Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Construção do léxico de aplicações

Miriam Sayão[1,2] Gustavo R. de Carvalho[2]

[1]Faculdade de Informática - Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre - RS - Brasil
[2]Departamento de Informática - Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro - RJ - Brasil
miriam@inf.puc-rio.br guga@les.inf.puc-rio.br

***Abstract**: Requirements engineering activities can involve the reading of documents which impact in some way the system being developed. Examples of such documents include organization's standards, laws, interviews summaries and also requirements document. In that process, one of most important tasks is the construction of a lexicon for the application, because in it the symbols of the domain are registered. The lexicon is a basis to promote the understanding among customers, users and software professionals. In this article we propose a strategy for automatic identification of symbols to compose a lexicon for the application, focusing in relevant actors and resources for the elicitation process. The proposed strategy evaluates documents handled in requirements engineering process; those documents are written in natural language, and the strategy aim to extract expressions that identify actors and resources to compose a lexicon for the application. Results from two case studies illustrate the application of our strategy.*

***Resumo:** Atividades associadas ao processo de requisitos podem envolver leitura de documentos que impactem de alguma forma o sistema em desenvolvimento. Exemplos de documentos incluem padrões da organização, legislação pertinente, atas de reuniões ou entrevistas ocorridas durante o processo de elicitação e também o próprio documento de requisitos sendo elaborado. Nesse processo, a construção do léxico da aplicação é tarefa prioritária, pois nele são registrados os símbolos próprios do domínio da aplicação. O léxico é a base para o entendimento entre clientes, usuários e profissionais de software. Neste artigo propomos uma estratégia para a identificação automática de símbolos com o objetivo de compor o léxico da aplicação, com o enfoque a identificação de atores e recursos relevantes para o processo de elicitação. A estratégia proposta avalia documentos manipulados no processo de requisitos e escritos em linguagem natural, extraindo expressões que identificam atores e recursos para o léxico da aplicação. Resultados obtidos em dois estudos de caso ilustram a aplicação da estratégia.*

Keywords: engenharia de requisitos, léxico ampliado da linguagem, léxico de aplicações, sintagmas nominais

1. Introdução

A engenharia em torno de requisitos é composta basicamente das atividades de elicitação, modelagem e análise (verificação e validação de requisitos). Durante o processo de elicitação, o engenheiro de requisitos utiliza diversas fontes visando obter as informações necessárias ao bom entendimento do problema. Fontes tradicionais de informação incluem clientes e futuros usuários do sistema, mas a leitura cuidadosa de documentos também é freqüente nesta etapa. Normas internas da organização, legislação pertinente à solução sendo buscada, documentação de outros sistemas são exemplos de documentos que devem ser avaliados. Documentos gerados durante o processo de elicitação também são freqüentemente consultados: registros de entrevistas, atas de reuniões, memorandos e mensagens trocados entre participantes e o próprio documento de requisitos são parte dessa documentação. Esses documentos são escritos em linguagem natural.

Durante o processo de elicitação termos próprios do domínio da aplicação são utilizados por clientes e usuários, e também estão presentes na documentação manipulada. Considerando que neste processo participam profissionais com diferentes funções e competências, termos técnicos ou de significado diverso do usual podem dar margem a diferentes interpretações para um mesmo termo ou expressão. A necessidade de um léxico para a aplicação é justificada pela necessidade de que todos os participantes compartilhem uma mesma compreensão dos termos próprios do domínio.

O léxico da aplicação pode assumir diferentes formatos, por exemplo, um glossário na sua forma mais simples, ou uma ontologia na sua forma mais elaborada. O léxico não é apenas uma exigência de processos de qualidade, mas se constitui também em fonte de consulta para os participantes do processo de requisitos. A estratégia aqui apresentada para a identificação automática de símbolos para o léxico é dirigida ao Léxico Ampliado da Linguagem (LAL), proposto por Leite [Leite93]. Símbolos incluídos no LAL correspondem a sujeitos, objetos, verbos ou estados, e serão detalhados na seção 2. Outras abordagens para o léxico, no entanto, também podem se beneficiar desta estratégia, pois os símbolos utilizados pelo LAL correspondem a entidades utilizadas em diferentes propostas.

A necessidade de um léxico abrangente é maior em ambientes distribuídos de desenvolvimento, onde os participantes estão geograficamente distantes, possuem diferenças culturais e estão sujeitos a dificuldades de comunicação. O léxico representa o conhecimento compartilhado, diminuindo dificuldades de entendimento em relação a termos próprios do domínio da aplicação. Mesmo entre países que compartilham uma mesma língua as diferenças culturais existentes também podem ser diminuídas com o uso do léxico. Isto possibilita que atividades do processo de requisitos, reconhecidamente intenso em atividades de comunicação, sejam menos sujeitas a problemas derivados de falhas na comunicação, contribuindo positivamente para as atividades relacionadas ao processo de requisitos.

Atividades de elicitação inserem-se no contexto de gerenciamento de requisitos, e estamos desenvolvendo uma estratégia para o gerenciamento de requisitos em ambientes distribuídos de desenvolvimento, utilizando uma arquitetura baseada em agentes de software [Sayão05]. Nesse contexto, o presente trabalho relata os primeiros resultados na exploração de técnicas básicas para subsidiar a construção dos

agentes, em particular do Agente Léxico, responsável por funções associadas ao tratamento de textos, e do Agente Construtor do Léxico, responsável por funções de manutenção do léxico da aplicação, visando à atualização da base de conhecimentos para o domínio da organização.

O restante deste artigo está organizado da seguinte maneira: na Seção 2 apresentamos os conceitos envolvidos na construção de glossários e léxicos. A seção 3 detalha o processo de construção automática proposto, segundo uma perspectiva de processo. Na seção 4 apresentamos dois estudos de caso utilizados para a demonstração da nossa estratégia. A Seção 5 contextualiza o trabalho com a literatura da área e na Seção 6 apresentamos as considerações finais.

2. O processo de requisitos e léxicos

Nosso desafio está relacionado à identificação de símbolos que irão compor o léxico da aplicação e contribuir para a construção do vocabulário do domínio da organização. Processos de desenvolvimento de software, como por exemplo o RUP (Rational Unified Process), definem glossários como um dos artefatos a serem gerados no processo de requisitos. Um glossário pode ser entendido como uma forma simplificada de léxico, estruturado de forma linear e contendo termos e suas definições. Os termos a serem inseridos num glossário são aqueles utilizados pelos participantes do processo para fazer referências às características da aplicação, visando facilitar o entendimento entre eles. O glossário deve ser construído durante a elaboração do modelo de negócios ou modelo de domínio. Já o Léxico Ampliado da Linguagem, ou LAL, é uma forma mais elaborada de registro de termos próprios do domínio da aplicação, fornecendo mais informações que simplesmente a definição de um termo. Detalharemos o LAL, segundo proposta de [Leite90] [Leite93].

2.1 LAL - Léxico Ampliado da Linguagem

Termos registrados no LAL são tipificados, e esta é a primeira diferença em relação a um glossário de termos do domínio da aplicação. Termos inseridos no LAL representam símbolos característicos do domínio da aplicação, e correspondem a um de quatro tipos: sujeito, objeto, verbo ou estado. Símbolos do LAL possuem noção e denotação. A noção de um símbolo é o que o define, e a denotação registra os impactos que o símbolo provoca ou recebe no domínio considerado. A tabela 1 [Leite90] registra os quatro tipos de símbolos e impactos correspondentes.

Sujeitos correspondem a entidades ativas, atores com papel relevante para a aplicação; um sujeito pode ser um ator, um componente ou um outro sistema com o qual deverá ocorrer interação. Verbos registram ações ou funcionalidades a serem desempenhadas pelos sujeitos ou pelo sistema em desenvolvimento, com algum impacto ou reflexo no ambiente operacional. Objetos são entidades passivas utilizadas ou necessárias a uma ação ou conjunto de ações, e estados são caracterizados por atributos significativos que registram valores em diferentes momentos da execução do sistema. A tabela 2 apresenta exemplo de entrada para o léxico de um sistema para bibliotecas de uma universidade.

Tabela 1 - Símbolos do LAL, noção e impactos [Leite90]

Tipo do símbolo	**Noção**	**Impacto**
Sujeito	quem é o sujeito	ações que executa
Verbo	quem realiza, quando acontece e quais os procedimentos	quais os reflexos das ações no ambiente e novos estados decorrentes
Objeto	definir o objeto e identificar outros objetos com os quais ele se relaciona	ações que podem ser aplicadas ao objeto
Estado	o que indica e ações que levaram a esse estado	identificar outros estados que podem ocorrer a partir do estado que se descreve

Tabela 2 - Exemplo de símbolo de um léxico do tipo LAL

Léxico Ampliado da Linguagem - Sistema de Bibliotecas
Usuário Tipo do símbolo: sujeito Noção: pessoa que pode utilizar a biblioteca; pode ser um aluno, professor ou funcionário da universidade Impactos: usuário é cadastrado no sistema usuário é retirado do cadastro de usuários usuário retira obras da biblioteca usuário devolve obras anteriormente retiradas usuário renova datas para devolução de obras anteriormente retiradas

3. Processo de extração de símbolos para um léxico

Neste artigo apresentamos o processo proposto para a extração de sujeitos e de objetos dos documentos da organização, utilizando uma abordagem baseada na utilização de sintagmas nominais. Sintagmas nominais são definidos como (a) classe gramatical com comportamento sintático de sujeito, de objeto direto e, se precedido de preposição, de adjunto adnominal ou de objeto indireto [Perini98] [Vieira01]; (b) enunciado que representa um conceito ou uma entidade (abstrata ou concreta) identificada por nomes próprios ou sintagmas nominais descritivos; pode ainda representar um papel [Liberato97].

No contexto da engenharia de software, Booch et al afirmam que um ator representa um papel que um ser humano, um dispositivo de hardware ou mesmo um outro sistema desempenha [Booch00]. Em documentos técnicos, atores são registrados através de identificadores que incluem nomes próprios, profissões e papéis. Em termos lingüísticos podemos associar atores a sintagmas nominais. Recursos correspondem a objetos que ocupam um espaço no mundo real ou virtual e são utilizados ou gerados por ações. Recursos e objetos podem ser identificados por substantivos e, portanto também podem ser lingüisticamente identificados por sintagmas nominais. O processo geral de extração de símbolos é baseado na extração de sintagmas nominais e é apresentado na Figura 1.

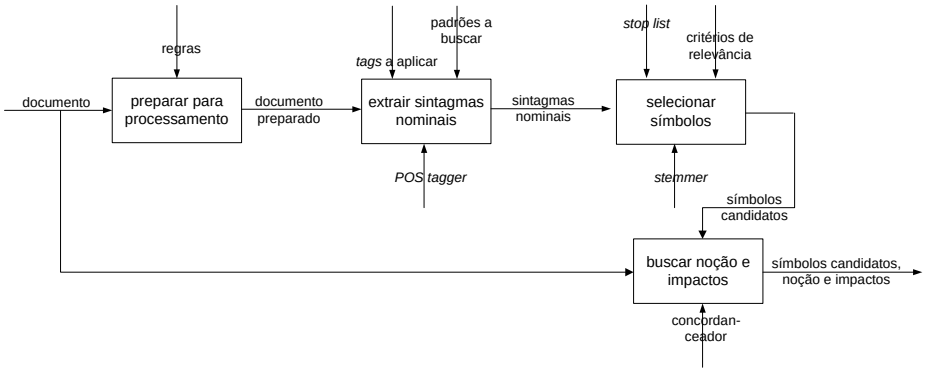

Figura 1 – Processo geral para extração de símbolos

3.1 Detalhamento do processo de extração de símbolos

A estratégia proposta é esquematizada utilizando uma perspectiva de processo, conforme Figura 1. O processo de extração de símbolos pode ser dividido em quatro atividades ou sub-processos. Inicialmente o documento é preparado para posteriormente ser manipulado por um *Part of Speech tagger* (*POS tagger*). Após a inserção das *tags* (etiquetas) são extraídos sintagmas nominais que correspondem a padrões pré-definidos; estes sintagmas nominais passarão por um processo de seleção, que irá considerar apenas aqueles que atendem a critérios de relevância estabelecidos. A etapa final extrai, para cada um dos sintagmas, contextos que possam vir a ser utilizados como noção e impactos. A seguir detalharemos cada um desses sub-processos.

3.1.1 Preparar para processamento

O trabalho de preparação dos documentos é necessário, pois as ferramentas utilizadas trabalham com arquivos do tipo texto puro. Os documentos da organização podem estar em diferentes formatos, como por exemplo, *pdf* (Portable Document Format), *doc* (Microsoft Word document) ou *rtf* (Rich Text Format). Documentos podem ainda incluir figuras e tabelas, que não serão processadas pelas ferramentas de tratamento de texto. O processo de preparação inicialmente retira figuras, transforma tabelas em texto, retira possíveis *tags* de formatação e gera texto puro, ou seja, arquivos do tipo *txt*.

Após a obtenção do arquivo em formato *txt*, faz-se necessário a *tokenização* do documento, ou seja, a colocação de um *token* por linha (inclusive dos caracteres de pontuação). A separação dos *tokens* é uma exigência do etiquetador utilizado e poderia ser dispensada caso a ferramenta utilizada não exigisse este formato.

3.1.2 Extrair sintagmas nominais

Sintagmas nominais possuem uma estrutura bem definida. Perini [Perini98] coloca que sintagmas nominais possuem duas estruturas básicas: a estrutura à esquerda do

núcleo do sintagma é composta por posições que podem ser ocupadas por determinantes, possessivos, quantificadores e outras classes de palavras. A estrutura à direita do núcleo é composta por modificadores, que por sua vez podem ser classes abertas ou mesmo outros sintagmas nominais. Neste trabalho não utilizamos a estrutura à esquerda do núcleo, pois não é usual que documentos organizacionais ou que utilizam linguagem técnica trabalhem com construções frasais mais sofisticadas, como usualmente acontece na literatura. Os padrões que estabelecemos para seleção de atores/sujeitos e objetos/recursos serão apresentados nas seções 3.2 e 3.3.

O processo de extração de sintagmas nominais é baseado na utilização de um etiquetador morfossintático, que analisa o texto pré-processado e associa uma etiqueta a cada um dos *tokens*. Após a etiquetagem são extraídos e contabilizados os sintagmas nominais que atendem a padrões pré-estabelecidos. Para a língua portuguesa existem alguns etiquetadores disponíveis, e para o nosso trabalho optamos pela utilização do QTAG, detalhado em [Mason97]. Apesar de outros etiquetadores possuírem precisão um pouco melhor [Lácio-Web06], priorizamos o QTAG pelo fato do texto etiquetado estar num formato semelhante ao XML, o que facilita a troca de informações entre diferentes aplicações.

O QTAG é um etiquetador probabilístico, desenvolvido originalmente para inserção de etiquetas em textos escritos na língua inglesa. A taxa de precisão das etiquetas inseridas é da ordem de 96,3% [Mason97]. A nova versão deste etiquetador, em linguagem Java, permite sua utilização em diversas outras línguas, através da construção de tabelas geradas pelo treino em corpora específico. Para a língua portuguesa, este etiquetador foi treinado com um corpus de aproximadamente 500 mil palavras, pelos pesquisadores T. Sardinha e R. Lima-Lopes, associados ao Lael/PUCSP. Conforme dados experimentais, a precisão deste etiquetador para textos em língua portuguesa é da ordem de 93% [Sardinha04]. O conjunto de etiquetas pode ser obtido em http://www2.lael.pucsp.br/corpora/etiquetagem.

3.1.3 Selecionar símbolos

O processo de seleção dos símbolos a partir dos sintagmas candidatos desconsidera sintagmas que contenham termos relacionados numa *stop list*. Esta *stop list* é composta prioritariamente por termos do domínio ou da língua geral que não são relevantes para o léxico. A etapa seguinte realiza a *stemização* dos sintagmas extraídos. Isto é necessário pois, nos nossos experimentos, identificamos a necessidade de agrupar singular e plural, feminino e masculino. Por exemplo, em nossos estudos de caso, foram separadamente extraídos e contabilizados os sintagmas nominais *gestor de escala* e *gestor de escalas*, o mesmo ocorrendo com os sintagmas *cessionária* e *cessionário*. Utilizamos o *stemmer* originalmente desenvolvido por V. Orengo [Orengo01] e modificado por M. A. L. Dias [Dias04]. Os sintagmas candidatos são agrupados e contabilizados após o processo de *stemização*, e adotamos o masculino singular para a representação, quando fosse o caso.

A última atividade do processo de seleção descarta os sintagmas cuja freqüência seja menor que quatro, conforme procedimento usual na área de extração de terminologia [Daile96]. Nos experimentos que realizamos, com documentos de diferentes tamanhos, o descarte de sintagmas com freqüência menor que quatro mostrou-se adequado aos nossos propósitos, porém estamos planejando a execução de

experimentos considerando diferentes valores visando confirmar ou modificar nossa escolha.

3.1.4 Buscar noção e impactos

A busca de contexto trabalha com cada um dos símbolos selecionados, varrendo o documento buscando identificar a presença do símbolo. Quando isto acontece, é extraído o parágrafo, pois este pode corresponder à definição ou a impactos do símbolo. Na identificação das sentenças é utilizado o radical (*stem*) do símbolo de busca; os parágrafos extraídos serão agrupados junto ao símbolo candidato, e utilizados posteriormente pelo engenheiro de requisitos para inserção no léxico da aplicação. Nos estudos de caso observamos que a identificação da noção (ou definição) dos símbolos é sensível ao contexto da aplicação. Esta etapa do nosso processo ainda está em fase de refinamento e não será detalhada neste artigo.

3.2 Extração de Sujeitos ou Atores: padrões utilizados

Na língua portuguesa funções ou papéis desempenhados por pessoas ou entidades são identificados por substantivos com terminações específicas. Alguns exemplos com terminação *ente*: gerente, presidente; terminação *or*: gestor, diretor, trabalhador; terminação *ário*: usuário, funcionário. Na extração de sintagmas nominais correspondendo a atores/sujeitos, utilizamos um conjunto de 82 padrões, pois para cada terminação consideramos o singular, o plural, feminino e masculino, na prática multiplicando cada terminação por quatro. Alguns exemplos de padrões utilizados são:

N*ente PRP* N* / N*entes PRP* N* / N*enta PRP* N* / N*entas PRP* N*
N*or PRP* N* / N*ores PRP* N* /N*ora PRP* N* /N*oras PRP* N*
N*eiro PRP* N*
N*ente N*/N*entes N* /N*enta N*/N*entas N*
N*or N*/N*ores N*
N*ora N*/N*oras N*

onde * indica qualquer quantidade de caracteres. O padrão **N*ente PRP N*** deverá extrair sintagmas compostos por um nome, preposição e nome, sendo que o primeiro nome deverá ter a terminação **ente**. Os padrões foram definidos empiricamente, após avaliação de textos etiquetados.

3.3 Extração de Objetos ou Recursos: padrões utilizados

A extração de recursos/objetos utilizou padrões mais genéricos que os utilizados na etapa de extração de atores/sujeitos. Foram utilizados nove padrões, também definidos empiricamente após avaliação de documentos etiquetados.

N* PRP* N* PRP* N*	N* PRP* N*	N* N*
N* PRP* N* N*	N* PRN* N*	N* PART*
N* CPR* N* N*	N* CPR* N*	N* ADJ*

4. Estudos de caso

Nossa proposta foi validada por dois estudos de caso, detalhados a seguir. Para apoio ao processo proposto, utilizamos um protótipo que consiste de programas escritos em Java, e uma biblioteca para tratamento de expressões regulares e padrões.

4.1 Estudo de caso: aplicação na área financeira

Este primeiro estudo de caso utilizou documentos públicos relacionados ao sistema SELIC, do Banco Central, disponíveis em http://www.bcb.gov.br/htms/spb/. Os documentos seguem uma estrutura padrão para o sistema SELIC. O total de palavras da porção de documentos avaliados é 5.461. A tabela 3 apresenta resultados parciais obtidos após a aplicação do processo proposto.

Tabela 3 - Símbolos do tipo sujeito e objeto para o sistema Selic

Sujeito	**Padrão**	**Freq.**	**Objeto**	**Padrão**	**Freq.**
cedente	N*ente	17	operação compromissada	N* PART*	35
clientes	N*entes	17	retorno de operação	N* PRP* N*	17
cessionário	N*ário	15	conta reservas	N* N*	16
correspondente	N*ente	10	pu de retorno	N* PRP* N*	16
instituição financeira	N* ADJ*	9	número de operação	N* PRP* N*	10
usuário	N*ário	9	reservas bancárias	N* N*	10
cedente da operação	N*ente CPR* N*	6	situação da operação	N* CPR* N*	9
cessionário da operação	N*ário CPR* N*	6	conta da instituição	N* CPR* N*	7
destinatário	N*ário	6	operação 1054/s	N* N*	7
emissor	N*or	6	pu da operação	N* CPR* N*	7
instituição liquidante	N*ão N*	6	endereço eletrônico	N* ADJ*	6

4.2 Estudo de caso 2: aplicação na área de recursos humanos

Este estudo de caso utilizou um documento de requisitos para um sistema de informação na área de recursos humanos. O documento segue padrões internos à organização, modelando requisitos com casos de uso e especificações suplementares. O total de casos de uso, neste documento, é de 60, e são 5 as especificações suplementares, relacionadas a documentação, treinamento e usabilidade. O total de palavras, neste segundo documento, é de 6.665. A tabela 4 apresenta parte do conjunto de sujeitos e objetos selecionados.

Tabela 4 - Símbolos do tipo sujeito e objeto para o sistema de gestão de escalas de trabalho

Sujeito	Padrão	Freq.
gestor	N*or, N*ores	101
usuário	N*ário	88
gestor de escalas	N*or PRP* N*	68
usuário final	N*ário N*	58
funcionário	N*ário, N*ários	39
administrador	N*or	28
colaborador	N*or, N*ores	22
administrador central	N*or N*	12
leitor	N*or	8
empresa	N*esa	7
administrador local	N*or N*	7

Objeto	Padrão	Freq.
banco de dados	N* PRP* N*	11
cálculo de estimativas	N* PRP* N*	10
sistema operacional	N* N*	8
postos de serviços	N* PRP* N*	7
parâmetros de pesquisa	N* PRP* N*	7
solicitações de processamento	N* PRP* N*	6
lista de postos de serviços	N* PRP* N* PRP* N*	6
sistema administrador	N* N*	5
relatório de ocorrencias	N* PRP* N*	5
log das ocorrencias	N* CPR* N*	5
habilita relatório	N* N*	5

4.3 Avaliação dos resultados preliminares

As taxas de *recall* e *precision* dos resultados preliminares estão sumarizadas na tabela 5. As avaliações foram executadas por equipes independentes, uma delas composta por especialistas do Banco central (estudo de caso 1) e a outra por engenheiros de software (estudo de caso 2). Mesmo tendo efetuado um número pequeno de estudos de caso, as taxas de *recall* e *precision* mostram a adequação da proposta para a construção do léxico para aplicações. Estamos iniciando novos experimentos com a aplicação da estratégia a documentos de diferentes domínios de aplicação visando obter um conjunto significativo de resultados para análise. Por outro lado, alguns aspectos da proposta estão em processo de revisão, visando melhorar os resultados já obtidos.

Tabela 5 - Valores de *recall* e *precision* obtidos nos estudos de caso

	ec1: suj	ec1: obj	ec2: suj	ec2: obj
recall	92%	100%	78%	78%
precision	75%	78%	84%	56%

O algoritmo de extração de sintagmas nominais está em processo de revisão de forma a evitar que sintagmas que atendem a mais de um padrão sejam recuperados em duplicidade. Algumas modificações nos padrões utilizados estão sendo experimentadas, visando à obtenção de melhores taxas de *recall* e *precision* em relação aos objetos recuperados. Também observamos alguns problemas decorrentes da inserção de etiquetas incorretas, principalmente em relação a objetos, pois estes utilizam padrões mais gerais que aqueles utilizados para sujeitos. Este é o caso do objeto *habilita relatório*, apresentado na tabela 4, onde a palavra *habilita* foi incorretamente etiquetada como Nome, quando a correta é Verbo. Uma análise mais detalhada das causas dos insucessos está programada para ser realizada após os ajustes já citados e após um conjunto maior de experimentos ser realizado.

5. Trabalhos relacionados

No contexto de processamento de linguagem natural e recuperação de informação, o uso de sintagmas nominais para recuperação de informações é encontrado em muitos trabalhos. [Parreiras03] propõe seu uso em indexação de textos científicos e [Perez03] os utiliza para a obtenção de conceitos que irão compor mapas conceituais. No contexto do processo de requisitos, trabalhos orientados à exploração de aspectos da linguagem natural ainda não são muito freqüentes, e selecionamos os descritos em [Harmain00] e [Boyd05], por buscarem diferentes abordagens no tratamento de sintagmas nominais em artefatos de requisitos.

Em [Harmain00] é descrito o CM-Builder, uma ferramenta *case* que avalia documentos de requisitos escritos em linguagem natural (língua inglesa) buscando relações semânticas nas sentenças de requisitos e visando á construção de um modelo inicial de classes em UML; as classes são derivadas de sintagmas nominais. O resultado inclui classes, atributos e relacionamentos, agrupados em modelos de classes e armazenados em arquivos cujo formato é adequado à manipulação posterior por ferramentas normalmente utilizadas nas etapas de projeto do sistema. Os autores enfatizam que os resultados devem ser vistos como apoio ao trabalho do analista ou engenheiro de software, devendo ser refinados para efetivamente contribuir para um modelo inicial de classes.

A abordagem proposta por Boyd et al [Boyd05] enquadra-se na linha de trabalho que utiliza um subconjunto controlado da linguagem natural para o registro de requisitos, com o objetivo de reduzir a ambigüidade. Esse trabalho investiga a expressividade sintática e semântica de um sub-conjunto da língua inglesa para uso em documentos de requisitos, focando especificamente nos verbos. Para corrigir falhas na identificação de entidades (recuperadas via sintagmas nominais) por parte do etiquetador utilizado, foi utilizado um dicionário gerado a partir de sintagmas nominais relacionados num glossário da aplicação.

Nosso trabalho aproxima-se do trabalho de [Harmain00], mas nossos objetivos são diferentes: visamos apoiar o processo de requisitos, enquanto eles buscam apoiar o processo de projeto do sistema. Utilizamos como fontes de informação não só documentos de requisitos, mas também outros documentos que sejam manipulados ou gerados no processo de requisitos – basta que tais documentos utilizem a linguagem natural. Enquanto a abordagem descrita em [Boyd05] utiliza um sub-conjunto restrito da língua natural, nossa abordagem não restringe o uso da língua portuguesa nos documentos que manipula. Nosso objetivo é identificar atores/sujeitos e recursos/objetos relevantes, de forma a apoiar a construção ou atualização do léxico da aplicação. O processo criado também é útil na construção de um léxico para o domínio da organização, pois em organizações utilizando desenvolvimento distribuído de software dificuldades derivadas de diferentes capacidades lingüísticas, diferenças culturais e *delays* de comunicação tornam mais presente a necessidade de um léxico abrangente.

Nossa abordagem utiliza também uma *stop list* constituída de termos não relevantes para o domínio considerado, permitindo ajustes nos termos a serem extraídos e gerando resultados mais precisos na avaliação futura de documentos de um mesmo domínio. Assim como enfatizado por Harmain e Gaizauskas [Hairmain00], consideramos que os resultados obtidos não prescindem do processo de

revisão e avaliação por especialistas do domínio; os sujeitos e objetos extraídos através do processo proposto constituem um subsídio importante para a construção ou atualização do léxico da aplicação.

6. Conclusões e trabalhos futuros

Neste artigo apresentamos um processo para a identificação automática de símbolos para a construção de um léxico para o domínio da aplicação. A extração é baseada na identificação de sintagmas nominais, através do uso de padrões identificados empiricamente em documentos manipulados no processo de requisitos. As etapas necessárias ao processo iniciam pela *tokenização* do documento, passam pela inserção de etiquetas que identificam a categoria da palavra e pela extração e consolidação de sintagmas nominais que atendem a padrões pré-definidos. Símbolos que atendem a critérios simples de freqüência são selecionados para compor o léxico da aplicação, e para cada um desses símbolos são ainda extraídos os contextos que os contém, para que o engenheiro de requisitos possa obter a definição e impactos desse símbolo.

Realizamos dois estudos de caso, um deles num documento de requisitos cedido por uma companhia que utiliza desenvolvimento distribuído de software; os participantes do processo de requisitos são geograficamente separados, o que aumenta a importância do léxico para compartilhamento de um mesmo entendimento a respeito dos símbolos próprios do domínio da aplicação. O outro estudo de caso utilizou documentos extraídos de um manual de usuário, mostrando que a estratégia é aplicável a diferentes tipos de documentos manipulados no processo de requisitos. A estratégia pode ser utilizada com outras finalidades, por exemplo, para buscar atores relevantes e definir classes candidatas para o processo de projeto.

As próximas etapas deste trabalho envolvem a realização de novos experimentos, visando avaliar a aplicação da proposta a documentos de diferentes tipos e domínios. Após isto está programado o refinamento do processo de obtenção da noção e impactos dos símbolos extraídos para o léxico. A etapa final está relacionada à estruturação de todo o processo em forma de serviços dos agentes Construtor do Léxico e Analisador Léxico, referidos na introdução deste artigo.

Este trabalho evidencia também que a maturidade já alcançada por métodos e técnicas da área de Processamento da Linguagem Natural possibilita sua aplicação no processo de desenvolvimento de software. A disponibilidade atual de um número maior de ferramentas que trabalham com a língua portuguesa contribui para isto, o que foi possível verificar neste trabalho.

Agradecimentos

Agradecemos à Tlantic Sistemas de Informação, pela cessão dos documentos de requisitos e aos revisores anônimos, cujas sugestões e questionamentos contribuíram para a melhoria deste trabalho. Também agradecemos a Will Lowe, pelo concordanceador do Yoshikoder.

Referências Bibliográficas

[Booch00] Booch, G. ; Rumbaugh, J. & Jacobson, I. "UML: guia do usuário". Rio de Janeiro: Campus, 2000. 472 p. ISBN 8535205624

[Boyd05] Boyd, S.; Zopwghi, D. & Farroukh, A. "Measuring the expressiveness of a Constrained Natural Language: an Empirical Study". In: 13th IEEE International Conference on Requirements Engineering (RE´05). Proceedings.

[Daile96] Daille, B. "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology". In: Klavans, J., Resnik, P. The Balancing ACT- Combining Symbolic and Statistical Approaches to Language, The MIT Press, 1996. pp. 49-66.

[Dias04] Dias, M. A. L. "Extração Automática de Palavras-Chave na Língua Portuguesa Aplicada a Dissertações e Teses da Área das Engenharias". Dissertação de Mestrado. Campinas: FEEC-UNICAMP, 2004.

[Harmain00] Harmain, H. M. & Gaizauskas, R. "CM-Builder: An Automated NL-based CASE Tool". In: 15th IEEE International Conference on Automated Software Engineering (ASE'2000), 2000. Proceedings. pp. 45-53.

[Lácio-Web06] Projeto Lácio-Web, disponível em http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm. Acesso em 28.07.06.

[Leite90] Leite, J.C.S.P.; Franco, A. P. "O uso de hipertexto na elicitação de linguagens de da aplicação". In: 4º Simpósio Brasileiro de Engenharia de Software, 1990. Anais. pp.124-133.

[Leite93] Leite, J.C.S.P. & Franco, A.P.M. "A Strategy for Conceptual Model Acquisition". In: First IEEE International Symposium on Requirements Engineering, San Diego, Ca, IEEE Computer Society Press, 1993. Proceedings. pp. 243-246

[Liberato97] Liberato, Y. G. "A estrutura do SN em português: uma abordagem cognitiva". Tese de doutorado, 1997. UFMG, Departamento de Lingüística, Belo Horizonte.

[Mason97] Mason, O. & Tufis, D. "Probabilistic Tagging in a Multi-lingual Environment: Making an English Tagger Understand Romanian". In: Third European TELRI Seminar, Montecatini, Italy, 1997. Proceedings.

[Orengo01] Orengo, V. M., Huyck, C. "A Stemming Algorithm for Portuguese Language". In: 8th Symposium on String Processing and Information Retrieval (SPIRE 2001), Laguna de San Raphael , Chile, (2001). Proceedings. pp. 186-193.

[Parreiras03] Parreiras, F. "O uso de sintagmas nominais como fonte de descritores para textos de periódicos científicos". Escola de Ciência da Informação. Belo Horizonte, 2003. Disponível em http://www.fernando.parreiras.nom.br/publicacoes/sn.pdf.

[Pérez03] Pérez, C. C. C.; Gasperin, C. & Vieira, R. "Extração semi-automática de conhecimento a partir de textos". In: IV Encontro Nacional de Inteligência Artificial (ENIA 2003), Campinas, 2003. Anais da SBC, 2003, v. 7, pp.193-202.

[Perini98] Perini, M. A. "Gramática descritiva do português". 3ªed. - São Paulo: Ática, 1998. 380p. ISBN 8508055501

[Sardinha04] Sardinha, A. P. B. "Lingüística de Corpus". São Paulo: Ed. Manole, 2004. 410 p.

[Sayão05] Sayão, M. & Leite, J. C. S. P. "Uso de Agentes no Processo de Requisitos em Ambientes Distribuídos de Desenvolvimento". In: Workshop de Engenharia de Requisitos, Lisboa, Portugal, 2005. Anais.

[Vieira01] Vieira, R. & Lima, V. L. S. (2001). "Lingüística Computacional: Princípios e Aplicações". In: As Tecnologias da Informação e a questão social: anais. Carlos Eduardo Ferreira (Ed.) Fortaleza, SBC. ISBN 85-88442-03-5 (v.2). pp 47-88.