Um projeto de metodologia para escolha automática de descritores para textos digitalizados utilizando sintagmas nominais

Renato Rocha Souza[1], Lidia Alvarenga[1]

[1]Escola de Ciência da Informação – Universidade Federal de Minas Gerais (UFMG)
Avenida Antônio Carlos, 6627 31270-010 Belo Horizonte, MG – Brasil

{rsouza,lidiaalvarenga}@eci.ufmg.br

Abstract. *It can be noticed that the indexing and representation strategies nowadays seems to be near the exhaustion, and it is worth to investigate new approaches to the indexing and information retrieving systems. Among these, a branch tries to consider the intrinsic semantics of the textual documents using noun phrases as descriptors instead of single keywords. We present in this article a methodology that is being developed in the scope of a doctorate research.*

Resumo. *Com o aparente esgotamento das estratégias atuais de representação e indexação de documentos, faz-se necessário investigar novas abordagens para sistemas de recuperação de informações. Dentre estas abordagens, há uma vertente que busca levar em conta a semântica intrínseca aos documentos textuais, e uma das formas de fazê-lo é através da utilização de sintagmas nominais como descritores, ao invés de palavras-chave. Uma metodologia para atingir tal propósito, que está sendo desenvolvida no escopo de uma tese de doutorado, é apresentada neste artigo.*

1. Introdução

Para lidar com os constantes e ininterruptos ciclos de criação e demanda de informação, há muito vêm sendo criados sistemas de recuperação de informações[1] que utilizam diversas tecnologias mecânicas e digitais de computação, para gerenciar grandes acervos de documentos. Podemos citar, dentre eles, a Internet, as intranets empresariais com seus portais corporativos, e as bibliotecas digitais.

Neste contexto, este artigo apresenta uma pesquisa em andamento, desenvolvida no âmbito do curso de doutorado do autor, no Programa de Pós Graduação em Ciência da Informação da Universidade Federal de Minas Gerais. A pesquisa pretende contribuir para enfrentar alguns dos muitos desafios que surgem quando lidamos com massivas quantidades de dados, como nos grandes acervos de documentos digitais,

[1] Entende-se, no escopo deste trabalho, que os sistemas de recuperação de informações são sistemas, usualmente baseados em tecnologias digitais, que lidam com a organização e o acesso aos itens de informação, desempenhando as atividades de representação, armazenamento e recuperação destes itens.

notadamente quando estes precisam ser regularmente organizados e pesquisados, recuperando em tempo hábil informação relevante para algum objetivo específico.

Com o aparente esgotamento[2] das estratégias tradicionais de busca em sistemas de recuperação de informações, entendemos que a melhoria da eficácia do serviço ao usuário dos sistemas depende dos resultados em diversas linhas de pesquisa, em todo o espectro da cadeia de processos de tratamento da informação. Temos como hipótese de trabalho que as principais frentes de atuação são as seguintes:

1. A exploração das informações semânticas e semióticas intrínsecas aos dados, de forma a expandir a compreensão das unidades e padrões de significado em textos, imagens e outras mídias;
2. O desenvolvimento de novas possibilidades de marcação semântica dos dados utilizando-se metalinguagens, criando espécies de índices acoplados aos próprios documentos com termos amplamente consensuais e não ambíguos, para que estes possam ser mais facilmente manipulados e identificados por computadores e outros dispositivos e, como conseqüência, pelos usuários;
3. O desenvolvimento de estratégias de apresentação da informação recuperada nas buscas sob formas altamente significativas, ou contextuais[3] - como em algumas interfaces gráficas – de forma que as relações entre os conceitos, e em conseqüência, os contextos, sejam evidentes; e também por estratégias que busquem estimular os vários órgãos sensoriais ao mesmo tempo – como nas ferramentas multimídias – para que a absorção das informações pelos usuários seja maior. Através destas interfaces e estratégias, as informações podem ser apresentadas de forma a possuírem conexões visuais aos seus contextos de origem, permitindo ao usuário refinar os resultados através da definição das conexões pertinentes e a exclusão das conexões geradas pelo ruído informacional;
4. A construção e manutenção de perfis personalizados de utilização, de forma que o SRI "aprenda" com a forma de trabalho do usuário e possa utilizar estas informações específicas para melhorar a estratégia de busca do SRI.

Uma abordagem completa para a organização e a recuperação de informações, visando a melhoria dos Sistemas de Recuperação atuais, deve unir estas estratégias e soluções, buscando:

- A indexação dos documentos utilizando representações mais significativas, de modo a aumentar e melhorar os pontos de acesso e a relevância das informações recuperadas;
- Prover uma forma adequada de apresentar as informações recuperadas aos usuários, de maneira que sejam intuitivas e facilmente compreensíveis;

[2] As estratégias tradicionais baseiam-se em modelagens dos documentos a partir da distribuição de suas palavras-chave. Embora existam propostas de avanços, parece haver um limite para a eficácia de tais estratégias.

[3] Informação apresentada sem desprezo do contexto que lhe confere sentido.

- Utilizar no processo de indexação padrões universais de registros de metadados para que os sistemas sejam interoperáveis entre si;
- Adaptar-se continuamente ao usuário, sendo preferível que possa aprender com a forma com que trabalha, de modo que as buscas sejam continuamente refinadas através de um trabalho de personalização.

Existem hoje diversas tentativas, mais ou menos coordenadas, de se abordar estas ações fundamentais, mas uma real integração demandaria a pesquisa em diferentes áreas do conhecimento e campos de pesquisa, como a ciência da informação, a lingüística, a ciência da computação, a sociologia, a antropologia, a comunicação, a psicologia cognitiva, entre outras.

De maneira isolada, há pesquisas em cada uma destas vertentes, mas é pouco explorada a utilização da semântica embutida nos próprios documentos, ou seja, das potencialidades intra-textuais da linguagem natural, para automatizar e melhorar as tarefas de indexação, organização e recuperação de informações.

Pesquisas nesta área incluem o uso de estruturas profundas da linguagem natural, como os sintagmas verbais e nominais, para indexação e recuperação [KURAMOTO, 1996 e 1999; MOREIRO et al, 2003]; e de ferramentas de representação de relacionamentos semânticos e conceituais, como os tesauros, para ampliar a gama de informações recuperadas e aferição de contextos [SPARCK JONES & WILLETT, 1997, pp. 15-20]; além de outras estratégias derivadas da lingüística e da ciência da informação. Todas estas estratégias são fortemente atreladas ao idioma, o que faz com que os possíveis resultados da pesquisa tenham uma aplicação circunscrita ao contexto da língua da comunidade em questão. As metodologias, entretanto, são generalizáveis e sua aplicabilidade a outras linguagens é perfeitamente possível.

Neste projeto, pretende-se apresentar uma metodologia para aproveitar o potencial de uso dos sintagmas nominais como descritores de documentos em processos de indexação. Parte-se da hipótese de que os sintagmas nominais, pelo maior grau de informação semântica embutida, podem vir a se tornar mais eficazes do que as palavras-chave usualmente extraídas e utilizadas como descritores em outros processos automatizados de representação de documentos, tais como os observados nos mecanismos de busca da Internet, ou em sistemas de leitura das palavras-chave fornecidas pelo autor dos documentos.

Alguns trabalhos anteriores se apresentam como marcos a partir dos quais se pretende avançar; dentre eles, a pesquisa sobre a viabilidade do uso dos sintagmas nominais para sistemas de recuperação de informações de KURAMOTO [1996 e 1999], e as ferramentas para marcação sintática do português e automatização da extração de sintagmas nominais desenvolvidas no âmbito dos projetos da Southern Denmark University [BICK, 2000], de VIEIRA [2000] e do PROJETO DIRPI [2001]. A partir destes resultados e ferramentas, pretende-se propor uma metodologia de escolha automática de descritores para documentos que utilize os sintagmas nominais em vez de palavras-chave para documentos textuais digitalizados em língua portuguesa.

2. Sintagmas nominais e sistemas de recuperação de informações

Entendemos por **sintagmas** certos grupos de palavras que fazem parte de seqüências maiores na estrutura de um texto, mas que mostram um grau de coesão entre eles

[PERINI, 1995]. Os constituintes ou sintagmas podem ou não ser facilmente identificáveis, sendo que por vezes é necessário recorrer a outros recursos para que seja feita a "demarcação" sintática. Perini acredita que a intuição "subjetiva, mas nem por isso duvidosa" que nos permite separar a oração em seus constituintes imediatos pode ser caracterizada através de critérios puramente formais [1985, pp. 42-43], mas há quem defenda que a identificação dos constituintes é somente completa através de uma abordagem cognitiva e amplamente contextual [LIBERATO, 1997], que só é esperada na análise do discurso[4] e na pragmática[5]; ou através de outros modelos gramaticais, como a análise transformacional [RUWET, 1975, pp.155-212 e 223-279]. Para a análise semântica, há também o problema das situações anafóricas, que ocorrem quando uma estrutura de uma oração se apresenta reduzida porque ocorre na vizinhança de outra estrutura oracional de certa forma paralela, dependendo desta para sua total compreensão [PERINI, 1986, p. 57].

De acordo com MIORELLI [2001], os sintagmas nominais podem ser entendidos – e tratados – de forma sintática, privilegiando a forma; ou semântica, buscando os significados maiores; cada uma com suas especificidades e implicações. A abordagem semântico-pragmática, utilizada por LIBERATO [1997], não prescinde de um "interpretador de contextos", natural na cognição humana, mas dificilmente implementado em heurísticas de inteligência artificial. A forma sintática, como analisada por PERINI [1986, 1995 e 1996] está mais relacionada à estrutura das orações em si, e é mais facilmente tratada computacionalmente. Assim como no trabalho de MIORELLI [2001], esta é a abordagem que será utilizada no âmbito deste projeto, da mesma forma que, provavelmente, em quaisquer abordagens, e com quaisquer ferramentas, que busquem a automatização de extração dos sintagmas nominais.

Sistemas de recuperação de informações usualmente adotam termos índices para indexação de documentos, sendo que estes termos índice são usualmente palavras-chave. Há uma idéia fundamental embutida de que a semântica dos documentos e das necessidades de informação do usuário podem ser expressas através destes conjuntos de palavras, o que é, claramente, uma grande simplificação do problema, porque grande parte da semântica do documento ou da requisição do usuário é perdida quando se substitui o texto completo por um conjunto de palavras [BAEZA-YATES & RIBEIRO-NETO, 1999, p.19].

Há, na literatura, registros de algumas tentativas de otimizar a organização dos documentos em SRIs através de um processamento aprofundado da linguagem natural dos documentos. Dentre elas, a identificação de "grupamentos de substantivos" (*noun groups*), ao invés de palavras-chave, se afigura uma boa estratégia para seleção de termos de indexação, uma vez que os substantivos costumam carregar a maior parte da semântica de um documento, ao invés de artigos, verbos, adjetivos, advérbios e conectivos. Esta proposta estabelece uma visão conceitual do documento [ZIVIANI, in BAEZA-YATES & RIBEIRO-NETO, 1999, pp.169-170]. Os grupamentos de substantivos são conjuntos de nomes nos quais a distância sintática no texto (medida

[4] Estuda a estrutura e a interpretação dos textos.

[5] Ocupa-se da relação dos enunciados lingüísticos com a situação extralingüística em que se inserem [PERINI, 1995].

pelo número de palavras entre dois substantivos) não excede um limite predefinido. Uma metodologia que extrapola esta proposta é a identificação dos sintagmas nominais e o seu uso como descritores, como proposto neste projeto.

SALTON & MCGILL [1983, pp. 90-94] discutem algumas abordagens teóricas para o uso de métodos lingüisticos na recuperação de informações ; entre elas, a análise da estrutura sintática (*parsing*) dos documentos de forma a identificar as estruturas sintagmáticas. Estes autores, entretanto, apontam as dificuldades intrínsecas ao processo de análise semântica através da análise sintática e exemplificam casos em que é impossível o reconhecimento não ambígüo de relações semânticas através dos componentes da sentença, sugerindo que um modelo baseado em gramáticas transformacionais poderia trazer melhores resultados. Neste ponto, parecem então concordar com LIBERATO [1997], que entende que a análise completa das estruturas semânticas só é possível através da análise cognitiva dos contextos. Ao indicar a maior eficácia relativa dos algoritmos de geração de frases baseadas em frequência de palavras, talvez apontem uma alternativa para a melhoria do algoritmo proposto neste trabalho. Outra alternativa apontada é a interferência humana no processo de desambiguação através de uma interface, o que seria pouco desejável num processo que pretende ser automático.

Um importante caminho de pesquisa que visa resolver os problemas de desambiguação semântica através da análise dos contextos é resolução de correferência, ou resolução anafórica [VIEIRA, 1998 e 2000; SANT'ANNA, 2000 ; ROSSI et al, 2001; GASPERIN et al, 2003]. Uma cadeia de correferência é uma seqüência de expressões em um discurso que se referem a uma mesma entidade, objeto ou evento. Essas cadeias são úteis para a representação semântica de um modelo de domínio, e podem melhorar a qualidade dos resultados em diversas aplicações de processamento de linguagem natural, como recuperação e extração de informações, geração automática de resumos, traduções automáticas, entre outros [ROSSI et al, 2001]. O processo de resolução de correferências envolve a identificação e extração dos sintagmas nominais.

LE GUERN e BOUCHÉ [apud KURAMOTO, 1999] apontam o sintagma nominal como a menor unidade de informação contida em um texto. O grupo de pesquisas SYDO, ao qual pertencem estes pesquisadores, tem como fundamento teórico a utilização de sintagmas nominais como descritores [Ibidem, 1996]. Ao trabalhar em parceria com este grupo, KURAMOTO [1999], em sua tese de doutorado, desenvolveu uma pesquisa fundamental para a consideração da utilização de sintagmas nominais como descritores. Já em um trabalho anterior, KURAMOTO [1996] vislumbrou a maquete proposta na tese e já apontava o potencial natural de organização dos sintagmas nominais que, se explorado convenientemente, poderia propiciar aos usuários maior facilidade no uso de um SRI e resultados mais precisos em resposta a um processo de busca de informação.

O sistema desenvolvido por Kuramoto pode ser considerado uma inspiração para o presente trabalho, na medida em que, em ambos, busca-se uma alternativa para uma melhor indexação utilizando-se sintagmas nominais. Entretanto, em sua maquete, "A extração dos sintagmas nominais foi realizada de forma manual, simulando uma extração automática. Este procedimento foi adotado em função da não-existência ainda de um sistema de extração automática de SN em acervos contendo documentos em língua portuguesa". [1996, p.6]. Ao menos um sistema deste tipo, entretanto, se

encontra hoje disponível, e foi disponibilizado para o propósito deste trabalho [GASPERIN et al, 2003]. Uma outra diferença fundamental é o objetivo. Se no projeto de Kuramoto buscava-se apresentar uma maquete de um SRI baseado em sintagmas nominais, o objetivo deste trabalho é propor uma metodologia de auxílio à indexação automática utilizando uma metodologia aplicada sobre os sintagmas nominais extraídos automaticamente. Diferenças a parte, o fundo filosófico é bastante comum.

3. O método proposto

Acreditamos que, neste ponto, todo o cabedal teórico necessário ao entendimento do contexto no qual se insere o projeto já tenha sido apresentado e possa ser corretamente entendido. Nesta altura, cabe apresentar a metodologia proposta para consecução dos objetivos.

A) Para o objetivo geral: "Propor uma metodologia para escolha semi-automática de descritores para documentos textuais digitalizados em língua portuguesa, utilizando as estruturas sintáticas e semânticas conhecidas como sintagmas nominais"; pretendemos perfazer os seguintes passos, em seguida explicitados e comentados:

1. Escolher um corpus considerável de textos publicados recentemente em meio eletrônico em revistas científicas da área de Ciência da Informação;
2. Analisar o corpus escolhido, retirar suas palavras-chave atribuídas pelos autores e informações adicionais de formatação, e extrair os sintagmas nominais do corpo do texto, utilizando as ferramentas detalhadas adiante;
3. Verificar a freqüência de incidência dos sintagmas nominais e adotar uma lógica para escolha dos mais significativos;

Neste ponto, talvez esteja uma das partes mais críticas da metodologia. A lógica para escolha dos sintagmas nominais relevantes está para ser estabelecida através da manipulação dos dados empíricos. Pode-se esperar, entretanto, que venha a ser derivada dos algoritmos de extração de palavras-chave baseados na lei de Zipf (freqüência simples com descarte dos picos, pesos relacionados à freqüência inversa nos documentos, valor discriminatório dos termos) apresentados anteriormente ou mesmo o algoritmo composto proposto por SALTON & MCGILL [1983, pp.71-75]. Há que se fazer as adaptações necessárias ao fato de não mais estarmos tratando de palavras-chave, mas sim de sintagmas nominais. Não são descartadas, entretanto, as metodologias de busca sequencial [BAEZA-YATES & RIBEIRO-NETO, 1999, pp. 209-215]. Espera-se que, após a obtenção de resultados satisfatórios em um pequeno subconjunto dos textos, o restante do corpus seja usado para validação da metodologia escolhida.

4. Verificar a incidência dos sintagmas nominais escolhidos em um tesauro na área da Ciência da Informação; separar os verificados em conjuntos doravante denominados: a) os que constam no tesauro e b) os que não constam no tesauro;
5. Comparar, separadamente, os sintagmas dos conjuntos a) e b) definidos acima com as palavras-chave escolhidas pelos autores dos textos e com o assunto do texto, na forma em que puder ser depreendido. Analisar os resultados;

6. Julgar, dentre aqueles que não constam do tesauro, quais deveriam constar; separar estes sintagmas nominais em conjuntos doravante denominados: c) sintagmas nominais que deveriam constar no tesauro; d) sintagmas nominais referentes a conceitos relevantes de áreas afins e; e) sintagmas nominais devem ser ignorados de forma semelhante às *stopwords*. Os sintagmas recolhidos em c) serão considerado para fins de validação dos próximos sintagmas, enquanto os sintagmas em d) serão analisados mais detidamente, pois a metodologia poderia ser ampliada com a utilização de tesauros de áreas correlatas. Os sintagmas em e) serão descartados das próximas análises;

A proposta metodológica deve sofrer alterações, na medida em que os dados empíricos forem manipulados e analisados. No entanto, este trabalho não teria sido possível sem as ferramentas de extração automática que, assim como os corpora, foram gentilmente cedidos pelos proprietários e desenvolvedores. Em seguida passamos à descrição destas ferramentas.

4. Ferramentas utilizadas

O trabalho de análise proposto na metodologia acima descrita é talvez a ponta do iceberg de todo o esforço computacional necessário, compreendido no processo. Para que seja possível a análise dos descritores, os sintagmas nominais tiveram que ser extraídos, no caso, automaticamente e de forma bastante veloz. Os textos dos corpora foram escolhidos pelo autor e transformados em formato de texto simples. Em seguida, foram submetidos sucessivamente ao processamento da ferramenta "Palavras" da Southern University of Denmark e o software "Palavras Extractor" desenvolvido em conjunto pela Universidade do Vale do Rio dos Sinos (Unisinos) de São Leopoldo e a Universidade de Évora, em Portugal. Os pesquisadores da Unisinos e da Universidade de Évora cederam, para os propósitos deste trabalho, uma interface integrada através da qual grande parte do processamento automático envolvido, inclusive o desempenhado pelo *site* dinamarquês, foi realizado, durante os meses de agosto e setembro de 2003. Em seguida vamos descrever em mais detalhes estas ferramentas.

4.1. O VISL e o processador "Palavras"

A Southern University of Denmark, desenvolveu e tornou público uma ferramenta de processamento morfo-sintático de textos digitalizados em português chamada "Palavras", que faz parte de um conjunto de ferramentas multilinguais chamado VISL (Virtual Interactive Sintax Learning), disponível no endereço da Internet: http://visl.sdu.dk/visl/. No VISL, várias ferramentas, para cada um dos idiomas suportados, operam em modo automático ou semi-automático, nos quais um usuário submete sentenças ou textos completos em uma das linguagens admitidas (dentre as quais, o português) e recebe de volta os textos marcados. As análises podem ser feitas em diferentes níveis (morfológico, sintático, semântico) e em várias formas de visualização, como textos simples, árvores sintáticas ou marcação com cores [BICK, 1996, 2001 e 2003]. O projeto VISL é altamente orientado a produtos e processos, uma vez que novas ferramentas tem sido constantemente disponibilizadas gratuitamente na Internet na medida em que os protótipos se mostram funcionais. O VISL é baseado em um emaranhado de páginas HTML, scripts CGI (commom gateway interface), Java e

PERL, e oferece uma interface gráfica que permite aos usuários uma diversidade de opções [VISL, 2003].

Uma das possibilidades de marcação oferecidas pelas ferramentas do *site* indica as categorias gramaticais e a função de cada palavra no contexto de uma oração. Através desta marcação e um processamento posterior, é possível extrair os sintagmas nominais das sentenças de um texto. Este pós-processamento pode ser feito manualmente, através da análise das funções e da proximidade das palavras, ou pode ser automatizado, o que é o objetivo da ferramenta "Palavras Extractor", descrita a seguir.

4.2. A extração automática de sintagmas nominais

A partir da ferramenta computacional "Palavras" do VISL, o Departamento de Lingüística Computacional Aplicada do Centro de Ciências Exatas e Tecnológicas da Universidade do Vale do Rio dos Sinos, sob a coordenação da professora doutora Renata Vieira, em parceria com o departamento de Informática da Universidade de Évora, de Portugal; desenvolveu, no escopo do projeto de cooperação DIRPI [PROJETO DIRPI, 2001], um conjunto de programas de interface e de pós-processamento dos resultados, chamados internamente de "Palavras Extractor". Os programas estabelecem um acesso ao *site* VISL, enviam textos para o analisador sintático PALAVRAS para o português [Bick, 2000 apud GASPERIN et al, 2003]. A saída desse analisador é convertida em um conjunto de arquivos XML: o arquivo de palavras (elementos <word>); um arquivo com as categorias morfo-sintáticas (POS - Part Of Speech) das palavras do corpus e um arquivo com as estruturas sintáticas das sentenças, representadas por "chunks" [GASPERIN et al, 2003]. A partir destes três arquivos XML gerados, pode-se trabalhar com mais facilidade e desenvoltura em comparação com o *output* do site VISL, pois através do uso de folhas de estilo (XSL) específicas, é possível então extrair os sintagmas nominais de qualquer texto ou corpus da língua portuguesa. Assim como são extraídos os sintagmas nominais, é possível extrair outras instâncias gramaticais, dependendo do interesse da pesquisa em questão, bastando para tanto o desenho de uma nova folha de estilo. Os sintagmas nominais utilizados neste projeto foram obtidos utilizando-se a folha de estilo específica para extração de sintagmas nominais, cedida pela pesquisadora da Unisinos Claudia Camerini Correa Perez, e o software XML SPY (http://www.altova.com), utilizado para aplicação da transformação XSL nos arquivos XML gerados. O resultado final são arquivos HTML contendo os sintagmas nominais na seqüência em que ocorrem no texto, desde os sintagmas nominais máximos até os sintagmas aninhados na estrutura máxima.

5. Conclusões

Ainda é cedo para extrair conclusões, mas a julgar pelos resultados preliminares, a proposta tem grandes chances de se tornar um método seguro para a atribuição de descritores, com variados graus de verossimilhança representacional, dependendo de características como:

O campo de conhecimento de que tratam os textos;

A qualidade e atualização do tesauro utilizado;

O tamanho dos corpora analisados previamente;

Na medida em que mais resultados forem alcançados, serão divulgados nos fóruns apropriados.

6. Referencias

BAEZA-YATES, R.; **RIBEIRO-NETO**, B. *Modern Information Retrieval.* New York: ACM Press, 1999. 511p.

BICK, Eckhard. *Parsers and its applications*. (s/d) Disponível na Internet: http://www.hum.au.dk/lingvist/lineb/home_uk.htm. Consultado em 07/2003

_____. *Automatic parsing of Portuguese*. In: Proceedings of II Encontro para o Processamento Computacional do Português Escrito e Falado, SBIA, 1996, Curitiba. Disponível na Internet: http://beta.visl.sdu.dk/~eckhard/postscript/curitiba.ps. Consultado em 07/2003

_____. *The VISL System*: research and applicative aspects of IT-based learning. In: Proceedings of NoDaLiDa, 2001,Uppsala. Disponível na Internet: http://stp.ling.uu.se/nodalida01/pdf/bick.pdf. Consultado em 07/2003

GASPERIN, Caroline Varaschin; **GOULART**, Rodrigo Rafael Vilarreal e **VIEIRA**, Renata. *Uma Ferramenta para Resolução Automática de Correferência.* In: Anais do XXIII Congresso da Sociedade Brasileira de Computação, VI Encontro Nacional de Inteligência Artificial, Vol VII. Campinas, 2003.

GASPERIN, Caroline Varaschin; **VIEIRA**, Renata; **GOULART**, Rodrigo Rafael Vilarreal e **QUARESMA**, Paulo. *Extracting XML chunks from Portuguese corpora.* In: Proceedings of the Workshop on Traitement automatique des langues minoritaires. 2003. Batz-sur-Mer.

PROJETO DIRPI: Desenvolvimento e Integração de Recursos para Pesquisa de Informação. Cooperação Científica e Técnica Luso-Brasileira. ICCTI/GRICES-CAPES, Universidade de Évora, Universidade Nova de Lisboa, Unisinos, PUC-RS. Julho de 2001.

KURAMOTO, Hélio. *Uma abordagem alternativa para o tratamento e a recuperação de informação textual*: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, 1996. Disponível na Internet: http://www.ibict.br/cionline/250296/25029605.pdf. Consultado em 07/2003.

_____. *Proposition d'un Système de Recherche d'Information Assistée par Ordinateur Avec application à la langue portugaise*. 1999. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière - Lyon 2, Paris, França.

LIBERATO, Yara G. *A Estrutura do Sintagma Nominal em Português:* uma abordagem Cognitiva. 1997. 203 f. Tese (Doutorado em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

MIORELLI, S. T. *Extração do Sintagma Nominal em sentenças em Português*. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

MOREIRO, José; **MARZAL**, Miguel Ángel; BELTRÁN, Pilar. *Desarrollo de un Método para la Creación de Mapas Conceptuales*. Anais do ENANCIB, Belo Horizonte, 2003.

PERINI, Mário A. *A Gramática Gerativa*: introdução ao estudo da sintaxe portuguesa. 2ª edição. Belo Horizonte: Vigília, 1985. 254 p.

_____. *Gramática descritiva do português*. 2ª edição. São Paulo: Editora Ática, 1995. 380p.

PERINI, Mário A.; **FRAIHA**, Sigrid; **FULGÊNCIO**, Lúcia; **BESSA NETO**, Regina. *O SN em português*: A hipótese mórfica. **Revista de Estudos de Linguagem** - UFMG, Belo Horizonte, Julho / Dezembro 1996. p. 43-56.

ROSSI, Daniela; **PINHEIRO**, Clarissa; **FEIER**, Nara e **VIEIRA**, Renata. *Resolução automática de Correferência em textos da língua portuguesa*. REIC Revista de Iniciação Científica da SBC, http://www.sbc.org.br/reic/, v. 1, n. 2, 2001.

RUWET, Nicolas *Introdução à Gramática Gerativa*. São Paulo: Perspectiva, Editora da Universidade de São Paulo, 1975. 357 p.

SALTON, Gerard e **MCGILL**, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983. 448 p.

SANT'ANNA, V. *Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita*. 100 f. 2000. Dissertação (Mestrado em Informática) – Instituto de Informática da PUC-RS – Porto Alegre.

SPARCK JONES, K. e **WILLETT**, P. (orgs.). *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997. 589p.

VIEIRA, R. *A review of the Linguistic literature on definite descriptions*. 1998. Acta Semiotica et Lingvistica, Vol. 7 : 219-258.

VIEIRA, R. et al. *Extração de Sintagmas Nominais para o Processamento de Co-referência*. 2000. Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR, 19-22 Novembro Atibaia SP.

VISL. *About VISL*. Disponível na Internet: http://visl.hum.sdu.dk/visl/about/index.html. Consultado em 05/2003.