, L.A. *Lingüística de texto: o que é e como se faz*. Série Debates I, UFPE. 1983.

MEADOW, C. T. Text Information Retrieval Systems. San Diego Academic Press, 1992.

MEDEIROS, J.B. *Redação Científica; a prática de fichamentos, resumos, resenhas*. São Paulo: Atlas 1997.

MENDONÇA, E. S. A Lingüística e a Ciência da Informação: Estudos de uma interseção. Revista Ciência da Informação, ol. 29, n. 3 pag. 50-70. 2000

MENG, W., LIU, K-P., YU, C. T., WU, W., RISHE, N. Estimating the Usefulness of Search Engines. ICDE, 1999.

MIORELLI, S. T. ED-CER: Extração do Sintagma Nominal em Sentenças em Português. Dissertação de Mestrado. Programa de Pós Graduação em Ciência da Computação, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, 2001.

MITKOV, R (ed.). The Oxford Handbook of Computational Linguistics. Oxford University Press; 2003.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. American Documentation, V. 2, n. 1, p. 20-32, 1951.

NEGROPONTE, N. A Vida Digital. Companhia das Letras, 1995.

NUNES, M. G. V.; DA SILVA, B. C. D.; RINO, L. H. M. ; OLIVEIRA JR, O. N.; MARTINS R. T.; MONTILHA, G. Introdução ao Processamento de Linguagens Naturais. Notas Didáticas do ICMC, São Carlos, 1999.

NÜRNBERG, P. J.; FURUTA, R.; LEGGETT, J. J.; MARSHALL, C. C.; SHIPMAN III, F. M. Digital Libraries: Issues and Architectures. Digital Libraries 95. 1995.

OKSEFJELL, S. & SANTOS D. Breve panorâmica dos recursos de português mencionados na Web. In Anais do III Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'98). Porto Alegre, 3-4 novembro 1998. pp.38-47.

OLIVEIRA, M. de & ARAÚJO, E. A. A produção de conhecimentos e a origem das bibliotecas. Texto Didático. ECI-UFMG, Belo Horizonte, 2005.

OLIVEIRA, M. de; MOTA, F. R. L. & ALVARADO, R. U. COMUNIDADE CIENTÍFICA E CIENTIFICIDADE DA CIÊNCIA DA INFORMAÇÃO. 8. - Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, 3: 1º encontro internacional de bibliotecários de língua portuguesa, Lisboa, 28 de Fevereiro a 3 de Março de 1990

OLTMANS, J. A. E. A Knowledge-Based Approach to Robust Parsing. PhD Thesis.Centre for Telematics and Information Technology (CTIT) P.O. Box 217, 7500 AE Enschede, The Netherlands, 1999.

O'NEILL, E. T.; LAVOIE, B. F.; BENNETT, R. Trends in the Evolution of the Public Web, 1998-2002. D-Lib Magazine April, 2003.

PAGE, L.; BRIN, S.; MOTWANI, R. and WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. In Proceedings of the 7th International World Wide Web Conference, pages 161-172, Brisbane, Australia, 1998.

PAULO, J. L.; CORREIA, M.; MAMEDE, N. J.; HAGÈGE, C. Using Morphological, Syntactical and Statistical Information for Automatic Term Acquisition. In: Advances in Natural Language Processing. Third International Conference, Proceedings, PorTAL 2002, Faro Portugal, June 23-26, 2002.

PEREIRA, M. B.; SOUZA, C. F. R.; NUNES, M. G. V. 2002. Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. Revista Eletrônica de Iniciação Científica. SBC. Março de 2002. Ano II, Volume II, Número

PERSIN, M. Document filtering for fast ranking. Proc. ACM SIGIR, Conf. Dublin, Ireland, 1994.

PERSIN, M; ZOBEL, J.; SACKS-DAVIS, R. Filtered Document Retrieval with Frequency-Sorted Indexes. Journal of American Society of Information Science. V. 47, n. 10, 1996.

PINKERTON, B. Finding what people want: Experiences with the WebCrawler. In Proc. 2nd World Wide Web Conf., 1994.

PÔSSAS, B.; ZIVIANI, N.; MEIRA, W.; RIBEIRO-NETO, B. Set-Based Model: A New Approach for Information Retrieval Proc. 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02), Tampere, Finland, August 2002.

PRATT, W. & FAGAN, L. The Usefulness of Dynamically Categorizing Search Results. Journal of the American Medical Informatics Association. Volume 7. 2000. http://www1.ics.uci.edu/~pratt/main.html

RAKHSHAN, A., HOLDER, L. B., and COOK, D.J. Structural Web Search Engine. Proceedings of the Sixteenth International Conference of the Florida AI Research Society, May 2003.

RANCHHOD, E. (ed.). Tratamento das Línguas por Computador. Uma introdução à lingüística computacional e suas aplicações, Lisboa: Caminho, 2001.

RANGANATHAN, S.P. Prolegomena to library classification. 1ed. London: Asia Publishing House, 1933.

RANGANATHAN, S.P. Faceted analysis. In: CHAN, L. M. et al. (Eds.). Theory of subject analysis: a sourcebook. Littleton, CO: Libraries Unlimited, 1985.

REHM, G. Towards Automatic Web Genre Identification -- A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. Proceedings of the Hawai'i International Conference on System Sciences, January 7-10, 2002, Big Island, Hawaii.

RIJSBERGEN, C. J. van. Information Retrieval. 1979. Disponível em: "http://www.dcs.gla.ac.uk/Keith/Preface.html"

RILOFF, E. Little words can make a big difference for text classification. Proceedings of the 18th annual international ACM SIGIR Conference on research and development in information retrieval, 1995.

ROBERTSON, S.E. Theories and models in information retrieval. Journal of Documentation. 33, pags 126-148. 1977.

ROSENFELD, L. & MORVILLE, P. Information Architecture for the World Wide Web: Designing Large-Scale Web Sites. O'Reilly & Associates; 2nd edition. August 15, 2002.

ROWLEY, J. Organizing Knowledge: An Introduction to Information Retrieval. Vermont Gower Publishing, 2ed. 1996.

RUGE, G. `Automatic detection of thesaurus relations for information retrieval applications', in Foundations of Computer Science: Potential - Theory - Cognition, pp. 499-506, 1997.

SALTON, G. Recent studies in automatic text analysis and document retrieval, Journal of the ACM, v. 20, n. 2, 1973.

SALTON, G. & McGILL, M. J. Introduction to Modern Information Retrieval. McGraw Hill, 1983.

SANTOS, D. "Processamento de linguagem natural: apresentação através das aplicações", in Elisabete Ranchhod (ed.), Tratamento das Línguas por Computador. Uma introdução à linguística computacional e suas aplicações, Lisboa: Caminho, pp.229-259, 2001.

SANTOS, D. "Working with Portuguese corpora", presentation at the University of Oslo, 22 October 2004, extending a presentation at ISLA, in Lisbon 1 October 2004.

SCHRADER, A. M. Two domain of information science: problems in conceptualization and in consensus-building. Information Services & Uses, North-Holland, v.6, p. 169-205, 1986.

SHANNON, C. & WEAVER, W. A Mathematical theory of communication. Univ. of Illinois Press. 1948

SINGER, M. Discourse Inference processes. *In:* GERNSBACHER, M. A. (Ed.). *Handbook of Psycholinguistics*. Academic Press, 1994. chapter 14, p.479-509.

SMEATON, A. F. Information Retrieval: Still Butting Heads with Natural Language Processing? InInformation Extraction: A Multidisciplinary Approach to an Emerging Information Technology M.T. Pazienza (Ed.), Springer-Verlag Lecture Notes in Computer Science # 1299, pp 115-138. 1997.

_____. 1999. in: Natural Language Information Retrieval, T. Strzalkowski (Ed.), Kluwer Academic Publishers, pp.99-111.

SOWA, J. F. Building Sharing and Merging ontologies. Tutorial 1999. Disponivel em http://users.bestweb.net/~sowa/ontology/ontoshar.htm

SPERBER, D.; WILSON, D. Relevance, communication and cognition. 2. ed. Oxford: Blackwell, 1995.

STRZALKOWSKI, T.; KARLGREN, J.; PEREZ-CARBALLO, J.; TAPANAINEN, P. and TILL, N.``Natural Language Information Retrieval: TREC-7 Report". In D. K. Harman, editor, Proceedings of the 7th Text Retrieval Conference, Gaithersburg, Maryland, November. National Institute of Standards and Technology. 1998.

SU, L. T. Value of search results as a whole as the best measure of information retrieval performance. Information Processing and Management Vol.34, nº 5, 557-579. 1998.

SUMMERS, R.; OPPENHEIM, C.; MEADOWS, J.; McKNIGHT, C.; KINNELL, M. Information Science in 2010: A Loughborough University View. Journal of The American Society for Information Science. 50(12) 1153-1162. 1999.

SVENONIUS, E. Ranganathan and classification science. Libri, Conpenhagen, v. 42, n. 3, p. 176-183. 1992.

TAKAHASHI, T. (org.). Sociedade da Informação no Brasil Livro Verde. Brasília Ministério da Ciência e Tecnologia, 2000.

TODIRASCU, A. 2001. Ontologies for Information Retrieval. TALN 2001. W3C. http://w3c.org. 1998.

TODOBR, 2004. www.todobr.com.br.

VOUTILAINEN, A. Nptool, a detector of English noun phrases. Disponível na Web em http://www.lingsoft.fi/doc/nptool/intro/. Visitado pela ultima vez em 04 de Abril de 2005.

WINOGRAD, T. Understanding Natural Language, (191 pp.) New York: Academic Press, 1972.

WEAVER, W. 'Translation' (1949). Repr. In: LOCKE, W. N. and BOOTH, A.D. (eds.) Machine Translation of Languages: fourteen essays, pag. 15-23, Cambridge, Mass.: Technnology Press of the Massachusetts Institute of Technology, 1955.

Web Server Survey. news.netcraft.com. Última visita em fevereiro de 2005.

WITTEN, I. H.; MOFFAT, A. & BELL, T. C. Managing Gigabytes. Morgan Kaufmann Publishers, Inc. Second Edition, 1999

WOLFRAM, D. Applications of Informetrics to Information Retrieval Research. Informing Science, Special Issue on Information Science research, Vol. 3, nº 2. 2000.

W3C World Wide Web Consortium. 2005. http://www.w3.org/

WU, M.-M. & SONNENWALD, D. H. Reflections on Information Retrieval Evaluation. 1999. http://pnclink.org/events-report/1999/Proceedings/wu-mm.pdf

YAHOO, www.yahoo.com.br.

ZHOU, Lina and ZHANG, Dongsong. 2003. NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval. Journal of the American Society for Information Science and Technology, 54(2):115-123.

ZIVIANI, N.; RIBEIRO-NETO, B.; LAENDER, A. H. F.; SILVA, A. S.; VELOSO, E. A.; GOLGHER, P. Cob-Web – A Crawler for the Brazilian Web. Sixth Symposium on String Processing and Information Retrieval (spire´ 99), Cancun, Mexico, 1999.

Anexos

Anexo 1
Textos Originais
(texto de exemplo 1)

ARTIGOS

Biblioteca Digital: a experiência do Impa

Maria Cecilia Pragana Chataignier
Bacharel em Arquitetura – UFRJ. Especialização em Análise de Sistemas – UERJ. Trabalha no Impa como bolsista do CNPq, na informatização da biblioteca.
E-mail: cecilia@impa.br

Margareth Prevot da Silva
Bacharel em Informática pela UFRJ. Mestre em Informática pela PUC-Rio, na área de Banco de Dados. Trabalha no Impa como bolsista pelo CNPq, na implantação do servidor de *preprints*.
E-mail: margot@impa.br

Resumo

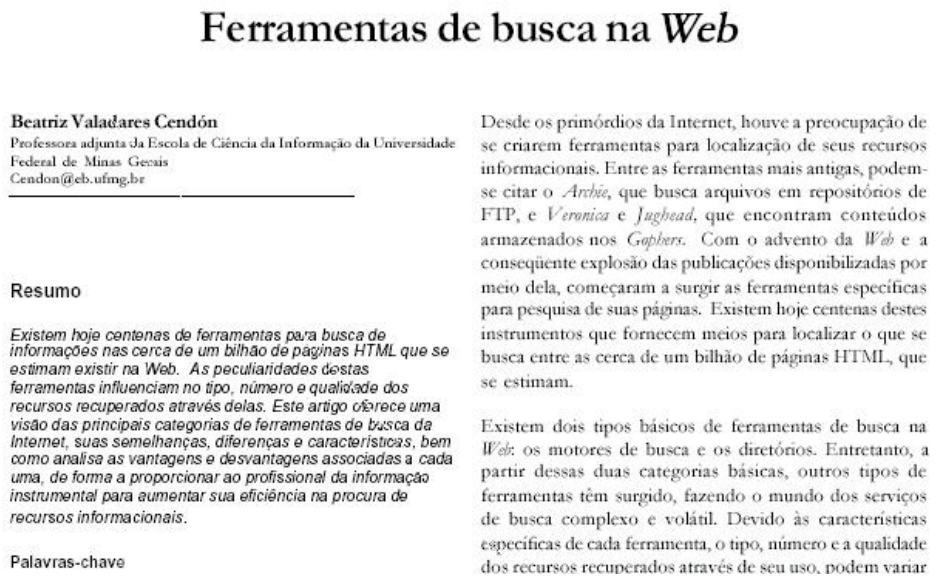
Relato da experiência do Impa na informatização de sua biblioteca, utilizando o software Horizon, e na construção de um servidor de preprints (dissertações de mestrado, teses de

INTRODUÇÃO

As bibliotecas são, com freqüência, descritas como o coração das universidades e institutos de pesquisas. Um bom acervo bibliográfico, em geral, reflete o nível da pesquisa que o instituto ou universidade desenvolve. A eficiência com a qual esse acervo é acessado e utilizado pelos leitores é de grande importância para o crescimento científico e tecnológico desses institutos e universidades e conseqüentemente do nosso país.

A Biblioteca do Instituto de Matemática Pura e Aplicada (Impa) possui notável acervo de cerca de 30 mil volumes em livros e 32 mil volumes em periódicos, correspondentes a mais de 600 títulos em assinaturas correntes, em matemática e áreas mais afins, tendo sido escolhida como biblioteca-base na área de matemática

(texto de exemplo 2)



Ferramentas de busca na *Web*

Beatriz Valadares Cendón
Professora adjunta da Escola de Ciência da Informação da Universidade Federal de Minas Gerais
Cendon@eb.ufmg.br

Resumo

Existem hoje centenas de ferramentas para busca de informações nas cerca de um bilhão de páginas HTML que se estimam existir na Web. As peculiaridades destas ferramentas influenciam no tipo, número e qualidade dos recursos recuperados através delas. Este artigo oferece uma visão das principais categorias de ferramentas de busca da Internet, suas semelhanças, diferenças e características, bem como analisa as vantagens e desvantagens associadas a cada uma, de forma a proporcionar ao profissional da informação instrumental para aumentar sua eficiência na procura de recursos informacionais.

Palavras-chave

Desde os primórdios da Internet, houve a preocupação de se criarem ferramentas para localização de seus recursos informacionais. Entre as ferramentas mais antigas, podem-se citar o *Archie*, que busca arquivos em repositórios de FTP, e *Veronica* e *Jughead*, que encontram conteúdos armazenados nos *Gophers*. Com o advento da *Web* e a conseqüente explosão das publicações disponibilizadas por meio dela, começaram a surgir as ferramentas específicas para pesquisa de suas páginas. Existem hoje centenas destes instrumentos que fornecem meios para localizar o que se busca entre as cerca de um bilhão de páginas HTML, que se estimam.

Existem dois tipos básicos de ferramentas de busca na *Web*: os motores de busca e os diretórios. Entretanto, a partir dessas duas categorias básicas, outros tipos de ferramentas têm surgido, fazendo o mundo dos serviços de busca complexo e volátil. Devido às características específicas de cada ferramenta, o tipo, número e a qualidade dos recursos recuperados através de seu uso, podem variar

Anexo 2
Textos da Coleção (Corpus)

(texto de exemplo 1)

Biblioteca Digital: a experiência do Impa

Maria Cecilia Pragana Chataignier
Bacharel em Arquitetura – UFRJ. Especialização em Análise de Sistemas – UERJ. Trabalha no Impa como bolsista do CNPq, na informatização da biblioteca.
E-mail: cecilia@impa.br
Margareth Prevot da Silva
Bacharel em Informática pela UFRJ. Mestre em Informática pela PUC-Rio, na área de Banco de Dados. Trabalha no Impa como bolsista pelo CNPq, na implantação do servidor de preprints.
E-mail: margot@impa.br

Palavras-chave
Biblioteca digital; Horizon; Impa; Automação; Preprint; Metadados; XML; Dublin core; Math-net; MPRESS; Harvest.

As bibliotecas são, com freqüência, descritas como o coração das universidades e institutos de pesquisas. Um bom acervo bibliográfico, em geral, reflete o nível da pesquisa que o instituto ou universidade desenvolve.
A eficiência com a qual esse acervo é acessado e utilizado pelos leitores é de grande importância para o crescimento científico e tecnológico desses institutos e universidades e conseqüentemente do nosso país.
A Biblioteca do Instituto de Matemática Pura e Aplicada (Impa) possui notável acervo de cerca de 30 mil volumes em livros e 32 mil volumes em periódicos, correspondentes a mais de 600 títulos em assinaturas correntes, em matemática e áreas mais afins, tendo sido escolhida como biblioteca-base na área de matemática pura e aplicada*.
Além desse acervo, a biblioteca possui um volume considerável de dissertações de mestrado e teses de doutorado (tanto publicações do próprio instituto, como de outras instituições), apostilas e atas de colóquios, bem como outras pré-publicações.

Anexo 3
Textos da Coleção (Corpus)

(texto de exemplo 2)

Ferramentas de busca na Web
Beatriz Valadares Cendón
Professora adjunta da Escola de Ciência da Informação da Universidade Federal de Minas Gerais
Cendon@eb.ufmg.br
Palavras-chave
Internet, Ferramentas de busca; Web; Diretórios; Motores de busca; Metamotores

Desde os primórdios da Internet, houve a preocupação de se criarem ferramentas para localização de seus recursos informacionais. Entre as ferramentas mais antigas, podemse citar o Archie, que busca arquivos em repositórios de FTP, e Veronica e Jughead, que encontram conteúdos armazenados nos Gophers. Com o advento da Web e a conseqüente explosão das publicações disponibilizadas por meio dela, começaram a surgir as ferramentas específicas para pesquisa de suas páginas. Existem hoje centenas destes instrumentos que fornecem meios para localizar o que se busca entre as cerca de um bilhão de páginas HTML, que se estimam.

Existem dois tipos básicos de ferramentas de busca na Web: os motores de busca e os diretórios. Entretanto, a partir dessas duas categorias básicas, outros tipos de ferramentas têm surgido, fazendo o mundo dos serviços de busca complexo e volátil. Devido às características específicas de cada ferramenta, o tipo, número e a qualidade dos recursos recuperados através de seu uso, podem variar enormemente. Para obter melhores resultados na busca de informações, o primeiro passo é entender as peculiaridades dos diferentes tipos de ferramentas de busca na Web. Este artigo oferece uma visão das principais categorias de ferramentas de busca da Internet, suas semelhanças, diferenças e características e analisa, também, as vantagens e desvantagens associadas a cada uma, de forma a proporcionar ao profissional da informação instrumental para aumentar sua eficiência na procura de recursos informacionais.

Anexo 4
Textos da Coleção processados sintaticamente

(texto de exemplo 1)

```
SOURCE: live
21. running text
A1
UTT:cl(fcl)
,
,
.
|-S:g(np)
| |-D:art('o' &lt;artd> F P)          As
| |-H:n('biblioteca' F P)   bibliotecas
|-P:v('ser' fin  PR 3P IND)          são
|-A:cl(acl)
|-A:g(pp)
| |-H:prp('com')  com
| |-D:n('freqüência' F S) freqüência
| |-H:prp('por' &lt;sam->)por
|-Cs:v('descrever' pcp  F P)        descritas
|-D:cl(acl)
  |-SUB:adv('como' &lt;rel>)        como
  |-SUB&lt;:g(np)
    |-D:art('o' &lt;artd> M S)        o
    |-H:n('coração' M S)   coração
    |-D:g(pp)
      |-H:prp('de' &lt;sam->)         de
      |-D:par
        |-CJT:g(np)
        | |-D:art('o' &lt;-sam> <artd> F P)    as
        | |-H:n('universidade' F P)   universidades
        |-CO:conj('e' &lt;co-prparg>)          e
        |-CJT:g(np)
          |-H:n('instituto' M P)         institutos
          |-D:g(pp)
            |-H:prp('de')       de
            |-D:n('pesquisa' F P)       pesquisas

SOURCE: live
22. running text
A1
UTT:cl(fcl)
=,
=,
.
|-S:par
  |-CJT:g(np)
  | |-D:art('um' &lt;arti> M S)       Um
  | |-D:adj('bom' M S)      bom
  | |-H:n('acervo' M S)     acervo
  | |-D:adj('bibliográfico' M S)      bibliográfico
  |-H:prp('com')  com
  |-A:g('em_geral' pp )     em_geral
  |-H:prp('por' &lt;sam->)por
  |-P:v('refletir' fin  PR 3S IND)    reflete
  |-Od:g(np)
    |-D:art('o' &lt;artd> M S)        o
    |-H:n('nível' M S)        nível
    |-D:g(pp)
```

```
      |-H:prp('de' &lt;sam->)          de
      |-D:g(np)
        |-D:art('o' &lt;-sam> <artd> F S)      a
        |-H:n('pesquisa' F S)          pesquisa
        |-D:cl(fcl)
        | |-Od:pron('que' indp  &lt;rel> M S) que
        | |-S:g(np)
        | | |-D:art('o' &lt;artd> M S)  o
        | | |-H:n('instituto' M S)       instituto
        | |-CO:conj('ou' &lt;co-subj>)          ou
        | |-CJT:n('universidade' F S)          universidade
        |-P:v('desenvolver' fin  PR 3S IND) desenvolve
          |-D:n('pesquisa' F P)       pesquisas

SOURCE: live
23. running text
A1
UTT:cl(fcl)
.
|-S:g(np)
| |-D:art('o' &lt;artd> F S)           A
| |-H:n('eficiência' F S)   eficiência
| |-D:cl(fcl)
| | |-A:g(pp)
| | | |-H:prp('com')          com
| | | |-D:pron('o_qual' indp  &lt;rel> F S)   a_qual
| | |-S:g(np)
| | | |-D:pron('esse' det  &lt;dem> M S)    esse
| | | |-H:n('acervo' M S)   acervo
| | |-P:vp
| | | |-D:v('ser' fin  PR 3S IND)    é
| | | |-H:v('acessar' pcp  M S)      acessado
| | |-CO:conj('e' &lt;co-pcv>)       e
| | |-P:v('utilizar' pcp  M S)          utilizado
| | |-A:par
| | | |-CJT:g(pp)
| | | |  |-H:prp('por' &lt;sam->)       por
| | | |  |-D:g(np)
| | | |    |-D:art('o' &lt;-sam> <artd> M P)   os
| | | |    |-H:n('leitor' M P)  leitores
| | |-P:v('ser' fin  &lt;ink> PR 3S IND)      é
| | |-Cs:g(pp)
| | | |-H:prp('de') de
| | | |-D:g(np)
| | | |  |-D:adj('grande' F S)          grande
| | | |  |-H:n('importância' F S)      importância
| | |-A:g(pp)
| | | |-H:prp('para')          para
| | | |-D:g(np)
| | | |  |-D:art('o' &lt;artd> M S)     o
| | | |  |-H:n('crescimento' M S)    crescimento
| | | |  |-D:par
| | | |  | |-CJT:adj('científico' M S)  científico
| | | |  | |-CO:conj('e' &lt;co-postnom>)      e
| | | |  | |-CJT:adj('tecnológico' M S)          tecnológico
| | | |  |-D:g(pp)
| | | |    |-H:prp('de' &lt;sam->)      de
| | | |    |-D:g(np)
| | | |      |-D:pron('esse' det  &lt;-sam> <dem> M P)          esses
| | | |      |-H:n('instituto' M P)       institutos
| | |-CO:conj('e' &lt;co-prparg>)  e
| | |-D:n('universidade' F P)        universidades
```

```
| |-CO:conj('e' &lt;co-advl>)        e
| |-CJT:adv('conseqüentemente' &lt;kc>)            conseqüentemente
|-Cs:g(pp)
  |-H:prp('de' &lt;sam->) de
  |-D:g(np)
    |-D:art('o' &lt;-sam> <artd> M S)        o
    |-D:pron('nosso' det  &lt;poss 1P> M S)          nosso
    |-H:n('país' M S)        país

SOURCE: live
24. running text
A1
UTT:cl(fcl)
((
))
,
|-S:g(np)
| |-D:art('o' &lt;artd> F S)           A
| |-H:n('biblioteca' &lt;prop> F S)Biblioteca
| |-D:g(pp)
|   |-H:prp('de' &lt;sam->)          de
|   |-D:par
|     |-CJT:g(np)
|     | |-D:art('o' &lt;-sam> <artd> M S)     o
|     | |-H:prop('Instituto_de_Matemática_Pura' M S)        Instituto_de_Matemática_Pura
|     |-CO:conj('e' &lt;co-prparg>)e
|     |-CJT:prop('Aplicada' M/F S)           Aplicada
|     |-D:v('ser' fin  PR 3S IND)    é
|-P:v('impar' fin  PR 3S IND)      Impa
|   |-CO:conj('e' &lt;co-pcv>)       e
|-P:v('possuir' fin  PR 3S IND)    possui
|-Od:g(np)
| |-D:adj('notável' M S)   notável
| |-H:n('acervo' M S)      acervo
| |-D:g(pp)
|   |-H:prp('de')   de
|   |-D:g(np)
|     |-D:g(ap)
|     | |-D:adv('cerca_de')cerca_de
|     | |-H:num('30' &lt;card> M P)          30
|     |-D:num('mil' &lt;card> M P) mil
|     |-H:n('volume' M P)  volumes
|-A:g(pp)
| |-H:prp('em')    em
| |-D:n('livro' M P)          livros
|-CO:conj('e' &lt;co-prparg>)      e
|-D:g(np)
| |-D:num('32' &lt;card> M P)      32
| |-D:num('mil' &lt;card> M P)     mil
| |-H:n('volume' M P)      volumes
|-A:g(pp)
| |-H:prp('em')    em
| |-D:adj('periódico' M P)           periódicos
|         |-H:prp('de' &lt;sam->)     de
|-D:n('correspondente' M/F P)    correspondentes
|-A:g(pp)
| |-H:prp('a_mais_de')    a_mais_de
| |-D:g(np)
|   |-D:num('600' &lt;card> M P)  600
|   |-H:n('título' M P)       títulos
|-A:g(pp)
| |-H:prp('em')    em
```

```
| |-D:g(np)
|   |-H:n('assinatura' F P)           assinaturas
|   |-D:adj('corrente' F P)correntes
|   |-D:pron('nosso' det  &lt;poss 1P> M S)          nosso
|-A:g(pp)
| |-H:prp('em')    em
| |-D:n('matemática' F S)           matemática
|-CO:conj('e' &lt;co-prparg>)      e
|-D:g(np)
  |-H:n('área' F P)          áreas
  |-D:g(ap)
    |-D:adv('muito' &lt;quant>)     mais
    |-H:adj('afim' M/F P)   afins
|-A:cl(icl)
  |-P:vp
  | |-D:v('ter' ger )           tendo
  | |-D:v('ser' pcp )          sido
  | |-H:v('escolher' pcp  F S)       escolhida
  |-SUB:adv('como' &lt;rel>)       como
  |-Od:n('base' F S)       biblioteca-base
  |-A:g(pp)
    |-H:prp('em' &lt;sam->)          em
    |-D:g(np)
      |-D:art('o' &lt;-sam> <artd> F S)       a
      |-H:n('área' F S)       área
      |-D:g(pp)
        |-H:prp('de')          de
        |-D:g(np)
          |-H:n('matemática' F S)    matemática
          |-D:par
            |-CJT:adj('puro' F S)       pura
            |-CO:conj('e' &lt;co-postnom>)    e
            |-CJT:v('aplicar' pcp  F S)          aplicada-

SOURCE: live
25. running text
A1
UTT:cl(fcl)
,
(
,
)
,
,
=bem_como       [bem_como] &lt;rel> ADV ù
.
|-A:g(advp)
| |-H:adv('além') Além
| |-D:g(pp)
|   |-H:prp('de' &lt;sam->)          de
|   |-D:g(np)
|     |-D:pron('esse' det  &lt;-sam> <dem> M S)   esse
|     |-H:n('acervo' M S)   acervo
|       |-D:art('o' &lt;-sam> <artd> M S)     o
|-S:g(np)
| |-D:art('o' &lt;artd> F S)          a
| |-H:n('biblioteca' F S)   biblioteca
|-P:v('possuir' fin  PR 3S IND)    possui
|-Od:g(np)
| |-D:art('um' &lt;arti> M S)         um
| |-H:n('volume' M S)      volume
| |-D:adj('considerável' M S)       considerável
```

```
| |-D:g(pp)
|   |-H:prp('de')   de
|   |-D:g(np)
|     |-H:n('dissertação' F P)        dissertações
|     |-D:g(pp)
|       |-H:prp('de')          de
|       |-D:par
|         |-CJT:n('mestrado' M S)   mestrado
|         |-CO:conj('e' &lt;co-prparg>)         e
|         |-CJT:g(np)
|           |-H:n('tese' F P) teses
|           |-D:g(pp)
|             |-H:prp('de')     de
|             |-D:n('doutorado' M S)  doutorado
|-D:g(np)
|-A:adv('tanto' &lt;quant>)          tanto
|-D:g(np)
| |-H:n('publicação' F P) publicações
| |-D:g(pp)
| | |-H:prp('de' &lt;sam->)          de
| | |-D:g(np)
| |   |-D:art('o' &lt;-sam> <artd> M S)       o
| |   |-D:pron('próprio' det  &lt;ident> M S)           próprio
| |   |-H:n('instituto' M S) instituto
| |-H:prp('a_mais_de')    a_mais_de
|-#FS-D:adv('como' &lt;rel> <ks>)          como
|-SUB&lt;:g(pp)
| |-H:prp('de')     de
| |-D:g(np)
|   |-D:pron('outro' det  &lt;diff> F P)        outras
|   |-H:n('instituição' F P)instituições
|   |-H:n('assinatura' F P)          assinaturas
|   |-D:adj('corrente' F P)correntes
|-D:n('apostila' F P)       apostilas
|-CO:conj('e' &lt;co-postnom>)   e
|-D:g(np)
| |-H:n('ata' F P) atas
| |-D:g(pp)
| | |-H:prp('de')   de
| | |-D:n('colóquio' M P)  colóquios
| |-D:g(ap)
|-A:cl(acl)
  | |-H:adj('afim' M/F P)   afins
  |-SUB&lt;:g(np)
  | |-D:pron('outro' det  &lt;diff> F P)        outras
  | |-H:n('publicação' F P)          pré-publicações
  | |-D:v('ter' ger )          tendo
```

Anexo 4
Textos da Coleção processados sintaticamente

(texto de exemplo 2)

```
SOURCE: live
5. running text
A1
UTT:cl(fcl)
,
.
|-D:prop('Metamotores_Desde' M/F S)   Metamotores_Desde
|-D:g(np)
| |-D:art('o' &lt;artd> M P)          os
| |-H:n('primórdio' M P)  primórdios
| |-D:g(pp)
|   |-H:prp('de' &lt;sam->)          de
|   |-D:g(np)
|     |-D:art('o' &lt;-sam> <artd> F S)       a
|     |-H:prop('Internet' F S)        Internet
|
|-P:v('haver' fin  &lt;ink> PS 1/3S IND)   houve
|-Od:g(np)
| |-D:art('o' &lt;artd> F S)          a
| |-H:n('preocupação' F S)        preocupação
|-D:prp('de')      de
|-Od:pron('se' pers  &lt;refl> M/F 3P ACC)         se
|-D:cl(icl)
  |-P:v('criar' inf  3P)      criarem
  |-Co:g(np)
    |-H:n('ferramenta' F P)          ferramentas
    |-D:g(pp)
      |-H:prp('para')         para
      |-D:g(np)
        |-H:n('localização' F S)      localização
        |-D:g(pp)
          |-H:prp('de')        de
          |-D:g(np)
            |-D:pron('seu' det  &lt;poss 3S> M P)      seus
            |-H:n('recurso' M P)       recursos
            |-D:adj('informacional' M P)       informacionais

SOURCE: live
6. running text
A1
A:g(pp)
,
=,
==,
==,
.
|-H:prp('entre')   Entre
|-D:g(np)
| |-D:art('o' &lt;artd> F P)          as
| |-H:n('ferramenta' F P) ferramentas
| |-D:g(ap)
|   |-D:adv('muito' &lt;quant>)     mais
|   |-H:adj('antigo' F P)   antigas
|     |-D:art('o' &lt;-sam> <artd> F S)       a
|-UTT:cl(icl)
|-S:n('podemse' M S)    podemse
```

```
|-P:v('citar' inf  3S)        citar
|-Od:g(np)
| |-D:art('o' &lt;artd> M S)          o
| |-H:prop('Archie' M S)  Archie
|-D:prp('de')       de
  |-D:cl(fcl)
    |-S:pron('que' indp  &lt;rel> M S)        que
    |-P:v('buscar' fin  PR 3S IND) busca
    |-Od:n('arquivo' M P) arquivos
    |-A:g(pp)
    | |-H:prp('em')           em
    | |-D:g(np)
    |   |-H:n('repositório' M P)       repositórios
    |   |-D:g(pp)
    |     |-H:prp('de')         de
    |     |-D:prop('FTP' M/F S)       FTP
    |     |-D:g(np)
    |-CO:conj('e' &lt;co-vfin>)       e
    |-Od:prop('Veronica' M/F S)   Veronica
    |-CO:conj('e' &lt;co-prparg>)  e
    |-Od:prop('Jughead' M/F S)    Jughead
  |-D:cl(fcl)
    |-S:pron('que' indp  &lt;rel> M S)        que
    |-P:v('encontrar' fin  PR 3P IND)         encontram
    |-Od:g(np)
    | |-H:n('conteúdo' M P)          conteúdos
    | |-D:v('armazenar' pcp  M P) armazenados
    |-A:g(pp)
      |-H:prp('em' &lt;sam->)         em
      |-D:g(np)
        |-D:art('o' &lt;-sam> <artd> M P)     os
        |-H:prop('Gophers' M P)     Gophers

SOURCE: live
7. running text
A1
UTT:cl(fcl)
====,
.
|-fCs:g(pp)
| |-H:prp('com')  Com
| |-D:g(np)
|   |-D:art('o' &lt;artd> M S)         o
|   |-H:n('advento' M S)   advento
|   |-D:g(pp)
|     |-H:prp('de' &lt;sam->)         de
|     |-D:g(np)
|       |-D:art('o' &lt;-sam> <artd> F S)      a
|       |-H:n('web' &lt;prop> F S)   Web
|-CO:conj('e')     e
|-D:g(np)
| |-D:art('o' &lt;artd> F S)          a
| |-D:adj('conseqüente' F S)       conseqüente
| |-H:n('explosão' F S)    explosão
| |-D:g(pp)
|   |-H:prp('de' &lt;sam->)          de
|   |-D:g(np)
|   | |-D:art('o' &lt;-sam> <artd> F P)       as
|   | |-H:n('publicação' F P)         publicações
|   | |-D:g(ap)
|   |   |-H:v('disponibilizar' pcp  F P)         disponibilizadas
|   |   |-D:g(pp)
```

```
|   |   | |-H:prp('por')        por
|   |   | |-D:adj('meio' M S)           meio
|   |   |-D:g(pp)
|   |   |   |-H:prp('de' &lt;sam->)     de
|   |   |   |-D:pron('ela' pers  &lt;-sam> F 3S NOM/PIV)       ela
|   |-Od:prop('Veronica' M/F S)   Veronica
|-P:vp
| |-D:v('começar' fin  PS/MQP 3P IND)   começaram
| |-SUB:prp('a')  a
| |-H:v('surgir' inf )          surgir
|-S:g(np)
| |-D:art('o' &lt;artd> F P)            as
| |-H:n('ferramenta' F P) ferramentas
| |-D:adj('específico' F P)             específicas
|-A:g(pp)
  |-H:prp('para')  para
  |-D:g(np)
    |-H:n('pesquisa' F S)  pesquisa
    |-D:g(pp)
      |-H:prp('de') de
      |-D:g(np)
        |-D:pron('seu' det  &lt;poss 3S> F P)          suas
        |-H:n('página' F P)  páginas

SOURCE: live
8. running text
A1
UTT:cl(fcl)
=======,
.
|-P:v('existir' fin  PR 3P IND)      Existem
|-A:adv('hoje')    hoje
|-S:g(np)
  |-H:n('centena' F P)     centenas
  |-D:g(pp)
  | |-H:prp('de' &lt;sam->)           de
  | |-D:g(np)
  | | |-D:pron('este' det  &lt;-sam> <dem> M P)    estes
  | | |-H:n('instrumento' M P)       instrumentos
  | | |-D:cl(fcl)
  | |   |-S:pron('que' indp  &lt;rel> M S)     que
  | |   |-P:v('fornecer' fin  PR 3P IND)       fornecem
  | |   |-Od:n('meio' M P)  meios
  | |   |-A:g(pp)
  | |     |-H:prp('para')       para
  | |     |-D:cl(icl)
  | |       |-P:v('localizar' inf )         localizar
  | |-Od:cl(fcl)
  |   |-Od:pron('o_que' indp  &lt;rel> M S)o_que
  |   |-S:pron('se' pers  M/F 3S/P ACC)    se
  |   |-P:v('buscar' fin  PR 3S IND)           busca
  |   |-A:g(pp)
  |     |-H:prp('entre')       entre
  |     |-D:g(np)
  |       |-D:art('o' &lt;artd> F P)    as
  |       |-D:g(ap)
  |       | |-D:adv('cerca_de')         cerca_de
  |       | |-H:num('um' &lt;card> M S)       um
  |       |-H:n('bilhão' M S)           bilhão
  |       |-D:g(pp)
  |         |-H:prp('de')       de
  |         |-D:g(np)
```

```
  |          |-H:n('página' F P)        páginas
  |          |-D:adj('html' &lt;prop> F S)      Html
  |-D:art('o' &lt;artd> F P)           as
  |          |-D:cl(fcl)
  |-Od:pron('que' indp  &lt;rel> M S)        que
  |-S:pron('se' pers  M/F 3S/P ACC)        se
  |-P:v('estimar' fin  PR 3P IND)  estimam
  |-D:g(np)

SOURCE: live
9. running text
A1
UTT:cl(fcl)
|-P:v('existir' fin  PR 3P IND)      Existem
|-S:g(np)
| |-D:num('dois' &lt;card> M P)   dois
| |-H:n('tipo' M P)           tipos
| |-D:adj('básico' M P)    básicos
| |-D:g(pp)
|   |-H:prp('de')   de
|   |-D:g(np)
|     |-H:n('ferramenta' F P)         ferramentas
|     |-D:g(pp)
|       |-H:prp('de')           de
|       |-D:n('busca' F S)   busca
|-A:g(pp)
| |-H:prp('em' &lt;sam->) em
| |-D:g(np)
|   |-D:art('o' &lt;-sam> <artd> F S)          a
|   |-H:n('web' &lt;prop> F S)       Web
|-:

SOURCE: live
10. running text
A1
PRED:g(np)
.
|-D:art('o' &lt;artd> M P) os
|-H:n('motor' M P)          motores
|-D:g(pp)
| |-H:prp('de')     de
| |-D:n('busca' F S)        busca
|-CO:conj('e')     e
|-PRED:g(np)
|-D:art('o' &lt;artd> M P) os
|-H:n('diretório' M P)      diretórios
|     |-D:g(pp)

SOURCE: live
11. running text
A1
UTT:cl(fcl)
,
==,
,
.
|-A:adv('entretanto' &lt;kc>)       Entretanto
|-H:n('motor' M P)          motores
|-A:g(pp)
| |-H:prp('a')       a
| |-D:cl(icl)
| | |-P:v('partir' inf )         partir
```

```
| | |-A:g(pp)
| | |-H:prp('de' &lt;sam->)          de
| | |-D:g(np)
| | |-D:pron('esse' det  &lt;-sam> <dem> F P)  essas
| | |-D:num('dois' &lt;card> F P)          duas
| | |-H:n('categoria' F P)          categorias
| | |-D:adj('básico' F P)          básicas
| |-H:prp('em' &lt;sam->)em
|-S:g(np)
| |-D:pron('outro' det  &lt;diff> M P)          outros
| |-H:n('tipo' M P)          tipos
| |-D:g(pp)
|   |-H:prp('de')   de
|   |-D:n('ferramenta' F P)          ferramentas
|-P:vp
| |-D:v('ter' fin  PR 3P IND)          têm
| |-H:v('surgir' pcp )          surgido
|       |-D:g(np)
|-A:cl(icl)
  |-P:v('fazer' ger )          fazendo
  |-Od:g(np)
    |-D:art('o' &lt;artd> M S)          o
    |-H:n('mundo' M S)    mundo
    |-D:g(pp)
      |-H:prp('de' &lt;sam->)          de
      |-D:g(np)
      | |-D:art('o' &lt;-sam> <artd> M P)     os
      | |-H:n('serviço' M P)          serviços
      | |-D:g(pp)
      |   |-H:prp('de')          de
      |   |-D:g(np)
      |     |-H:n('busca' F S)          busca
      |     |-D:par
      |       |-CJT:adj('complexo' M S)          complexo
      |       |-CO:conj('e' &lt;co-postnom>)  e
      |       |-CJT:adj('volátil' M/F S)          volátil
      |-H:prp('de') de

SOURCE: live
12. running text
A1
UTT:cl(fcl)
,
=,
,
.
|-A:g(pp)
| |-H:prp('devido_a' &lt;sam->)   Devido_a
| |-D:g(np)
|   |-D:art('o' &lt;artd> <-sam> F P)          as
|   |-H:n('característica' F P)          características
|   |-D:adj('específico' F P)          específicas
|   |-D:g(pp)
|     |-H:prp('de') de
|     |-D:g(np)
|       |-D:pron('cada' det  &lt;quant> F S)cada
|       |-H:n('ferramenta' F S)          ferramenta
|       |-H:n('categoria' F P)          categorias
|-S:par
| |-CJT:g(np)
| | |-D:art('o' &lt;artd> M S)          o
| | |-H:n('tipo' M S)          tipo
```

```
| |-H:n('tipo' M P)           tipos
| |-CJT:n('número' M S) número
| |-CO:conj('e' &lt;co-subj>)        e
| |-CJT:g(np)
|   |-D:art('o' &lt;artd> F S)          a
|   |-H:n('qualidade' F S) qualidade
|   |-D:g(pp)
|     |-H:prp('de' &lt;sam->)         de
|     |-D:g(np)
|       |-D:art('o' &lt;-sam> <artd> M P)     os
|       |-H:n('recurso' M P)           recursos
|       |-D:v('recuperar' pcp  M P) recuperados
|-A:g(advp)
| |-H:adv('através')         através
| |-D:g(pp)
|   |-H:prp('de')   de
|   |-D:g(np)
|     |-D:pron('seu' det  &lt;poss 3S> M S)           seu
|     |-H:n('uso' M S)       uso
|         |-H:prp('de')         de
|-P:vp
| |-D:v('poder' fin  PR 3P IND)    podem
| |-H:v('variar' inf )          variar
|-A:adv('enorme')          enormemente
      |       |-CO:conj('e' &lt;co-postnom>)  e

SOURCE: live
13. running text
A1
UTT:cl(fcl)
==,
.
|-A:g(pp)
| |-H:prp('para')  Para
| |-D:cl(icl)
|   |-P:v('obter' inf )         obter
|   |-Od:g(np)
|   | |-D:adj('bom' M P)    melhores
|   | |-H:n('resultado' M P)          resultados
|   |-A:g(pp)
|   | |-H:prp('em' &lt;sam->)         em
|   | |-D:g(np)
|   |   |-D:art('o' &lt;-sam> <artd> F S)      a
|   |   |-H:n('busca' F S)   busca
|   |   |-D:g(pp)
|   |     |-H:prp('de')         de
|   |     |-D:n('informação' F P)     informações
|   |-H:n('tipo' M S)         tipo
|-S:g(np)
| |-D:art('o' &lt;artd> M S)          o
| |-D:adj('primeiro' &lt;NUM-ord> M S)    primeiro
| |-H:n('passo' M S)        passo
|-P:v('ser' fin  PR 3S IND)          é
|-Cs:cl(icl)
  |-P:v('entender' inf )     entender
  |-Od:g(np)
  | |-D:art('o' &lt;artd> F P)          as
  | |-H:n('peculiaridade' F P)        peculiaridades
  | |-D:g(pp)
  |   |-H:prp('de' &lt;sam->)         de
  |   |-D:g(np)
  |     |-D:art('o' &lt;-sam> <artd> M P)     os
```

```
  |     |-D:pron('diferentes' det  &lt;quant> M P)    diferentes
  |     |-H:n('tipo' M P)      tipos
  |     |-D:g(pp)
  |       |-H:prp('de')        de
  |       |-D:g(np)
  |         |-H:n('ferramenta' F P)   ferramentas
  |         |-D:g(pp)
  |           |-H:prp('de')     de
  |           |-D:n('busca' F S)        busca
  |-A:g(pp)
    |-H:prp('em' &lt;sam->)          em
    |-D:g(np)
      |-D:art('o' &lt;-sam> <artd> F S)       a
      |-H:n('web' &lt;prop> F S)    Web
        |-D:pron('seu' det  &lt;poss 3S> F P)          suas

SOURCE: live
14. running text
A1
UTT:par
==========,
==,
==,
==,
====,
.
|-CJT:cl(fcl)
  |-S:g(np)
  | |-D:pron('este' det  &lt;dem> M S)      Este
  | |-H:n('artigo' M S)      artigo
  |-P:v('oferecer' fin  PR 3S IND)oferece
  |-Od:par
  | |-CJT:g(np)
  | | |-D:art('um' &lt;arti> F S)      uma
  | | |-H:n('visão' F S)     visão
  | | |-D:g(pp)
  | | |   |-H:prp('de' &lt;sam->)       de
  | | |   |-D:par
  | | |   | |-CJT:g(np)
  | | |   |   |-D:art('o' &lt;-sam> <artd> F P)  as
  | | |   |   |-D:adj('principal' F P)     principais
  | | |   |   |-H:n('categoria' F P)      categorias
  | | |   |   |-D:g(pp)
  | | |   |     |-H:prp('de')      de
  | | |   |     |-D:g(np)
  | | |   |       |-H:n('ferramenta' F P)           ferramentas
  | | |   |       |-D:g(pp)
  | | |   |         |-H:prp('de')  de
  | | |   |         |-D:g(np)
  | | |   |           |-H:n('busca' F S)   busca
|-| | |   |           |-D:g(pp)
|-| | |   |             |-H:prp('de' &lt;sam->)      de
  |-| |   |             |-D:g(np)
  |-| |   |               |-D:art('o' &lt;-sam> <artd> F S)  a
    |- |   |               |-H:prop('Internet' F S)    Internet
    |-  |-D:art('o' &lt;-sam> <artd> M P)    os
  | |     |           |-CJT:g(np)
  |-|     |             |-D:pron('seu' det  &lt;poss 3S> F P)       suas
    |-    |             |-H:n('semelhança' F P)  semelhanças
    |-    |-H:prp('de')         de
  | |-Od:n('diferença' F P)           diferenças
  | |-CO:conj('e' &lt;co-prparg>)   e
```

```
  | |-Od:n('característica' F P)     características
  |-CO:conj('e' &lt;co-vfin> <co-fmc>)     e
  |-CJT:cl(fcl)
    |-P:v('analisar' fin  PR 3S IND)          analisa
    |-H:prp('em' &lt;sam->)          em
    |-A:adv('também')     também
    | |-D:art('o' &lt;-sam> <artd> F S)       a
    |-Od:g(np)
    | |-D:art('o' &lt;artd> F P)       as
    | |-H:n('vantagem' F P)          vantagens
    |-CO:conj('e' &lt;co-acc>)       e
    |-CJT:g(np)
      |-H:n('desvantagem' F P)     desvantagens
      |-D:g(ap)
        |-H:v('associar' pcp  F P)    associadas
        |-D:g(pp)
          |-H:prp('a')          a
          |-D:g(np)
            |-D:pron('cada' det  &lt;quant> M/F S)     cada
            |-H:num('um' &lt;card> F S)        uma
        |-A:g(pp)
          |-H:prp('de_forma_a')       de_forma_a
          |-D:cl(icl)
            |-P:v('proporcionar' inf )  proporcionar
            |-Od:g(pp)
              |-H:prp('a' &lt;sam->)    a
              |-D:g(np)
                |-D:art('o' &lt;-sam> <artd> M S)         o
                |-H:n('profissional' M S)          profissional
                |-D:g(pp)
                  |-H:prp('de' &lt;sam->)          de
                  |-D:g(np)
                    |-D:art('o' &lt;-sam> <artd> F S)      a
|-                  |-H:n('informação' F S)       informação
|-                  |-D:adj('instrumental' F S)   instrumental
|-                  |-D:g(pp)
|-                    |-H:prp('para')       para
  |-                  |-D:cl(icl)
  |-                    |-P:v('aumentar' inf )       aumentar
    |-                  |-Od:g(np)
    |-                    |-D:pron('seu' det  &lt;poss 3S> <si> F S)          sua
      |-                  |-H:n('eficiência' F S)     eficiência
      |-                  |-D:g(pp)
      |-                    |-H:prp('em' &lt;sam->)           em
        |-                  |-D:g(np)
        |-                    |-D:art('o' &lt;-sam> <artd> F S)         a
          |-                  |-H:n('procura' F S)   procura
          |-                  |-D:g(pp)
          |-                    |-H:prp('de')de
            |-                  |-D:g(np)
            |-                    |-H:n('recurso' M P)          recursos
              |-                  |-D:adj('informacional' M P)          informacionais
```

Anexo 5
Extração de Sintagmas Nominais

(texto de exemplo 1)

Salvar NP's: sim

[Frase **21**]

1 - **S:g** (**np**)
2 - **D:art** (**'o' <artd> F P**) **As**
2 - **H:n** (**'biblioteca' F P**) **bibliotecas**
1 - **P:v** (**'ser' fin PR 3P IND**) **são**
1 - **A:cl** (**acl**)
1 - **A:g** (**pp**)
2 - **H:prp** (**'com'**) **com**
2 - **D:n** (**'freqüência' F S**) **freqüência**
2 - **H:prp** (**'por' <sam->**) **por**
1 - **Cs:v** (**'descrever' pcp F P**) **descritas**
1 - **D:cl** (**acl**)

Frase: As bibliotecas com freqüência por

NP's:
1 - As bibliotecas ...arquivado!

[Frase **22**]

1 - **S:par** ()

Frase:

NP's:
(nenhum)

[Frase **23**]

1 - **S:g** (**np**)
2 - **D:art** (**'o' <artd> F S**) **A**
2 - **H:n** (**'eficiência' F S**) **eficiência**
2 - **D:cl** (**fcl**)
3 - **A:g** (**pp**)
4 - **H:prp** (**'com'**) **com**
4 - **D:pron** (**'o_qual' indp <rel> F S**) **a_qual**
3 - **S:g** (**np**)
4 - **D:pron** (**'esse' det <dem> M S**) **esse**
4 - **H:n** (**'acervo' M S**) **acervo**
3 - **P:vp** ()
4 - **D:v** (**'ser' fin PR 3S IND**) **é**
4 - **H:v** (**'acessar' pcp M S**) **acessado**
3 - **CO:conj** (**'e' <co-pcv>**) **e**

3 - P:v ('utilizar' pcp M S) utilizado
3 - A:par ()
4 - CJT:g (pp)
4 - H:prp ('por' <sam->) por
4 - D:g (np)
4 - D:art ('o' <-sam> <artd> M P) os
4 - H:n ('leitor' M P) leitores
3 - P:v ('ser' fin <ink> PR 3S IND) é
3 - Cs:g (pp)
4 - H:prp ('de') de
4 - D:g (np)
4 - D:adj ('grande' F S) grande
4 - H:n ('importância' F S) importância
3 - A:g (pp)
4 - H:prp ('para') para
4 - D:g (np)
4 - D:art ('o' <artd> M S) o
4 - H:n ('crescimento' M S) crescimento
4 - D:par ()
5 - CJT:adj ('científico' M S) científico
5 - CO:conj ('e' <co-postnom>) e
5 - CJT:adj ('tecnológico' M S) tecnológico
4 - D:g (pp)
4 - H:prp ('de' <sam->) de
4 - D:g (np)
4 - D:pron ('esse' det <-sam> <dem> M P) esses
4 - H:n ('instituto' M P) institutos
3 - CO:conj ('e' <co-prparg>) e
3 - D:n ('universidade' F P) universidades
2 - CO:conj ('e' <co-advl>) e
2 - CJT:adv ('conseqüentemente' <kc>) conseqüentemente
1 - Cs:g (pp)

Frase: A eficiência com a_qual esse acervo é acessado e utilizado por os leitores é de grande importância para o crescimento científico e tecnológico de esses institutos e universidades e conseqüentemente

NP's:
1 - A eficiência com a_qual esse acervo é acessado e utilizado por os leitores é de grande importância para o crescimento científico e tecnológico de esses institutos e universidades e conseqüentemente ...arquivado!

[Frase **24**]

1 - S:g (np)
2 - D:art ('o' <artd> F S) A
2 - H:n ('biblioteca' <prop> F S) Biblioteca
2 - D:g (pp)
2 - H:prp ('de' <sam->) de
2 - D:par ()
2 - CJT:g (np)

3 - D:art ('o' <-sam> <artd> M S) o
3 - H:prop ('Instituto_de_Matemática_Pura' M S) Instituto_de_Matemática_Pura
2 - CO:conj ('e' <co-prparg>) e
2 - CJT:prop ('Aplicada' M/F S) Aplicada
2 - D:v ('ser' fin PR 3S IND) é
1 - P:v ('impar' fin PR 3S IND) Impa
2 - CO:conj ('e' <co-pcv>) e
1 - P:v ('possuir' fin PR 3S IND) possui
1 - Od:g (np)
2 - D:adj ('notável' M S) notável
2 - H:n ('acervo' M S) acervo
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:g (np)
2 - D:g (ap)
3 - D:adv ('cerca_de') cerca_de
3 - H:num ('30' <card> M P) 30
2 - D:num ('mil' <card> M P) mil
2 - H:n ('volume' M P) volumes
1 - A:g (pp)
2 - H:prp ('em') em
2 - D:n ('livro' M P) livros
1 - CO:conj ('e' <co-prparg>) e
1 - D:g (np)
2 - D:num ('32' <card> M P) 32
2 - D:num ('mil' <card> M P) mil
2 - H:n ('volume' M P) volumes
1 - A:g (pp)
2 - H:prp ('em') em
2 - D:adj ('periódico' M P) periódicos
2 - H:prp ('de' <sam->) de
1 - D:n ('correspondente' M/F P) correspondentes
1 - A:g (pp)
2 - H:prp ('a_mais_de') a_mais_de
2 - D:g (np)
2 - D:num ('600' <card> M P) 600
2 - H:n ('título' M P) títulos
1 - A:g (pp)
2 - H:prp ('em') em
2 - D:g (np)
2 - H:n ('assinatura' F P) assinaturas
2 - D:adj ('corrente' F P) correntes
2 - D:pron ('nosso' det <poss 1P> M S) nosso
1 - A:g (pp)
2 - H:prp ('em') em
2 - D:n ('matemática' F S) matemática
1 - CO:conj ('e' <co-prparg>) e
1 - D:g (np)
1 - A:cl (icl)

Frase: A Biblioteca de o Instituto_de_Matemática_Pura e Aplicada é e notável acervo de

cerca_de 30 mil volumes em livros e 32 mil volumes em periódicos de correspondentes a_mais_de 600 títulos em assinaturas correntes nosso em matemática e

NP's:
1 - A Biblioteca de o Instituto_de_Matemática_Pura e Aplicada é ...arquivado!
2 - notável acervo de cerca_de 30 mil volumes em livros e ...arquivado!
3 - 32 mil volumes em periódicos de ...arquivado!
4 - correspondentes a_mais_de 600 títulos em assinaturas correntes nosso em matemática e ...arquivado!

[Frase **25**]

1 - A:g (advp)
2 - H:adv ('além') Além
2 - D:g (pp)
2 - H:prp ('de' <sam->) de
2 - D:g (np)
2 - D:pron ('esse' det <-sam> <dem> M S) esse
2 - H:n ('acervo' M S) acervo
2 - D:art ('o' <-sam> <artd> M S) o
1 - S:g (np)
2 - D:art ('o' <artd> F S) a
2 - H:n ('biblioteca' F S) biblioteca
1 - P:v ('possuir' fin PR 3S IND) possui
1 - Od:g (np)
2 - D:art ('um' <arti> M S) um
2 - H:n ('volume' M S) volume
2 - D:adj ('considerável' M S) considerável
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:g (np)
2 - H:n ('dissertação' F P) dissertações
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:par ()
2 - CJT:n ('mestrado' M S) mestrado
2 - CO:conj ('e' <co-prparg>) e
2 - CJT:g (np)
2 - H:n ('tese' F P) teses
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:n ('doutorado' M S) doutorado
1 - D:g (np)
1 - A:adv ('tanto' <quant>) tanto
1 - D:g (np)
2 - H:n ('publicação' F P) publicações
2 - D:g (pp)
3 - H:prp ('de' <sam->) de
3 - D:g (np)
3 - D:art ('o' <-sam> <artd> M S) o
3 - D:pron ('próprio' det <ident> M S) próprio

3 - H:n ('instituto' M S) instituto
2 - H:prp ('a_mais_de') a_mais_de
1 - #FSD:adv ('como' <rel> <ks>) como
1 - SUB<:g (pp)
2 - H:prp ('de') de
2 - D:g (np)
2 - D:pron ('outro' det <diff> F P) outras
2 - H:n ('instituição' F P) instituições
2 - H:n ('assinatura' F P) assinaturas
2 - D:adj ('corrente' F P) correntes
1 - D:n ('apostila' F P) apostilas
1 - CO:conj ('e' <co-postnom>) e
1 - D:g (np)
2 - H:n ('ata' F P) atas
2 - D:g (pp)
3 - H:prp ('de') de
3 - D:n ('colóquio' M P) colóquios
2 - D:g (ap)
1 - A:cl (acl)

Frase: Além de esse acervo o a biblioteca um volume considerável de dissertações de mestrado e teses de doutorado tanto publicações de o próprio instituto a_mais_de como de outras instituições assinaturas correntes apostilas e atas de colóquios

NP's:
1 - a biblioteca ...arquivado!
2 - um volume considerável de dissertações de mestrado e teses de doutorado ...arquivado!
3 - tanto ...arquivado!
4 - publicações de o próprio instituto a_mais_de como de outras instituições assinaturas correntes ...arquivado!
5 - apostilas e ...arquivado!
6 - atas de colóquios ...arquivado!

FIM

Anexo 5
Extração de Sintagmas Nominais

(texto de exemplo 2)

Salvar NP's: sim

[Frase **1**]

0 - **PRED:g** (**np**)
1 - **H:n** (**'ferramenta' F P**) **Ferramentas**
1 - **D:g** (**pp**)
2 - **H:prp** (**'de'**) **de**
2 - **D:g** (**np**)
2 - **H:n** (**'busca' F S**) **busca**
2 - **D:g** (**pp**)
2 - **H:prp** (**'em' <sam->**) **em**
2 - **D:g** (**np**)
2 - **D:art** (**'o' <-sam> <artd> F S**) **a**
2 - **H:prop** (**'Web_Beatriz_Valadares_Cendón_Professora' M S**) **Web_Beatriz_Valadares_Cendón_Professora**
2 - **D:adj** (**'adjunto' F S**) **adjunta**
2 - **D:g** (**pp**)
2 - **H:prp** (**'de' <sam->**) **de**
2 - **D:g** (**np**)
2 - **D:art** (**'o' <-sam> <artd> F S**) **a**
2 - **H:prop** (**'Escola_de_Ciência_da_Informação' F S**) **Escola_de_Ciência_da_Informação**
2 - **D:g** (**pp**)
2 - **H:prp** (**'de' <sam->**) **de**
2 - **D:g** (**np**)
2 - **D:art** (**'o' <-sam> <artd> F S**) **a**
2 - **H:prop** (**'Universidade_Federal' F S**) **Universidade_Federal**
1 - **D:prp** (**'de'**) **de**
1 - **D:g** (**np**)
1 - **H:n** (**'ferramenta' F P**) **Ferramentas**
1 - **D:g** (**pp**)

Frase: Ferramentas de busca em a Web_Beatriz_Valadares_Cendón_Professora adjunta de a Escola_de_Ciência_da_Informação de a Universidade_Federal de Ferramentas

NP's:
1 - Ferramentas de busca em a Web_Beatriz_Valadares_Cendón_Professora adjunta de a Escola_de_Ciência_da_Informação de a Universidade_Federal de ...arquivado!
2 - Ferramentas ...arquivado!

[Frase **2**]

0 - **PRED:n** (**'web' F S**) **Web**

Frase: Web

NP's:
1 - Web ...arquivado!

[Frase **3**]

0 - **PRED:n** (**'diretório' M P**) **Diretórios**

Frase: Diretórios

NP's:
1 - Diretórios ...arquivado!

[Frase **4**]

0 - **PRED:g** (**np**)
1 - **H:n** (**'motor' M P**) **Motores**
1 - **D:g** (**pp**)
2 - **H:prp** (**'de'**) **de**
2 - **D:n** (**'busca' F S**) **busca**
2 - **H:n** (**'busca' F S**) **busca**

Frase: Motores de busca busca

NP's:
1 - Motores de busca busca ...arquivado!

[Frase **5**]

1 - **D:prop** (**'Metamotores_Desde' M/F S**) **Metamotores_Desde**
1 - **D:g** (**np**)
2 - **D:art** (**'o' <artd> M P**) **os**
2 - **H:n** (**'primórdio' M P**) **primórdios**
2 - **D:g** (**pp**)
2 - **H:prp** (**'de' <sam->**) **de**
2 - **D:g** (**np**)
2 - **D:art** (**'o' <-sam> <artd> F S**) **a**
2 - **H:prop** (**'Internet' F S**) **Internet**
2 - **H:prop** (**'Web_Beatriz_Valadares_Cendón_Professora' M S**) **Web_Beatriz_Valadares_Cendón_Professora**
1 - **P:v** (**'haver' fin <ink> PS 1/3S IND**) **houve**
1 - **Od:g** (**np**)
2 - **D:art** (**'o' <artd> F S**) **a**
2 - **H:n** (**'preocupação' F S**) **preocupação**
1 - **D:prp** (**'de'**) **de**
1 - **Od:pron** (**'se' pers <refl> M/F 3P ACC**) **se**
1 - **D:cl** (**icl**)

Frase: Metamotores_Desde os primórdios de a Internet Web_Beatriz_Valadares_Cendón_Professora a preocupação de se

NP's:
1 - os primórdios de a Internet Web_Beatriz_Valadares_Cendón_Professora ...arquivado!
2 - a preocupação de se ...arquivado!

[Frase **6**]

1 - H:prp ('entre') Entre
1 - D:g (np)
2 - D:art ('o' <artd> F P) as
2 - H:n ('ferramenta' F P) ferramentas
2 - D:g (ap)
2 - D:adv ('muito' <quant>) mais
2 - H:adj ('antigo' F P) antigas
2 - D:art ('o' <-sam> <artd> F S) a
1 - UTT:cl (icl)
1 - S:n ('podemse' M S) podemse
1 - P:v ('citar' inf 3S) citar
1 - Od:g (np)
2 - D:art ('o' <artd> M S) o
2 - H:prop ('Archie' M S) Archie
1 - D:prp ('de') de

Frase: Entre as ferramentas mais antigas a podemse o Archie de

NP's:
1 - as ferramentas mais antigas a ...arquivado!
2 - podemse ...arquivado!
3 - o Archie de ...arquivado!

[Frase **7**]

1 - fCs:g (pp)
2 - H:prp ('com') Com
2 - D:g (np)
2 - D:art ('o' <artd> M S) o
2 - H:n ('advento' M S) advento
2 - D:g (pp)
2 - H:prp ('de' <sam->) de
2 - D:g (np)
2 - D:art ('o' <-sam> <artd> F S) a
2 - H:n ('web' <prop> F S) Web
1 - CO:conj ('e') e
1 - D:g (np)
2 - D:art ('o' <artd> F S) a
2 - D:adj ('conseqüente' F S) conseqüente
2 - H:n ('explosão' F S) explosão
2 - D:g (pp)
2 - H:prp ('de' <sam->) de
2 - D:g (np)
3 - D:art ('o' <-sam> <artd> F P) as
3 - H:n ('publicação' F P) publicações

3 - D:g (ap)
3 - H:v ('disponibilizar' pcp F P) disponibilizadas
3 - D:g (pp)
4 - H:prp ('por') por
4 - D:adj ('meio' M S) meio
3 - D:g (pp)
3 - H:prp ('de' <sam->) de
3 - D:pron ('ela' pers <-sam> F 3S NOM/PIV) ela
2 - Od:prop ('Veronica' M/F S) Veronica
1 - P:vp ()
2 - D:v ('começar' fin PS/MQP 3P IND) começaram
2 - SUB:prp ('a') a
2 - H:v ('surgir' inf) surgir
1 - S:g (np)
2 - D:art ('o' <artd> F P) as
2 - H:n ('ferramenta' F P) ferramentas
2 - D:adj ('específico' F P) específicas
1 - A:g (pp)

Frase: Com o advento de a Web e a conseqüente explosão de as publicações disponibilizadas por meio de ela Veronica começaram a surgir as ferramentas específicas

NP's:
1 - a conseqüente explosão de as publicações disponibilizadas por meio de ela Veronica ...arquivado!
2 - as ferramentas específicas ...arquivado!

[Frase **8**]

1 - P:v ('existir' fin PR 3P IND) Existem
1 - A:adv ('hoje') hoje
1 - S:g (np)

Frase: hoje

NP's:
(nenhum)

[Frase **9**]

1 - P:v ('existir' fin PR 3P IND) Existem
1 - S:g (np)
2 - D:num ('dois' <card> M P) dois
2 - H:n ('tipo' M P) tipos
2 - D:adj ('básico' M P) básicos
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:g (np)
2 - H:n ('ferramenta' F P) ferramentas
2 - D:g (pp)
2 - H:prp ('de') de

2 - D:n ('busca' F S) busca
1 - A:g (pp)
2 - H:prp ('em' <sam->) em
2 - D:g (np)
2 - D:art ('o' <-sam> <artd> F S) a
2 - H:n ('web' <prop> F S) Web
1 - : ()

Frase: dois tipos básicos de ferramentas de busca em a Web

NP's:
1 - dois tipos básicos de ferramentas de busca em a Web ...arquivado!

[Frase **10**]

0 - PRED:g (np)
1 - D:art ('o' <artd> M P) os
1 - H:n ('motor' M P) motores
1 - D:g (pp)
2 - H:prp ('de') de
2 - D:n ('busca' F S) busca
1 - CO:conj ('e') e
1 - PRED:g (np)
1 - D:art ('o' <artd> M P) os
1 - H:n ('diretório' M P) diretórios
2 - D:g (pp)

Frase: os motores de busca e os diretórios

NP's:
1 - os ...arquivado!
2 - motores de busca e ...arquivado!
3 - os ...arquivado!
4 - diretórios ...arquivado!

[Frase **11**]

1 - A:adv ('entretanto' <kc>) Entretanto
1 - H:n ('motor' M P) motores
1 - A:g (pp)
2 - H:prp ('a') a
2 - D:cl (icl)
3 - P:v ('partir' inf) partir
3 - A:g (pp)
3 - H:prp ('de' <sam->) de
3 - D:g (np)
3 - D:pron ('esse' det <-sam> <dem> F P) essas
3 - D:num ('dois' <card> F P) duas
3 - H:n ('categoria' F P) categorias
3 - D:adj ('básico' F P) básicas
2 - H:prp ('em' <sam->) em

1 - S:g (np)
2 - D:pron ('outro' det <diff> M P) outros
2 - H:n ('tipo' M P) tipos
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:n ('ferramenta' F P) ferramentas
1 - P:vp ()
2 - D:v ('ter' fin PR 3P IND) têm
2 - H:v ('surgir' pcp) surgido
2 - D:g (np)
1 - A:cl (icl)

Frase: Entretanto motores a partir de essas duas categorias básicas em outros tipos de ferramentas têm surgido

NP's:
1 - motores a partir de essas duas categorias básicas em ...arquivado!
2 - outros tipos de ferramentas ...arquivado!

[Frase **12**]

1 - A:g (pp)
2 - H:prp ('devido_a' <sam->) Devido_a
2 - D:g (np)
2 - D:art ('o' <artd> <-sam> F P) as
2 - H:n ('característica' F P) características
2 - D:adj ('específico' F P) específicas
2 - D:g (pp)
2 - H:prp ('de') de
2 - D:g (np)
2 - D:pron ('cada' det <quant> F S) cada
2 - H:n ('ferramenta' F S) ferramenta
2 - H:n ('categoria' F P) categorias
1 - S:par ()
2 - CJT:g (np)
3 - D:art ('o' <artd> M S) o
3 - H:n ('tipo' M S) tipo
2 - H:n ('tipo' M P) tipos
2 - CJT:n ('número' M S) número
2 - CO:conj ('e' <co-subj>) e
2 - CJT:g (np)
2 - D:art ('o' <artd> F S) a
2 - H:n ('qualidade' F S) qualidade
2 - D:g (pp)
2 - H:prp ('de' <sam->) de
2 - D:g (np)
2 - D:art ('o' <-sam> <artd> M P) os
2 - H:n ('recurso' M P) recursos
2 - D:v ('recuperar' pcp M P) recuperados
1 - A:g (advp)
2 - H:adv ('através') através

2 - D:g (pp)
2 - H:prp ('de') de
2 - D:g (np)
2 - D:pron ('seu' det <poss 3S> M S) seu
2 - H:n ('uso' M S) uso
2 - H:prp ('de') de
1 - P:vp ()
2 - D:v ('poder' fin PR 3P IND) podem
2 - H:v ('variar' inf) variar
1 - A:adv ('enorme') enormemente

Frase: Devido_a as características específicas de cada ferramenta categorias o tipo tipos número e a qualidade de os recursos recuperados através de seu uso de podem variar enormemente

NP's:
(nenhum)

[Frase **13**]

1 - A:g (pp)
2 - H:prp ('para') Para
2 - D:cl (icl)
2 - P:v ('obter' inf) obter
2 - Od:g (np)
3 - D:adj ('bom' M P) melhores
3 - H:n ('resultado' M P) resultados
2 - A:g (pp)
3 - H:prp ('em' <sam->) em
3 - D:g (np)
3 - D:art ('o' <-sam> <artd> F S) a
3 - H:n ('busca' F S) busca
3 - D:g (pp)
3 - H:prp ('de') de
3 - D:n ('informação' F P) informações
2 - H:n ('tipo' M S) tipo
1 - S:g (np)
2 - D:art ('o' <artd> M S) o
2 - D:adj ('primeiro' <NUM-ord> M S) primeiro
2 - H:n ('passo' M S) passo
1 - P:v ('ser' fin PR 3S IND) é
1 - Cs:cl (icl)

Frase: Para obter melhores resultados em a busca de informações tipo o primeiro passo

NP's:
1 - o primeiro passo ...arquivado!

[Frase **14**]

1 - CJT:cl (**fcl**)
5 - D:g (**pp**)
5 - H:prp (**'de' <sam->**) **de**
2 - H:n (**'informação' F S**) **informação**
2 - D:adj (**'instrumental' F S**) **instrumental**
2 - D:g (**pp**)
2 - H:prp (**'para'**) **para**

Frase: de informação instrumental para

NP's:
(nenhum)

FIM

Anexo 6
Classes do Experimento-Piloto

(Autores dos textos da coleção do Experimento-Piloto)

Project: preexperimento

Class Autor

Concrete Class Extends

AGENTE

Direct Instances:

None

Direct Subclasses:

1. Guilherme Ataíde Dias
2. Eliane Maria Stuart Garcez
3. Maria Nélida González de Gómez
4. Gregório J. Varvakis Rados
5. Ilza Leite Lopes
6. Maria Lourdes Blatt Ohira
7. Noêmia Schoffen Prado
8. Maurício Barcellos Almeida
9. Rubén Urbizagástegui Alvarado
10. Beatriz Valadares Cendón
11. Patricia Zeni Marchiori
12. Karina Moutinho
13. Paulo C. Cunha Filho
14. Alessandra Marques de Lima
15. Dinah Aguiar Población
16. Daisy Pires Noronha
17. Yara Rezende
18. Janete Fernandes Silva
19. Marta Araújo Tavares Ferreira
20. Mônica Erichsen Nassif Borges
21. Sergio Luis da Silva

22. Nadia Aurora Peres Vanti
23. Oswaldo H. Yamamoto
24. Paulo Rogério Meira Menandro
25. Sílvia Helena Koller
26. Anna Carolina LoBianco
27. Cláudio Simon Hutz
28. José Lino de Oliveira Bueno
29. Maria do Carmo Guedes
30. Paulo César Rodrigues Borges
31. Max F. Cohen
32. Antonio Cesar Ferreira Guimarães

Anexo 7
Classes do Experimento

(núcleos das proposições)

Project: experimentofinal

Class Predicador

Direct Subclasses:

1. incrementar
2. propiciou
3. são poucos
4. engines ou Internet
5. utilizarem
6. estão
7. utilizarem 1998 estes serviços os estudos de Jansen et al. 1998
8. engines sobre consultas feitas em o
9. Excite e
10. engines Jansen_&_Pooch sobre 2000 consultas
11. acreditam que de utilizando um framework comum será possível comparar os resultados entre diferentes estudos válidos possibilitando responder a questões como
12. são comuns para quaisquer usuários de este tipo de sistema estudos
13. são
14. osusuáriosqueparticiparamdosestudosfazem
15. são de populações diferentes
16. é praticamente nulo
17. encontramos
18. passou a ser reconhecida como elementochave em todos_os segmentos de a sociedade EP
19. surgiram
20. advêm de uma necessidade bem real
21. ser
22. consumido

23. têm sido realizados sobre information literacy
24. características
25. está em sua infância
26. consumido um território ainda indefinido 12. Sendo um conceito dinâmico o constantemente
27. é
28. apresenta
29. codificada e socialmente contextualizadas que podem ser comunicadas estando se portanto um indissociadas de a comunicação
30. pode ser definida como
31. têm emergido
32. apresenta acomodando novos significados ser
33. tem sido proposta7 15 incluindo a cultural ler tecnológica ser acadêmica usar marginal etc. complexos aspectos compartimentalizados de literacy termos exclusivos discutir
34. seriam
35. parece ser
36. direcionados a
37. é objetivo de este artigo propor uma tradução de a expressão nem resolver eventuais questões de gênero

2507. aplicam- se o também a a própria terminologia de a terminologia de a lexicografia e de a lexicologia
2508. Confrontem- se lexicologia à_guisa_de exemplificação
2509. refere- se a o conjunto de vocábulos de um universo de discurso dicionários
2510. é lexicologia
2511. definidos e
2512. organizados em_forma_de dicionário dois
2513. Busca- se assim
2514. Consideram- se aqui apenas um
2515. levando em_conta entre outros elementos os níveis de atualização e de abstração de a linguagem verbal
2516. parecem pertinentes
2517. misma
2518. parecem diccionario y glosario ponderações
2519. embargo
2520. diferencias entre ellos a la
2521. diferencias
2522. radica considerar el nivel linguístico del que forma parte el corpus estudiado
2523. parecem
2524. dato se basa la lengua
2525. tendremos
2526. pertenece
2527. considera la
2528. habla por
2529. podemos diferenciar los

2530. parte diferenciar 2531. conta 2532. têm conceituado 2533. medida 2534. vão além a o afirmar em a nota explicativa de 2535. é Quem em Instrumentação 3 o 2536. sofisticados para biotecnologia aparelhagens 2537. advém 2538. precisa tomar diretrizes específicas 2539. têm encontrado em a informação mais especificamente em a informação científica e técnica e 2540. era tecnológica em 2541. Trabalhar 2542. podem ser levantados

Anexo 7
Classes do Experimento

Project: experimentofinal

Class OBJETO-AFETADO

Direct Subclasses:

1. a publicação de o Journal_des_Sçavans
2. os
3. precursores ou e
4. de
5. o quadro de a produção científica
6. 1999
7. mais_de 600 mil periódicos científicos em todo_o_mundo Biojone many 2001 parts
8. estimandose que sejam escritos diariamente Royal_Society entre seis e sete_mil artigos científicos para alimentar- los e Trzesniak que 2001
9. Tal profusão de títulos
10. um conjunto de problemas Biojone
11. avaliações sobre a qualidade de os artigos
12. dados sobre o impacto
13. a dificuldade de indexação em bases internacionais reconhecidas e questões vinculadas a a língua
14. aspectos que merecem a atenção de a comunidade científica
15. o quadro
16. a participação brasileira produção
17. todas_as áreas de conhecimento em periódicos indexados por o Institute