OntoLP: Engenharia de Ontologias em Língua Portuguesa

Luiz Carlos Ribeiro Junior[1], **Renata Vieira**[1]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

lucarijr@gmail.com, renata.vieira@pucrs.com.br

Abstract. *The continuous growth of digital information resources of many kinds (texts, images, videos, services) points to the need of consistent knowledge representation structures for modeling, storing, accessing and communicating about these resources. In this context, ontologies are being claimed as an important technology. Ontologies wide adoption is, however, difficult due to the large cost of building them from scratch. This problem instigate a new area of research related to ontology learning from texts. For this kind of problem, many of the tools and methods needed are language dependent, since they rely on linguistic knowledge. Efforts are needed for each language. This paper presents a tool (OntoLP) developed as a plug-in to the ontology editor Protégé which analyzes a given domain corpus of Portuguese texts and suggests concept candidates and their hierarchy to the ontology engineer, based on the knowledge presented in the texts.*

Resumo. *O crescimento das bases digitais e fontes de dados de diversas naturezas (textos, imagens, vídeos, serviços) tem apontado para a necessidade de estruturas mais consistentes de representaçãao do conhecimento para modelar, armazenar, acessar, e comunicar sobre esses recursos. Nesse contexto, as ontologias aparecem como a tecnologia apropriada ao problema. A principal dificuldade na sua utilização em larga escala está no seu processo de construção extremamente custoso. Essa característica estimula pesquisas visando automatizar a tarefa, as quais, em sua maioria, consideram as bases textuais como fontes de conhecimento. As ferramentas e métodos são, nesse caso, dependentes de idioma, pois baseiam-se fortemente no uso de informações linguísticas. Sendo assim, para viabilizar a utilização de ontologias em larga escala, são necessários estudos específicos para cada língua. Este trabalho apresenta a ferramenta OntoLP, desenvolvida como um plug-in para o ambiente Protégé, que faz a análise de um corpus de domínio em língua portuguesa e sugere candidatos a conceitos e hierarquias ao engenheiro de ontologia com base no conhecimento representado nos textos.*

1. Introdução

Em diversas áreas relacionadas à tecnologia de informação, constata-se o crescimento tanto no interesse científico-tecnológico, como na necessidade prática de uso intensivo de meta-dados na organização semântica de recursos, documentos e serviços. Isso é percebido tanto em contextos restritos (ambientes corporativos) como em contextos mais amplos (web). Uma das tecnologias fundamentais para o desenvolvimento dessa área são as ontologias. As ontologias são responsáveis pela organização semântica de meta-dados

e possibilitam uma estruturação lógica de conteúdos que podem ser manipuladas de forma inteligente.

A pesquisa nessa área tem apresentado um grande interesse da comunidade científica e ganhou novo fôlego com o projeto da Web Semântica. As questões de pesquisa relacionadas a ontologias são variadas, perpassando diversas áreas (tanto dentro como fora da computação). Essas questões são importantes para o crescente conjunto de aplicações baseadas em conhecimento, tais como, Pesquisa Semântica para Multimídia, Sistemas de Perguntas e Respostas, Serviços Web e Web Semântica.

A vasta gama de aplicações é uma das razões do grande interesse por métodos de construção, manutenção, mapeamento e aprendizado de ontologias. O desenvolvimento dessa área no Brasil, não pode deixar de considerar a importância das questões relacionadas à língua portuguesa. A maioria dos trabalhos nessa área, mesmo que desenvolvida por pesquisadores brasileiros, tomam como base ontologias em língua inglesa, pela simples razão da carência de ontologias na língua portuguesa. Preencher essa lacuna é necessário, e esse projeto se propõe a colaborar para o desenvolvimento nessa direção.

Entendemos que o problema de construção de ontologias em língua portuguesa tem relação com os seguintes desafios propostos no documento "Grandes Desafios da Computação no Brasil: 2006-2016":

a) Gestão da Informação em grandes volumes de dados multimídia distribuídos (ontologias como ferramenta de meta-dados e indexação semântica de dados multimída);

b) Modelagem computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem-natureza (ontologias como ferramenta de modelagem de domínios);

d) Acesso participativo e universal do cidadão brasileiro ao conhecimento (ontologias como ferramenta auxiliar na comunicação);

Nesse contexto necessitamos de soluções que são dependentes de língua. Não poderemos embasar gestão da informação, modelagem de sistemas complexos e a comunicação com o cidadão, apenas com soluções desenvolvidas para a língua Inglesa. Reconhecemos que parte do que é desenvolvido na área de processamento de linguagem natural e recuperação de informações é independente de língua (soluções por métodos estatísticos), mas há uma forte parcela do desenvolvimento dessa área que não o é.

A disponibilidade de ontologias e bases de dados terminológicos e semânticos para embasar a pesquisa nessa área é essencial. Para dar um exemplo da abrangência e relevância dessa área, veja-se o esforço do governo no desenvolvimento da "Lista de Assuntos do Governo" (LAG[1]):

"Segundo a Organization for the Advancement of Structured Information Standards (OASIS) os governos de todas as esferas no mundo são os maiores produtores e consumidores de dados e informações. Órgãos do governo disponibilizam informações e serviços em portais e sítios da web, porém o grande volume e a complexidade da estrutura governamental podem tornar a localização da informação uma tarefa difícil, até mesmo impossível para os cidadãos. A LAG, foi criada para ajudar os cidadãos a en-

[1]Disponível em https://www.governoeletronico.gov.br/anexos/lista-de-assuntos-do-governo-lag-v1.0

contrar informações, independentemente do conhecimento da estrutura do governo ou de qual órgão o assunto é responsabilidade. O foco da LAG é o CIDADÃO. O esquema tem por objetivo ser intuitivo para os cidadãos que buscam assuntos do seu interesse na larga faixa de informações do setor público. Portanto a LAG: a) prefere a linguagem do leigo ao jargão do serviço público ou termos técnicos; b) não supõe que o cidadão tenha conhecimento prévio das responsabilidades de cada nível ou órgão governamental. Procura ser independente da estrutura governamental, devendo sobreviver às mudanças de estruturas e organogramas; c) o uso comum é mais importante do que a precisão acadêmica, quando se está escolhendo nomes ou posições relativas aos cabeçalhos.

Para que a LAG consiga cumprir o seu objetivo, deve ser constantemente atualizada. Sugestões, correções e críticas para melhorá-la serão sempre bem-vindas."

Essas questões podem ser entendidas no contexto dos desafios mencionados acima, padronização de vocabulário e mais do que isso, definição clara de conceitos. Portanto, ontologias são e, ao que tudo indica, serão cada vez mais importantes ao uso prático de aplicações relacionadas a gestão de informação, modelagem de domínios complexos, muitas vezes com particularidades nacionais como a ecologia ou meio ambiente, e na relação com o cidadão de forma geral.

A construção de ontologias é conhecidamente um processo demorado e difícil, que muitas vezes requer constante atualizacão. Por isso pesquisadores da área vêm estudando formas de automatizar a tarefa. Pesquisas nesse sentido possuem como ponto fundamental a fonte de conhecimento utilizada na execução do processo. Muitos dos esforços realizados atualmente na área propõem a construção de ontologias a partir de um conjunto de textos de um domínio específico ([Suchanek et al. 2006, Baségio 2006, Ryu and Choi 2006]), visto que a linguagem é a primeira forma de transferência de conhecimento entre os seres humanos.

Nessa linha, este trabalho propõe e avalia métodos de extração de termos candidatos a conceitos de textos da língua portuguesa, com objetivo de auxiliar engenheiros de ontologias especialmente dedicados a desenvolverem ontologias em língua portuguesa. Os métodos desenvolvidos constituem a ferramenta OntoLP, implementado como um *plug-in* para o editor de ontologias Protégé, um ambiente bastante utilizado na comunidade científica e que dá suporte à construção de ontologias seguindo as tecnologias da Web Semântica, como por exemplo, a construção de ontologias OWL *Web Ontology Language*, conforme o padrão definido pelo *World Wide Web Consortium* (W3C)[2].

O trabalho está organizado da seguinte forma, a Seção 2 descreve a ferramenta OntoLP e a metodologia proposta para a extração de termos. Na Seção 3 são apresentados os experimentos de avaliação dessa ferramenta. Na Seção 4, apresentamos a conclusão do trabalho.

2. A ferramenta OntoLP

As abordagens atuais de construção de ontologias a partir de textos baseiam-se fortemente no uso de informações lingüísticas, característica que as torna dependentes de idioma. Sendo assim, para que todos tenham acesso aos benefícios da utilização de ontologias em

[2]http://www.w3.org/

larga escala, são necessários estudos específicos para cada língua. Nesse sentido, quando comparado com outros idiomas, principalmente o inglês, pouco foi feito para o português.

O processo de construção automática de ontologias divide-se basicamente em cinco etapas ([Buitelaar et al. 2005]): 1) Extração de Termos Candidatos a Conceitos de um domínio; 2) Identificação da Relação Hierárquica entre os Termos; 3) Identificação de Relações Não-Hierárquicas; 4) Identificação de Instâncias e 5) Extração de Regras (Axiomas). Nesse processo, a Extração de Termos é considerada uma tarefa central a qualquer abordagem de construção de ontologias, visto que termos são realizações lingüísticas de conceitos de uma área específica, sendo vitais para as fases seguintes.

Existem três principais abordagens para a identificação de termos: estatística, lingüística e híbrida. Na primeira, cada documento pertencente ao corpus é considerado simplesmente como um vetor de termos e sua freqüência de ocorrência. Na segunda é necessário que os textos estejam anotados com informações lingüísticas. Na última abordagem é feito o casamento entre as duas metodologias anteriores.

O processo de extração de termos da ferramenta de auxílio a construção de ontologias, OntoLP, é baseado numa série de métodos híbridos, incluindo informações semânticas prototípicas para auxiliar o processo.

A Figura 1 apresenta uma visão geral da arquitetura do módulo de extração de termos. A extração é realizada em duas etapas:

- CorpusXCES: leitura do Corpus previamente anotado com informações linguísticas pelo parser PALAVRAS ([Bick 2000]), contendo informações morfológias, sintáticas e semânticas, representadas no formato XCES/PLN-BR. O corpus é utilizado como fonte de conhecimento para a construção da ontologia.
- Extração de Termos: nesta etapa são aplicados diferentes métodos, visando à extração de termos simples (unigramas) e complexos (n-gramas, onde n>1). Para auxiliar a extração dos termos foi adicionado um módulo anterior baseado em informações semânticas, originalmente proposto neste trabalho.

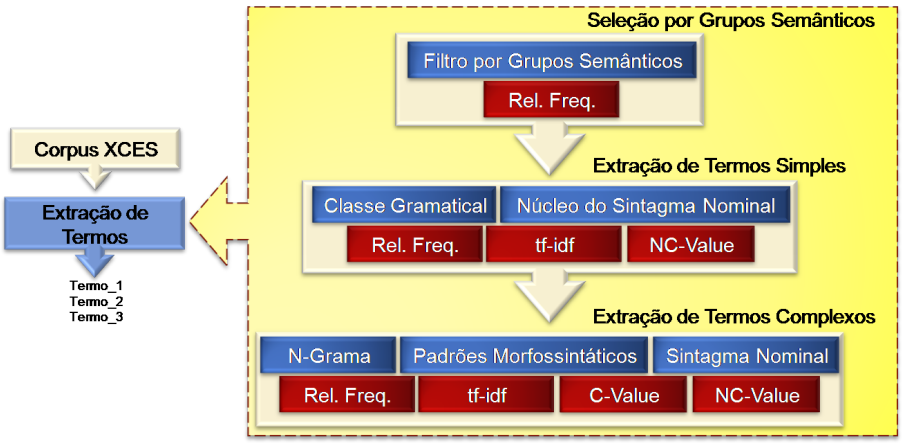

Figura 1. Arquitetura geral do módulo de extração de termos candidatos a conceitos

No plug-in OntoLP foram somados aos métodos de extração um conjunto de funcionalidades que visam auxiliar o engenheiro de ontologias nas fases que necessitam de

interação com a ferramenta. A interface de extração divide-se em três partes, relacionadas as sub-etapas da extração de termos apresentada na Figura 2.

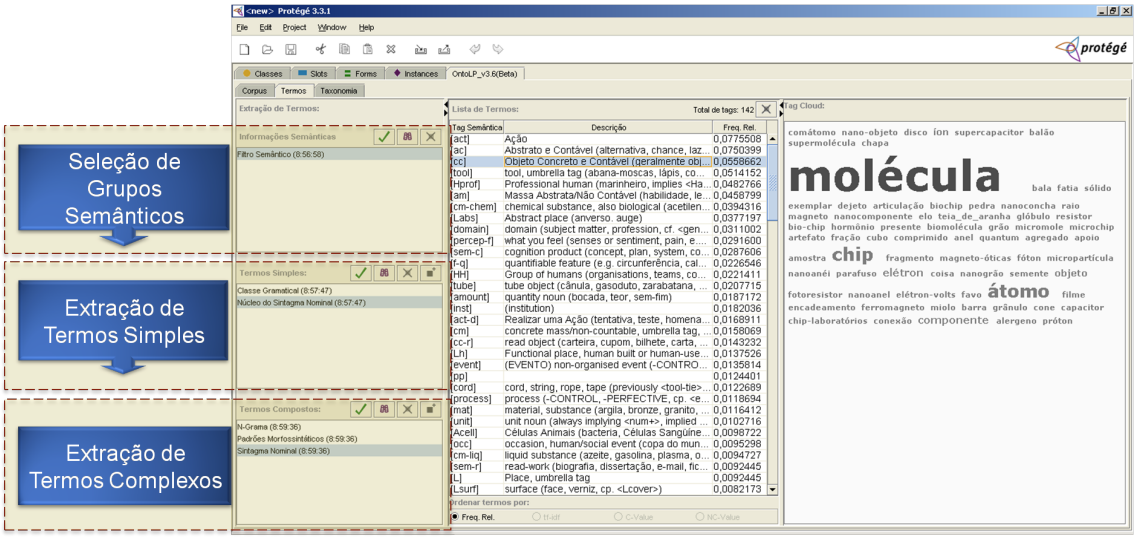

Figura 2. Mapeamento da Interface de Extração de Termos e das sub-etapas propostas para a tarefa

Nas seções 2.1 e 2.2 são detalhadas essas sub-etapas.

2.1. Seleção de Grupos Semânticos

Uma técnica bastante comum de ser aplicada a sistemas de extração de termos é a remoção de *stopwords*[3]. Esse processo costuma apresentar melhoras significativas nos resultados dos métodos. Entretanto, a construção de uma lista de *stopwords* pode constituir um passo manual extra indesejável na tarefa de extração.

A etapa de Seleção de Grupos Semânticos (Figura 2) proposta aqui baseia-se nas informações semânticas disponibilizadas pelo PALAVRAS. Essas são informações prototípicas que classificam nomes comuns em classes gerais, por exemplo, a *tag* "<an>" atribuída ao substantivo "olho", indica que a palavra pertence à classe "Anatomia". Dessa forma, os substantivos etiquetados com uma mesma *tag* são agrupados em conjuntos semânticos, por exemplo:

- Grupo <an> (Anatomia): {testa, ouvido, neurônio, cintura, olho, pé, mão}.
- Grupo <L> (Lugar): {território, campo, mundo, zona, área, rio, mar}.
- Grupo <H> (Humano): {pessoa, assassino, vítima, torcedor, arqueólogo}.

Os grupos semânticos podem ainda apresentar subdivisões, como pode ser observado no exemplo da Figura 3: 1) lugares e lugares relacionados a água; 2) anatomia e anatomia de movimento e 3) humano e humano profissional.

Com base nesses conjuntos foi criado o método Filtro por Grupo Semântico, que emprega os seguintes passos: a) Extração dos Grupos Semânticos: as *tags* semânticas presentes no corpus de entrada são extraídas; b) Cálculo de relevância dos Grupos

[3] *Stopwords* trata-se de uma lista de termos que não possuem relevância semântica. Geralmente é composta por classes gramaticais como artigos, preposições e advérbios.

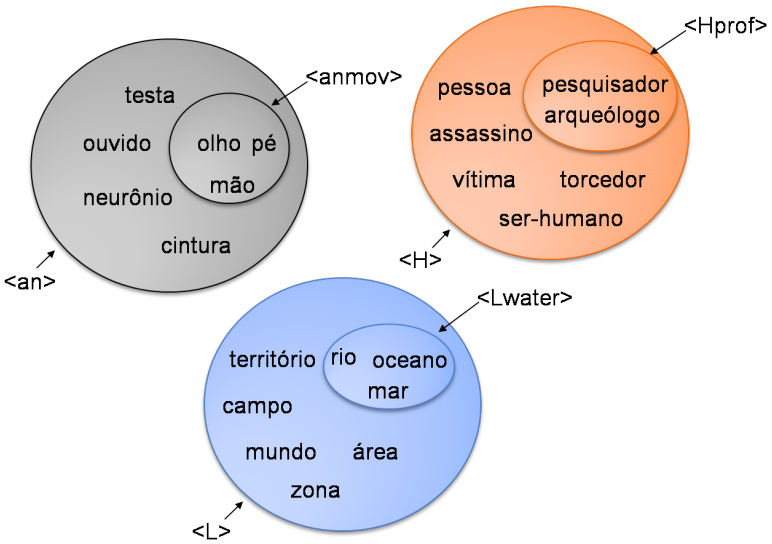

Figura 3. Exemplos de grupos e sub-grupos semânticos

Semânticos: o cálculo de freqüência relativa (FR) é aplicado a lista de *tags* semânticas extraída anteriormente. A lista é apresentada ao engenheiro ordenada conforme essa medida; c) Exclusão dos Grupos Irrelevantes: o engenheiro exclui os grupos semânticos que considera não ter relação com o domínio representado pelo corpus de entrada.

Quando o engenheiro exclui um determinado grupo estará automaticamente eliminando todos os termos pertencentes a ele. Sendo assim, esse método, através de uma abordagem semi-automática, possibilita a construção de uma lista de *stopwords* específica para um domínio. A seleção correta dos grupos depende do conhecimento do engenheiro de ontologia sobre a área, a ferramenta auxília o engenheiro ao mostrar as ocorrências dos termos de cada grupo e sua relevância pelo método de "*tag clouds*"[4], Figura 4. Nesse exemplo são mostrados os termos relacionados à etiqueta Hprof (profissão humana), destacada em (1), como pode se observar, em relação a essa etiqueta os termos pesquisador e cientista são os mais freqüentes, conforme (2) na figura. Cabe salientar que o corpus considerado neste exemplo era constituído de artigos do caderno de Ciências da Folha de São Paulo, por isso as profissões pesquisador e cientista aparecem com maior evidência.

2.2. Extração de Termos Simples e Complexos

A etapa de extração de termos foi dividida em identificação de termos simples e complexos. Em ambos os casos foram implementados métodos híbridos para a execução da tarefa. A Tabela 1 descreve resumidamente cada método e sua execução.

Para o cálculo de relevância dos termos foram utilizadas quatro medidas estatísticas: FR e *tf-idf*, descritas em [Manning and Schütze 1999], *NC-Value* e *C-Value*, propostas por [Frantzi et al. 1998]. A métrica *C-Value* é exclusiva para o cálculo da relevância de termos complexos, enquanto as demais são aplicadas tanto para termos simples quanto complexos.

Durante o processo de extração os métodos recebem a lista de Grupos Semânticos gerada na etapa anterior e percorrem o corpus selecionando os termos considerados aptos

[4]http://en.wikipedia.org/wiki/Tag_cloud

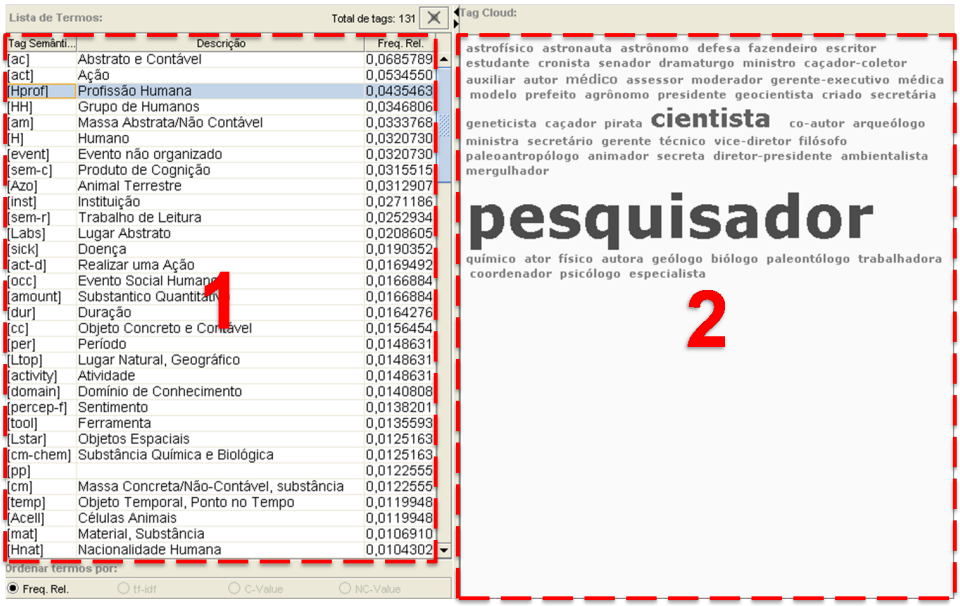

Figura 4. *Tag clouds* relacionada ao grupo semântico Hprof (humano-profissão)

Tabela 1. Métodos de extração de termos implementados

Método	Execução	Tipo de Termo
Classe Gramatical	Extrai classes gramaticais definidas pelo engenheiro	Simples
Núcleo do Sintagma Nominal	Extrai apenas termos considerados núcleo de um SN	Simples
N-Grama	Extrai conjuntos de palavras de tamanho 'n'	Complexos
Padrões Morfossintáticos	Extrai padrões morfossintáticos, semelhante a expressões regulares	Complexos
Sintagma Nominal	Extrai somente termos que constituem SN's	Complexos

e que pertençam a pelo menos um grupo presente na lista de entrada. A lista de termos extraída pelos métodos é submetida às medidas de relevância. Após o cálculo, os termos são re-organizados em ordem decrescente conforme essas medidas. A lista final pode ser editada pelo engenheiro.

3. Experimentos

Os experimentos realizados neste trabalho foram divididos em duas partes: (1) Avaliação dos Métodos de Extração de Termos e (2) Avaliação do Uso de Informações Semânticas na Extração de Termos. Na primeira foram analisados quais os melhores pares de métodos/medidas em termos de Precisão (P), Revocação (R) e *F-measure* (F). Para isso foi necessário um conjunto de recursos constituído por um corpus de domínio e uma lista de termos de referência, nesse caso da área de Ecologia ([Zavaglia et al. 2007]). O corpus é constituído por textos extraídos de parte dos livros "A Economia da Natureza" e "Ecologia", além de revistas presentes no projeto LácioWeb[5]. O CórpusEco conta com um total de 260.921 palavras. A construção da lista de termos foi feita através de critério semântico (manualmente), sendo extraídos 694 termos. Além disso, foram utilizados dois glossários especializados e mais 1105 termos extraídos do Dicionário On-Line do Jornal do Meio Ambiente[6]. Finalmente, foram eliminados os termos duplicados nas listas e

[5]http://www.nilc.icmc.usp.br/lacioweb/
[6]http://www.jornaldomeioambiente.com.br/

feita a intersecção com o CórpusEco, restando um total de 520 termos divididos em 322 unigramas, 136 bigramas e 62 trigramas.

A segunda etapa avaliou o impacto da utilização dos Grupos Semânticos nos pares que obtiveram melhores resultados no primeiro experimento. Nessa avaliação foram convidados dois grupos de pesquisa com experiência em extração de termos para participar do processo. Os recursos utilizados para esse experimento são descritos na Seção 3.2.

3.1. Avaliação da Etapa de Extração dos Termos

Nesse experimento foram consideradas todas as combinações possíveis entre as técnicas disponíveis no plug-in (métodos/medidas). Para a avaliação foi utilizado um corte que seleciona apenas os mil primeiros termos. Esse corte foi selecionado pois seus valores se aproximam da *F-measure* média calculada para diferentes faixas de corte. Os resultados obtidos são demonstrados na Tabela 2, considerando listas de uni, bi e trigramas. Na tabela, além do valor obtido por cada métrica, são apresentados o total de termos corretos extraídos pelos pares método/medida (Corretos), considerando o corte aplicado, e o total de termos na lista de referência (Ref.). O total geral de termos extraídos e corretos são exibidos junto ao nome do método.

Para a extração de unigramas houve um empate entre quatro pares: Classe Gramatical/*NC-Value(tf-idf)*, Núcleo do SN/FR, Núcleo do SN/*NC-Value(tf-idf)* e Núcleo do SN/*NC-Value*(FR). Para demonstrar o desempenho desses pares de uma forma mais ampla, o gráfico 5 apresenta a *F-measure* obtida por cada um em diferentes faixas de termos. Conforme é possível perceber, no geral o par Classe Gramatical/*tf-idf* foi o que obteve os melhores resultados.

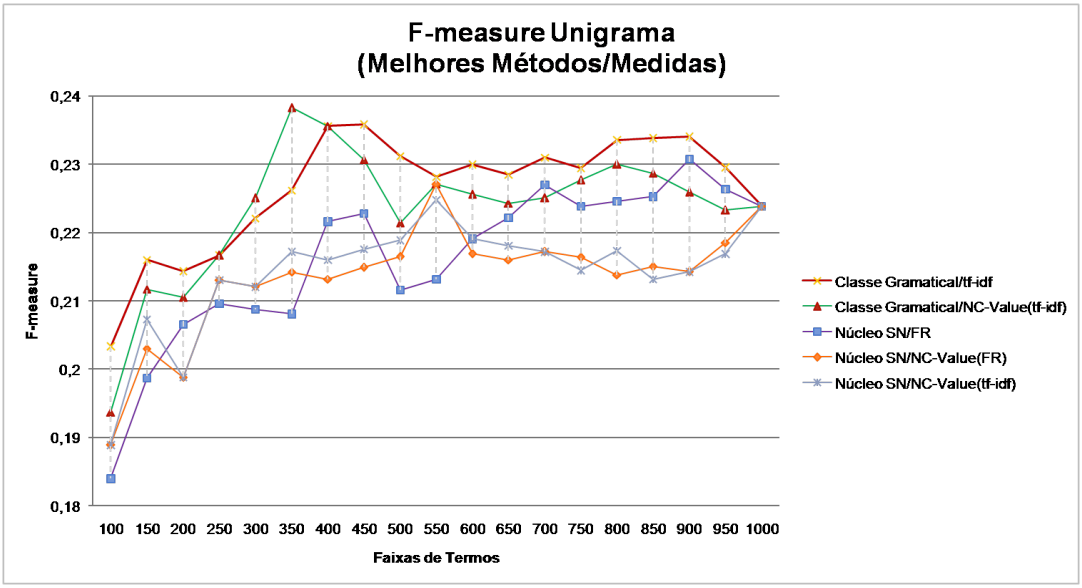

Figura 5. Resultados obtidos pelos métodos para diferentes faixas de termos durante a extração de unigramas (*F-measure*)

Com relação aos bigramas, os melhores pares foram: Padrões Morfossintáticos/FR e Padrões Morfossintáticos/*C-Value*. Nesse caso, cabe salientar que a medida *C-Value* foi projetada para listas com variações no tamanho dos termos. Quando

Tabela 2. Resultados para a extração de unigramas considerando uma faixa de 1000 termos

Método (total/corretos)	Medida	Corretos	P	R	F	Corte	Ref.
Classe Gramatical (5742/272)	tf-idf	145	14,5%	46,33%	22,09%	1000	322
	FR	146	14,6%	46,65%	22,24%		
	NC-Value(tf-idf)	**147**	**14,7%**	**46,96%**	**22,39%**		
	NC-Value(FR)	146	14,6%	46,65%	22,24%		
Núcleo do Sintagma Nominal (4527/258)	tf-idf	*142*	*14,2%*	*45,37%*	*21,63%*		
	FR	**147**	**14,7%**	**46,96%**	**22,39%**		
	NC-Value(tf-idf)	**147**	**14,7%**	**46,96%**	**22,39%**		
	NC-Value(FR)	**147**	**14,7%**	**46,96%**	**22,39%**		
N-Grama (9570/99)	tf-idf	53	5,3%	38,97%	9,33%	1000	136
	FR	52	5,2%	38,24%	9,15%		
	C-Value	52	5,2%	38,24%	9,15%		
	NC-Value(tf-idf)	38	3,8%	27,94%	6,69%		
	NC-Value(FR)	39	3,9%	28,68%	6,87%		
	NC-Value(C-Value)	39	3,9%	28,68%	6,87%		
Padrões Morfossintáticos (6455/97)	tf-idf	56	5,6%	41,18%	9,86%		
	FR	**57**	**5,7%**	**41,91%**	**10,04%**		
	C-Value	**57**	**5,7%**	**41,91%**	**10,04%**		
	NC-Value(tf-idf)	45	4,5%	33,09%	7,92%		
	NC-Value(FR)	43	4,3%	31,62%	7,57%		
	NC-Value(C-Value)	43	4,3%	31,62%	7,57%		
Sintagma Nominal (4760/74)	tf-idf	51	5,1%	37,5%	7,93%		
	FR	51	5,1%	37,5%	8,98%		
	C-Value	51	5,1%	37,5%	8,98%		
	NC-Value(tf-idf)	39	3,9%	28,68%	6,87%		
	NC-Value(FR)	*37*	*3,7%*	*27,21%*	*6,51%*		
	NC-Value(C-Value)	*37*	*3,7%*	*27,21%*	*6,51%*		
N-Grama (8046/39)	tf-idf	23	2,3%	37,1%	4,33%	1000	62
	FR	24	2,4%	38,71%	4,52%		
	C-Value	24	2,4%	38,71%	4,52%		
	NC-Value(tf-idf)	19	1,9%	30,65%	3,58%		
	NC-Value(FR)	18	1,8%	29,03%	3,39%		
	NC-Value(C-Value)	18	1,8%	29,03%	3,39%		
Padrões Morfossintáticos (9195/49)	tf-idf	**29**	**2,9%**	**46,77%**	**5,46%**		
	FR	28	2,8%	45,16%	5,27%		
	C-Value	12	1,2%	19,35%	2,26%		
	NC-Value(tf-idf)	20	2%	32,26%	3,77%		
	NC-Value(FR)	20	2%	32,26%	3,77%		
	NC-Value(C-Value)	21	2,1%	33,87%	3,95%		
Sintagma Nominal (4455/41)	tf-idf	23	2,3%	37,1%	4,33%		
	FR	26	2,6%	41,94%	4,9%		
	C-Value	*10*	*1%*	*16,13%*	*1,88%*		
	NC-Value(tf-idf)	23	2,3%	37,1%	4,33%		
	NC-Value(FR)	23	2,3%	37,1%	4,33%		
	NC-Value(C-Value)	21	2,1%	33,87%	3,95%		

isto não ocorre, a medida se comporta como o cálculo da FR. Sendo assim, foram escolhidos os pares Padrões Morfossintáticos/FR e Padrões Morfossintáticos/*tf-idf* para a comparação de desempenho. O segundo par foi selecionado, pois obteve o resultado mais próximo do melhor.

No gráfico 6 é apresentada a comparação entre os pares selecionados. Conforme é possível perceber o par Padrões Morfossintáticos/*tf-idf* obteve piores resultados somente para as faixas de 550, 600 e 650 termos, além da utilizada no corte (1000).

Finalmente, para a avaliação de extração dos trigramas, o melhor par foi Padrões Morfossintáticos/*tf-idf*. Contudo, ele foi comparado com o segundo e terceiro par baseado nos resultados, nesse caso, Padrões Morfossintáticos/FR e SN/FR respectivamente. A Figura 7 apresenta a comparação entre eles. O resultado obtido por cada método/medida para as diferentes faixas de termos confirmou o melhor desempenho do Padrões Morfos-

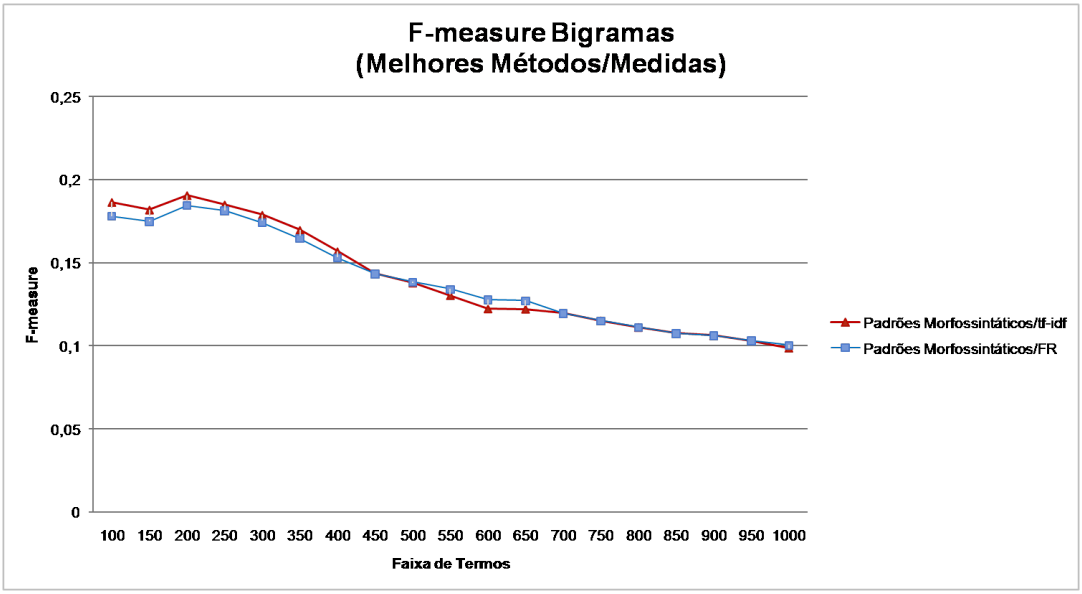

Figura 6. Resultados obtidos pelos métodos para diferentes faixas de termos durante a extração de bigramas (*F-measure*)

sintáticos/*tf-idf*, sendo inferior somente para os 150 e 200 termos.

Ressaltamos que a referência é composta de um número limitado de termos, o que explica a diminuição da medida 'F' conforme o aumento do número de termos incluídos para análise.

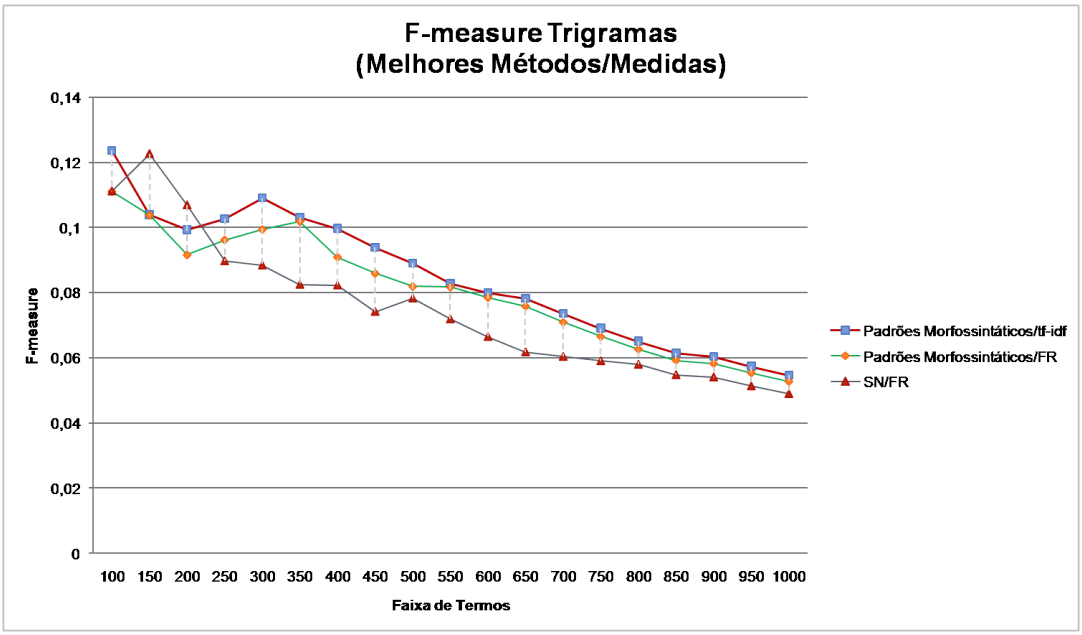

Figura 7. Resultados obtidos pelos métodos para diferentes faixas de termos durante a extração de trigramas (*F-measure*)

Com base nos resultados descritos acima, foram selecionados os pares: Classe Gramatical/*tf-idf* (unigramas), Padrões Morfossintáticos/*tf-idf* (bi e trigramas) para serem avaliados no experimento seguinte.

3.2. Avaliação do OntoLP feita por Usuários

A estratégia utilizada na Seleção de Grupos Semânticos torna necessária a intervenção do usuário durante o processo. Dessa forma, os resultados obtidos são dependentes do seu nível de conhecimento sobre o domínio. Conseqüentemente, optou-se por uma avaliação qualitativa da ferramenta feita por usuários experientes nas tarefas de extração de termos. O experimento teve como objetivo avaliar se a etapa de Seleção de Grupos Semânticos proposta realmente pode auxiliar o processo de extração.

Para cumprir o objetivo foi dada preferência por pesquisadores com experiência em construção de ontologias. Esses usuários receberam listas de termos constituídas por uni, bi e trigramas, para que excluíssem os irrelevantes. Sendo assim, os avaliadores deveriam possuir familiaridade com o domínio em questão, tornando-os aptos a fazer essa seleção.

Os experimentos foram executados por dois grupos de pesquisa, ambos relacionados à área de extração de termos e/ou organização hierárquica de conceitos. Para a tarefa foram utilizados dois corpora, de acordo com a experiência de cada grupo. Os grupos convidados para a tarefa e seu respectivo corpus foram:

- GETerm (UFSCar): este grupo foi constituído por dois avaliadores, um aluno de doutorado e um aluno de IC, todos da área de Lingüística. O grupo trabalha atualmente no desenvolvimento do projeto *NanoTerm* ([Aluísio 2005]), cujo corpus foi utilizado nos experimentos. O corpus do projeto é constituído por documentos extraídos de diversas fontes, incluindo textos informativos, científico e técnico administrativo, compondo um total geral de 2.565.490 palavras, distribuídas em 1057 textos.
- Grupo TERMISUL (UFRGS): neste grupo o experimento foi realizado por um avaliador, bolsista de IC da Ciência da Computação, supervisionado por um Doutor da área de Lingüística. Atualmente o grupo trabalha com o corpus *JPED* ([Coulthard 2005]), utilizado nos experimentos. O *JPED* é constituído por 283 textos em português extraídos do Jornal de Pediatria, totalizando 785.448 palavras.

Para a avaliação foram definidas uma série de tarefas de extração de termos. Ao final, cada avaliador gerou uma lista de Grupos Semânticos que considerou relevante para o domínio, utilizando o método Filtro por Grupos Semânticos. Para cada conjunto foram geradas três listas de termos, constituídas por uni, bi e trigramas. Finalmente, cada avaliador editou essas listas, restando apenas termos relevantes para o domínio. As listas finais foram utilizadas no cálculo de precisão dos métodos de extração de termos.

Nessa avaliação foram considerados apenas os melhores pares método/medida conforme o experimento anterior. Os resultados são apresentados na Tabela 3 para unigramas, bigramas e trigramas. Nela é demonstrado: o método acompanhado da medida (Método/Medida); o total de termos corretos extraídos (Corretos); a faixa de termos considerada (Corte); e a precisão de cada par método/medida (P). Além disso, a coluna "GS" indica os resultados com e sem a aplicação da etapa de Seleção de Grupos Semânticos.

Na maioria dos resultados a Precisão foi melhor quando aplicada Seleção de Grupos Semânticos. Como esperado, no domínio de Nanotecnologia & Nanociência (N&N) os resultados foram inferiores aos obtidos no domínio de Pediatria, provavelmente

por ser uma área nova e muito específica. Essas características influenciaram também na diferença entre os resultados alcançados com e sem a aplicação das informações semânticas. Para Pediatria o aumento foi de 17,33% para unigramas, enquanto para Nanotecnologia & Nanociência o aumento médio foi de 3,66%. Para os bigramas a melhora foi de 20,67% (Pediatria) e 2,67% (N&N). Finalmente, para trigramas as melhora nos resultados foram de 6,66% (Pediatria) e 0,165% (N&N).

Tabela 3. Resultado da extração de termos com e sem a Seleção de Grupos Semânticos

	Unigramas					
	Método/Medida	Avaliador	Corretos	P	GS	Corte
NanoTerm	Classe Gramatical/ tf-idf	Av1	26	17,33%	S	150
			19	12,67%	N	
		Av2	59	39,33%	S	
			55	36,67%	N	
JPED		Av3	108	72%	S	
			82	54,67%	N	
	Bigramas					
	Método/Medida	Avaliador	Corretos	P	GS	Corte
NanoTerm	Padrães Morfossintáticos/ tf-idf	Av1	19	12,67%	S	150
			17	11,33%	N	
		Av2	57	38%	S	
			51	34%	N	
JPED	Padrões Morfossintáticos/ tf-idf	Av3	139	92,67%	S	
			108	72%	N	
	Trigramas					
	Método/Medida	Avaliador	Corretos	P	GS	Corte
NanoTerm	Padrões Morfossintáticos/ tf-idf	Av1	17	11,33%	S	150
			17	11,33%	N	
		Av2	45	30%	S	
			43	29,67%	N	
JPED	Padrões Morfossintáticos/ tf-idf	Av3	83	55,33%	S	
			73	48,67%	N	

A geração de hierarquia de conceitos está em fase de avaliação. A Figura 8 ilustra o módulo de auxílio à construção de hierarquias. A interface oferece ao engenheiro uma sugestão de hierarquia (coluna 1) de acordo com diferentes métodos, e que pode ser exportada para a composição da ontologia em construção (coluna 2).

4. Conclusão

Entendemos que a pesquisa na área de ontologias, de maneira geral, está fortemente ligada ao desenvolvimento de várias das áreas indicadas como desafios da computação no Brasil (gestão da informação, modelagem e comunicação). A contextualização, localização ou nacionalização das questões relacionadas à gestão de informação, modelagem e comunicação com o cidadão é uma questão especialmente relevante neste cenário. Esse desafio envolve atenção à língua portuguesa.

Atualmente vemos o desenvolvimento de tecnologias para auxiliar na engenharia de ontologias, que tomam textos como fonte de conhecimento de um domínio. Para desenvolvermos esse tipo de tecnologia problemas particulares da língua devem ser considerados e pesquisados.

O desenvolvimento da área, de forma assim contextualizada, requer ainda a construção de infra-estrutura básica de pesquisa. Faz-se necessário construir e disponibilizar recursos que possam ser compartilhados entre os pesquisadores da área e fomentar a

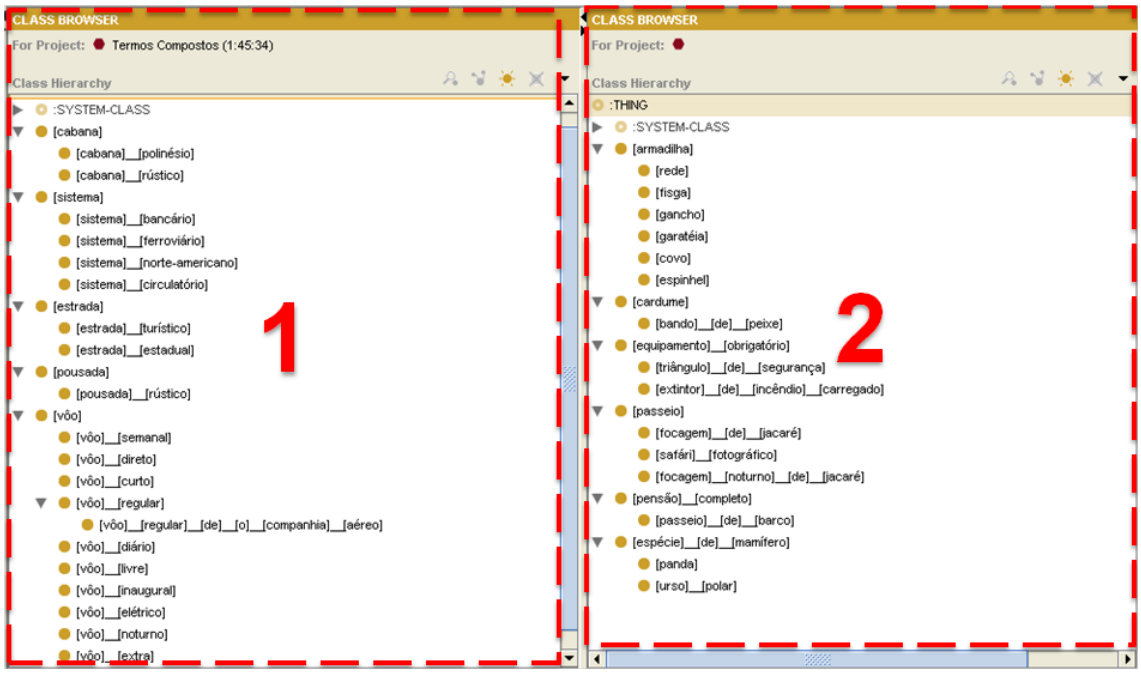

Figura 8. Geração de hierarquia

área. A carência de recursos básicos, de certa forma, contribui para a inércia da pesquisa do tratamento computacional do português, por outro lado, não vemos um interesse global, geral, internacional em desenvolver tecnologia com esse viés. Por isso, acreditamos que esse é um dos problemas a ser encarado como desafio para a computação no Brasil.

Consideramos o papel importante da língua específica utilizada no país, mas também reconhecemos que a globalização exige formas de comunicação mais universais. Em relação à aplicação e uso de ontologias, tem-se como um dos grandes problemas da área, o problema de integração de diferentes ontologias. E, apesar do trabalho estar centrado inicialmente na importância do processamento da língua portuguesa, a tendência é que os problemas aqui relacionados (gestão, modelagem, comunicação) venham a requerer, de forma mais intensa no seu desenvolvimento futuro, o tratamento de bases multilingües. Mas na pluralidade de línguas, é importante termos a presença da língua portuguesa, e da tecnologia para o seu tratamento específico.

O trabalho apresentado aqui se insere nesse contexto. Descrevemos um estudo de métodos de extração de termos a partir de textos da língua portuguesa, empregados na construcão de uma ferramenta de auxílio à engenharia de ontologias em língua portuguesa. Os métodos foram implementados em um plug-in para o ambiente Protégé, e uma avaliação inicial foi realizada. Em breve a ferramenta será disponibilizada para a comunidade científica. Pretendemos não apenas facilitar o trabalho de engenheiros de ontologia, mas estimular a pesquisa na área, através do compartilhamento de recursos.

Agradecimentos. Este trabalho teve apoio parcial da CAPES, CNPq e FAPERGS.

Referências

Aluísio, S. (2005). Desenvolvimento de uma estrutura conceitual (ontologia) para a área de nanociência e nanotecnologia. Technical Report 2004.1.34165.1.6, Universidade de

São Paulo.

Baségio, T. (2006). Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do brasil. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS.

Bick, E. (2000). *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology learning from text: An overview. In P-Buitelaar, Cimiano, P., and Magnini, B., editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Coulthard, R. J. (2005). The application of corpus methodology to translation: the jped parallel corpus and the pediatrics comparable corpus. Master's thesis, Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina.

Frantzi, K. T., Ananiadou, S., and ichi Tsujii, J. (1998). The c-value/nc-value method of automatic recognition for multi-word terms. In *ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 585–604, London, UK. Springer-Verlag.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Ryu, P.-M. and Choi, K.-S. (2006). Taxonomy learning using term specificity and similarity. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 41–48, Sydney, Australia. Association for Computational Linguistics.

Suchanek, F. M., Ifrim, G., and Weikum, G. (2006). Leila: Learning to extract information by linguistic analysis. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 18–25, Sydney, Australia. Association for Computational Linguistics.

Zavaglia, C., Aluísio, S., das Graças Volpe Nunes, M., and de Oliveira, L. M. (2007). Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. In *Anais do 5º Workshop em Tecnologia da Informação e da Linguagem Humana, TIL'2007*, pages 1575–1584, Rio de Janeiro, Brasil.