Luiz Carlos Ribeiro Junior

OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa

São Leopoldo

2008

Luiz Carlos Ribeiro Junior

OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

Orientador:
Renata Vieira

UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

São Leopoldo

2008

Dedicatória

Dedico este trabalho aos meus pais. Obrigado pelo amor e paciência.

Agradecimentos

Em primeiro lugar, agradeço a Deus, por sempre me indicar o melhor caminho a seguir.

Em seguida, agradeço a minha orientadora, professora Renata Vieira, pelo interesse constante no crescimento dos seus orientados. Durante esses dois anos você abriu portas importantes para nosso futuro, nos instigou a trabalhar em projetos paralelos e se fez presente em todos os momentos, acompanhando de perto nossos passos. Hoje vejo que essas atitudes influenciaram positivamente este trabalho.

Quanto à realização deste trabalho, agradeço a equipe do LEL (Zeca, Jonatan, Paty 2, César, Míriam e Vinícius), sem vocês não seria possível realizá-lo no prazo de dois anos! Agradeço ainda a Sandrinha, a quem quero deixar um MUITO OBRIGADO pela amizade e pelas tantas dicas que me forneceu.

Agradeço aos amigos que fiz durante o mestrado, ao longo desses dois anos passamos por momentos de angústia, de felicidade e, principalmente, de estudos. SOBREVIVEMOS! A vocês, Cony, Paty, Roberto (Roberval), Luciano (espichado) e Paulo, que sigamos sempre amigos e que conquistemos nossos sonhos.

Nesses agradecimentos, não podem faltar os amigos que fiz no NILC, obrigado Ane, Helena, Gawa, Elô, Thiago C., Jorge, Lívia, Lúcia, Teo e, especialmente, Thiago P. e Élen, com vocês viciei-me em "Imagem e Ação"! Gostaria de agradecer também a professora Sandra Aluísio, que dedicou longas horas do seu tempo trocando idéias e dando dicas valiosas para a realização desse trabalho!

De todos os grupos de pesquisa citados, deixo um agradecimento especial ao GPSI. Nele fiz amizades e ganhei uma base científica que auxiliou muito a realização deste trabalho. Obrigado aos professores Daniel, Stanley e Ramiro, pelos conhecimentos e conselhos, e aos antigos bolsistas e agora mestres ou quase mestres, Borges, Pezaum, Primo, Piltcher, Costinha, Marcos e Silva!

Agradeço também a Marcela, que desde o início de minha graduação serviu de inspiração para as minhas conquistas. Sem dúvida, não fosse por você eu não teria trilhado caminhos que considero terem me feito uma pessoa melhor. Obrigado!

Não poderia faltar um agradecimento a minha família. OBRIGADO PAI, MÃE, MANA E AMANDICA (que em breve poderá ler esse agradecimento)! Vocês são os maiores responsáveis por eu completar mais essa etapa da minha vida!

Finalmente, agradeço a CAPES pelo apoio financeiro durante o mestrado.

Anexos

Anexo A - Resultados obtidos em experimentos preliminares aplicando restrições aos termos

Anexo B - Manual do OntoLP

Anexo C - Taxonomias do processo de avaliação da organização hierárquica

Anexo D - Resultados obtidos por (CIMIANO; HOTHO; STAAB, 2005) para a Organização Hierárquica de Conceitos

Anexo E - Questionário de Avaliação do plug-in OntoLP

Resumo

O crescimento da Internet provoca a necessidade de estruturas mais consistentes de representação do conhecimento disponível na rede. Nesse contexto, a Web Semântica e as ontologias aparecem como resposta ao problema. Contudo, a construção de ontologias é extremamente custosa, o que estimula diversas pesquisas visando automatizar a tarefa. Em sua maioria, essas pesquisas partem do conhecimento disponível em textos. As ferramentas e métodos são, nesse caso, dependentes de idioma. Para que todos tenham acesso aos benefícios da utilização de ontologias em larga escala, estudos específicos para cada língua são necessários. Nesse sentido, pouco foi feito para o Português.

Este trabalho procura avançar nas questões concernentes à tarefa para a língua portuguesa, abrangendo o desenvolvimento e a avaliação de métodos para a construção automática de ontologias a partir de textos. Além disso, foi desenvolvida uma ferramenta de auxílio à construção de ontologias para a língua portuguesa integrada ao ambiente largamente utilizado Protégé.

Palavras-chave: Construção de Ontologia, Ontologias, Web Semântica, Processamento de Linguagem Natural.

Abstract

The internet evolution is in need of more sophisticated knowledge management techniques. In this context, the Semantic Web and Ontologies are being developed as a way to solve this problem. Ontology learning is, however, a difficult and expensive task. Research on ontology learning is usually based on natural language texts. Language specific tools have to be developed. There is no much research that considers specifically the portuguese language.

This work advances in these questions and it considers portuguese in particular. The development and evaluation of methods are presented and discussed. Besides, the developed methods were integrated as a plug-in of the widely used ontology editor Protégé.

Keywords: Ontology Learning, Ontologies, Semantic Web, Natural Language Processing.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

EAT – *Extração Automática de Termos*
FR – *Freqüência Relativa*
HTML – *Hypertext Markup Language*
MI – *Mutual Information*
NP – *Noun Phrase*
NS – *Name Space*
OWL – *Web Ontology Language*
PLN – *Processamento de Linguagem Natural*
SN – *Sintagma Nominal*
TF-IDF – *Term Frequency-Inverse Document Frequency*
XML – *eXtensible Markup Language*
W3C – *World Wide Web Consortium*

Sumário

Capítulo 1

Introdução

Através do prêmio *Millennium Technology Prize* concedido ao inglês Tim Berners-Lee, criador da Web, foi consolidada a importância dessa invenção para a humanidade. Ainda assim, pesquisas na área continuam a evoluir, tendo como um dos exemplos mais expressivos a Web Semântica[1].

A Web Semântica é um projeto ambicioso e prevê a reengenharia da Web atual. A idéia principal é o uso intensivo de meta-dados na organização semântica de documentos e serviços distribuídos pela rede. Atualmente, pesquisadores de diferentes áreas têm se dedicado a estudos relacionados ao seu desenvolvimento. Em alguns países da Europa, assim como nos Estados Unidos, investimentos significativos em pesquisas relacionadas à área vêm sendo realizados.

Segundo (BERNERS-LEE; HENDLER; LASSILA, 2001), para que os objetivos da Web Semântica sejam atingidos, computadores devem ter acesso a coleções estruturadas de informação e conjuntos de regras que possam ser usadas na condução de raciocínio automático. O modelo de representação de informações adotado na área são as chamadas Ontologias. O estudo dessas estruturas é tão importante que, de acordo com (MAEDCHE; STAAB, 2000), sua proliferação é um dos principais fatores para o sucesso da Web

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

Semântica.

Apesar de sua visibilidade estar recentemente muito atrelada à criação da Web Semântica, as ontologias vêm sendo aplicadas a diferentes áreas da computação. As questões de pesquisa relacionadas a essas estruturas são variadas: metodologias de construção de ontologias, mapeamento de ontologias, medidas de similaridade entre ontologias, e a construção automática de ontologias, sendo essa última o foco deste trabalho.

A principal dificuldade na utilização de ontologias em larga escala é seu processo de construção extremamente custoso. Para a realização da tarefa, geralmente é necessária a presença de um especialista no domínio a ser representado. Além disso, em muitos casos essa pessoa precisa ser treinada para familiarizar-se com a estrutura de representação. Devido a essa e outras questões, estudos relacionados à construção automática de ontologias são de vital importância, prevendo auxiliar e agilizar o processo de criação das mesmas.

Muitos dos esforços realizados atualmente na área propõem a construção de ontologias a partir de um conjunto relevante de textos de um domínio específico. Segundo (BUITELAAR; OLEJNIK; SINTEK, 2003), este parece um caminho acertado, visto que a linguagem é a primeira forma de transferência de conhecimento entre os seres humanos.

Como será discutido ao longo do trabalho, o estado da arte em construção automática de ontologias a partir de textos baseia-se fortemente no uso de informações lingüísticas. Essa característica torna os métodos propostos dependentes das particularidades do idioma. Sendo assim, para que todos tenham acesso às novas tecnologias provenientes da aplicação de ontologias é essencial que pesquisas específicas para as diferentes línguas sejam realizadas.

Quando comparado a outros idiomas, a pesquisa relacionada à construção automática de ontologias para o Português é bastante escassa. Essa escassez gera uma dificuldade adicional à tarefa, a obtenção de recursos para a execução de experimentos. Nesse sentido, a partir desse trabalho quatro pontos-chave para a área foram tratados: 1) estudo de novos métodos para a realização da tarefa; 2) desenvolvimento de módulos

automáticos para a avaliação das etapas do processo; 3) organização de recursos para realização de novas pesquisas; 4) desenvolvimento de uma ferramenta de auxílio à construção de ontologias. Cada um desses pontos é discutido durante o trabalho. Portanto, o objetivo deste trabalho é propor e avaliar técnicas para a construção automática de ontologias a partir de textos da língua portuguesa com base em técnicas propostas para outras línguas. Além disso, visa ao desenvolvimento de uma ferramenta de auxílio à realização da tarefa com base nos métodos desenvolvidos.

O texto se organiza da seguinte forma: no capítulo 2 são apresentadas as áreas correlatas ao trabalho, como Web Semântica, Ontologias e Processamento de Linguagem Natural (PLN), e os trabalhos relacionados. A metodologia para a construção de ontologias a partir de textos proposta é detalhada no Capítulo 3, enquanto no Capítulo 4 é apresentado o plug-in de auxílio à tarefa. A avaliação das etapas de construção compreendidas por esse trabalho, bem como a avaliação do plug-in como ferramenta, são apresentados no Capítulo 5. Finalmente, no Capítulo 6, são feitas as considerações finais e discutidos trabalhos futuros.

Capítulo 2

Fundamentos

Nesse capítulo é feita uma discussão sobre os temas Web Semântica e Ontologias. Além disso, são introduzidos alguns conceitos de PLN importantes para o entendimento do trabalho. Em seguida, são apresentados alguns métodos utilizados na construção automática de ontologias e a aplicação deles em abordagens propostas por outros trabalhos.

2.1 Web Semântica

Nos últimos anos o termo Web Semântica tornou-se muito familiar aos profissionais e pesquisadores da Internet e Web. Grande parte das evoluções tecnológicas propostas para essas áreas estão sendo impulsionadas por pesquisas de novos padrões para a construção de páginas Web. A idéia é substituir o formato atual, baseado na *Hypertext Markup Language* (HTML), por uma linguagem que possibilite enriquecer os documentos Web com informações semânticas. Dessa maneira é possível descrever a informação que está sendo armazenada, permitindo que máquinas realizem processamento automático sobre os documentos.

Atualmente, muitos pesquisadores vêm apostando na Web Semântica como a propulsora de uma revolução na forma como seres humanos relacionam-se entre si e com as

máquinas. Segundo (BERNERS-LEE; HENDLER; LASSILA, 2001), através da Web Semântica computadores serão capazes de compreender conteúdos de páginas disponíveis na Web, realizar negociações com o mínimo de auxílio dos seres humanos, entre outras tarefas.

Embora a Web Semântica pareça um grande salto em termos de funcionalidade, ela nada mais é do que uma extensão da Web atual, na qual as informações são disponibilizadas com seu significado bem definido, possibilitando que computadores e pessoas trabalhem de forma cooperativa (BERNERS-LEE; HENDLER; LASSILA, 2001). Portanto, como afirma (GRAU, 2004), a idéia é que a próxima geração da Web (Web Semântica) combine tecnologias já existentes na Web atual com formalismos de representação do conhecimento, visando prover uma infra-estrutura que permita processar, descobrir e filtrar melhor os dados disponibilizados on-line.

A necessidade de reestruturar a Web surgiu por dois motivos principais, seu crescimento e o formato utilizado para disponibilizar os dados. Desde que foi criada, a Web vem crescendo de forma exponencial, impulsionada pela quantidade cada vez maior de dispositivos conectados à Internet e pelo barateamento do *hardware* e formas de acesso. Essas facilidades fazem com que cada vez mais pessoas disponibilizem informações sobre assuntos diversos, ficando acessíveis instantaneamente. Entretanto, o formato usado para representar tais informações é projetado exclusivamente para leitura por pessoas, limitando a navegação somente a capacidade humana de vasculhar esse imenso espaço.

Para amenizar essa limitação surgiram os mecanismos de busca, que recebem palavras-chaves relacionadas aos interesses do usuário e realizam cálculos estatísticos, ou outros, sobre uma base de páginas previamente indexadas, retornando as mais relevantes conforme os termos usados na consulta. Apesar do sucesso desses mecanismos, eles são o exemplo mais conhecido de um dos principais problemas da Web atual, a chamada sobrecarga de informação (*information overload*). Conforme (CHEN, 1994), a sobrecarga de informação ocorre quando uma pessoa, ao realizar uma consulta, obtém um número excessivo de informações como resposta e não consegue absorvê-las ou tratá-las, tendo

que examinar todos os documentos resultantes para encontrar o que procura.

Apesar das constantes atualizações nos mecanismos de busca, para alguns pesquisadores a solução definitiva para o problema da sobrecarga é a alteração do formato de definição dos documentos Web. De acordo com (BERNERS-LEE; HENDLER; LASSILA, 2001), o conteúdo da Web foi inicialmente projetado para leitura por humanos, não para a manipulação dos dados através de programas de computador. Segundo (FERNEDA, 2003), as limitações da linguagem HTML refletem diretamente na qualidade da informação recuperada. Essas limitações motivaram a criação da *eXtensible Markup Language* (XML) que vem se tornando o novo padrão das páginas Web.

O padrão XML é uma das tecnologias de base para a construção da Web Semântica, conforme apresentado na Figura 1, chamada de "*Semantic Web Layer Cake*". Entre as camadas destacadas, a que possui maior relação com esse trabalho é a *Ontology Vocabulary*, responsável pela representação das ontologias. As demais camadas são descritas em "http://www.w3.org/".

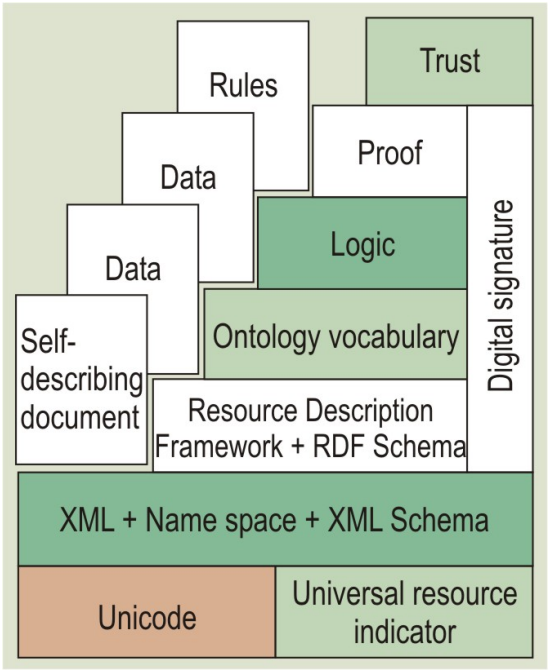


Figura 1: *Semantic Web Layer Cake* (HENDLER, 2001).

O objetivo da Web Semântica é desenvolver tecnologias e padrões projetados para auxiliar máquinas a entender as informações disponibilizadas online, possibilitando assim

que elas descubram informações valiosas para o usuário, integrem dados, naveguem pela rede e automatizem algumas tarefas (W3C, 2001).

2.2 Ontologia

O termo ontologia tem sua origem no idioma grego sendo onto (ser) + logia (estudo). Inicialmente, o termo era utilizado exclusivamente na área de filosofia. Nesta área, as ontologias estão focadas no fornecimento de sistemas de categorização para a organização da realidade (GUARINO, 1998).

Nos últimos anos ontologias vêm sendo consideradas importantes em diversas áreas da Ciência da Computação, em especial na Inteligência Artificial, na Lingüística Computacional, em Banco de Dados e na Web Semântica. Entretanto, na computação o termo não compartilha o mesmo significado que possui na filosofia. Nesse contexto, segundo (GUARINO; GIARETTA, 1995; GUARINO, 1998), o termo refere-se a um artefato de engenharia que, em uma visão simplista, pode ser descrito como uma hierarquia de conceitos relacionados entre si através de uma classificação de parentesco (hipernímia[1] e hipônimo[2]), também chamada de taxonomia. De uma forma mais sofisticada, axiomas apropriados são adicionados para expressar outros relacionamentos entre os conceitos e para restringir sua interpretação. A definição mais conhecida para o termo diz que: "*ontologia é considerada uma especificação formal e explícita de uma conceitualização compartilhada*" (GRUBER, 1993, 1995). Entretanto, para a correta interpretação do conceito é necessário ter em mente que (FENSEL, 2001):

- Conceitualização: é uma visão simplificada e abstrata do mundo que se deseja representar por algum propósito. Portanto, são os objetos, os conceitos e outras entidades existentes em alguma área de interesse e os relacionamentos existentes entre elas.

[1] É uma relação de parentesco onde um conceito é super-classe (mais geral) de outro. Por exemplo, "veículo" é hiperônimo de "automóvel".

[2] É a relação inversa a hipernímia, onde um conceito é sub-classe (mais específico) de outro.

- Explícita: significa que o tipo dos conceitos e as restrições disponibilizadas devem ser definidas de maneira explícita.

- Formal: refere-se ao fato de que uma ontologia deve ser passível de processamento automático.

- Compartilhada: reflete a noção de que uma ontologia captura o conhecimento consensual, não sendo restrito a um indivíduo, mas aceito por um grupo.

Deste modo, como afirma (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999), ontologias tratam da organização de objetos, suas propriedades e seus relacionamentos em um determinado domínio de conhecimento. Além disso, disponibilizam termos potencialmente úteis para descrever o conhecimento sobre um domínio específico.

Segundo (GUARINO, 1998), é possível diferenciar ontologias de acordo com sua generalidade, nesse caso, temos:

- Ontologias de topo ou de senso comum: descrevem conceitos bastante gerais como espaço, tempo, matéria, objeto, evento, ação etc., que são independentes de um problema ou domínio particular.

- Ontologias de domínio: descrevem o vocabulário relacionado a um domínio particular, especializando conceitos introduzidos nas ontologias de topo. Exemplos comuns são ontologias de medicina, automobilismo, computação, entre outras.

- Ontologias de tarefa: descrevem tarefas de um domínio como processos, planos, metas e escalonamentos através de uma visão funcional.

- Ontologias de aplicação: descrevem conceitos que dependem de um domínio e de uma tarefa particular, portanto, geralmente são uma especialização de ontologias de domínio e tarefa. Esses conceitos freqüentemente correspondem aos papéis desempenhados por entidades do domínio enquanto executam uma certa atividade.

Guarino diz ainda que ontologias de domínio e de tarefa especializam os termos presentes nas ontologias de topo, e que por sua vez, ontologias de aplicação se utilizam dos termos e regras das ontologias de domínio e de tarefas. Sendo assim, Guarino sugere uma relação de dependência entre os níveis de generalidade, demonstrada na Figura 2.

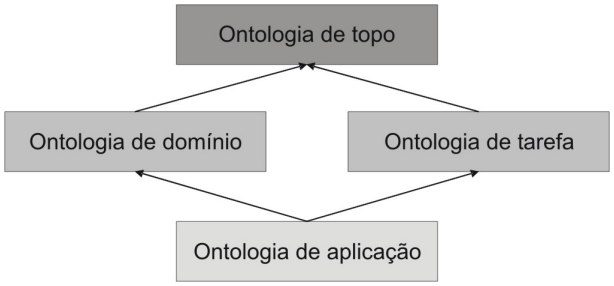


Figura 2: Ontologias de acordo com seu nível de generalidade

Apesar de serem aplicadas a diversas áreas, ontologias têm um papel especialmente importante para a Web Semântica. De acordo com (BERNERS-LEE; HENDLER; LASSILA, 2001), para o funcionamento da Web Semântica computadores devem ter acesso a coleções estruturadas de informação e conjuntos de regras que possam usar para conduzir raciocínio automático, sendo esse o principal desafio da área. Além disso, segundo (FENSEL, 2001), a comunidade acadêmica acredita que, em um futuro próximo, todo o negócio na Web deverá disponibilizar suas páginas através de uma estrutura ontológica.

Uma das principais dificuldades relacionadas a ontologias é a sua construção, sendo também um dos principais problemas para o estabelecimento da Web Semântica. Por esse motivo muitos trabalhos (FALBO; MENEZES; ROCHA, 1998; FERNANDEZ; GOMEZ-PEREZ; JURISTO, 1997; HOLSAPPLE; JOSHI, 2002; KISHORE; ZHANG; RAMESH, 2004; STUDER; BENJAMINS; FENSEL, 1998; USCHOLD; KING, 1995) propõem diferentes metodologias para a construção de ontologias. Apesar disso, ainda não existe uma metodologia considerada a "melhor". Outro contratempo é que essas metodologias propõem a construção manual de ontologias apenas com o auxílio de algumas ferramentas durante o processo. Porém, como citado em (BREWSTER; CIRAVEGNA; WILKS, 2003; MAEDCHE; STAAB, 2001), a construção manual de ontologias é um processo complexo, tedioso e de alto custo, e por ser extremamente artesanal é também propensa a erros. Sendo assim, diversos trabalhos

vêm propondo a construção automática ou semi-automática de ontologias para agilizar o processo e auxiliar na atualização das mesmas.

2.3 Processamento de Linguagem Natural

A área de Processamento de Linguagem Natural (PLN) visa desenvolver técnicas que possibilitem gerar e analisar textos escritos em um idioma humano. Esse é um grande desafio, visto que computadores estão aptos a compreender instruções escritas através de uma linguagem própria, as quais utilizam regras fixas e estruturas lógicas bem definidas. Em contrapartida, quando tratamos com linguagem natural, nos deparamos com características como ambigüidade e interpretações dependentes de contexto, entre outras.

Segundo (PERINI, 1985), a descrição da linguagem natural compõe-se basicamente de três elementos: a descrição formal, a descrição semântica e o sistema que relaciona ambos os planos. A primeira compreende os elementos fonológicos, morfológicos e sintáticos; enquanto a segunda se relaciona a todos os elementos anteriores através das regras de interpretação semântica. As regras fonológicas, morfológicas e sintáticas definem as construções possíveis na língua, enquanto as semânticas relacionam as construções e seus significados.

Os níveis relacionados a esse trabalho são os morfológico, sintático e semântico, isto porque o nível fonológico está relacionado à linguagem falada. Uma breve explicação sobre os níveis considerados é feita abaixo:

- O nível Morfológico é relativo à formação das palavras, sua composição em morfemas e as possibilidades de variação dessas.
- O nível Sintático trata das regularidades e equívocos na ordenação das palavras e estruturação das frases.
- O nível Semântico está relacionado ao estudo do significado das palavras e como

esse significado pode combinar com o significado das sentenças.

Ter acesso a esses níveis de informação é importante para a tarefa de construção de ontologias, visto que, grande parte dos métodos propostos tira proveito desses dados (seção 2.4). Sendo assim, duas ferramentas tornam-se fundamentais: sistemas que realizem a tarefa de anotação dos textos automaticamente e linguagens que possibilitem representar tais informações.

A seção 2.3.1 trata da tarefa de anotação de textos de uma forma geral. Já a seção 2.3.2 descreve o que são sintagmas e a quais as estruturas que os compõem. Finalmente, a seção 2.3.3 apresenta dois dos principais formatos de representação de informações lingüísticas.

2.3.1 Anotação de Textos

A anotação de textos consiste em disponibilizar informações lingüísticas para que sistemas possam utilizá-las. Para essa tarefa existem diversas ferramentas implementadas que abrangem diferentes línguas e a executam de forma automática (DECLERCK, 2002; BASILI; PAZIENZA; VELARDI, 1996; BICK, 2000). No geral, tais sistemas disponibilizam dois níveis diferentes de informações: *part-of-speech* (PoS) e estruturas sintagmáticas. O primeiro está relacionado ao processo de determinar a correta classe sintática para uma palavra particular, dado seu contexto atual (substantivos, verbos, adjetivos etc.). Em alguns casos, informações como forma canônica, gênero, número e grau podem acompanhar as informações de *part-of-speech*. O segundo refere-se às estruturas sintáticas da sentença e são descritos na seção abaixo.

2.3.2 Sintagmas

Os sintagmas são formados por um conjunto de elementos que constituem uma unidade significativa dentro da oração, e que mantêm entre si relação de dependência e

de ordem. Eles se organizam em torno de um elemento fundamental chamado núcleo, que pode, por si só, constituir um sintagma (SILVA; KOCH, 2001). É com base na categoria gramatical do seu elemento núcleo que os sintagmas são classificados:

- Sintagma nominal (SN): pode ter como núcleo um nome, um pronome (pro) ou um substantivo (N). Pode ser constituído por um determinante (Det), um pré-modificador e um pós-modificador (Mod), conforme a regra apresentada em 2.1. Um determinante em sua forma básica pode ser constituído por um artigo, numeral ou pronome. Já os modificadores são formados por um sintagma adjetival ou um sintagma preposicional.

$$SN \rightarrow \left\{ \begin{array}{c} (Det)\ \ (Mod)\ \ N\ \ (Mod) \\ pro \\ \Delta \end{array} \right\} \tag{2.1}$$

- Sintagma preposicional (SP): é constituído de uma preposição (prep) seguida de um SN. A regra para um sintagma preposicional é apresentada em 2.2.

$$SP \rightarrow prep + SN \tag{2.2}$$

- Sintagma adjetival (SA): tem como núcleo um adjetivo que pode vir acompanhado de outros elementos: intensificadores (intens) e modificadores adverbiais (SP_A), antepostos ao núcleo, e sintagmas preposicionais (SP_C), pospostos a ele, conforme apresentado na regra 2.3.

$$SA \rightarrow (intens)\ (SP_A)\ Adj\ (SP_c) \tag{2.3}$$

- Sintagma verbal (SV): tem como núcleo um verbo e apresenta configurações diversificadas. Entretanto, um sintagma verbal é o único que sempre desempenha a mesma função em uma frase, o de predicado.

Um exemplo da estrutura de uma oração (O) em sintagmas é apresentado na Figura 3. A representação é feita através de uma árvore chamada "indicador sintagmático".

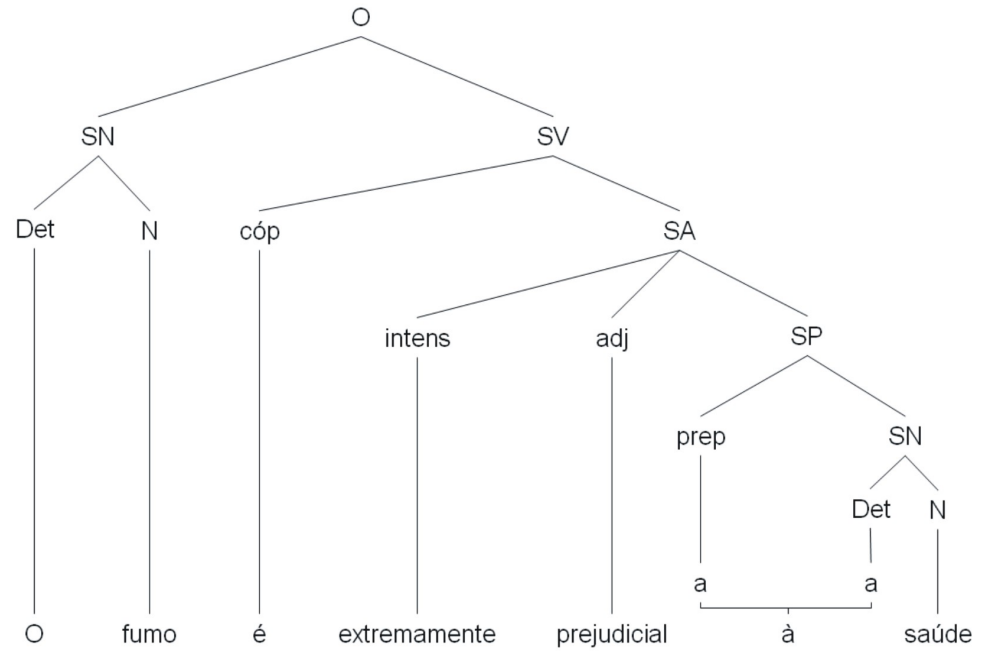


Figura 3: Exemplo de estruturação de uma sentença em sintagmas (SILVA; KOCH, 2001).

Por fim, cabe salientar que uma sentença, em sua forma mais básica, obrigatoriamente, deve possuir um sintagma nominal e um verbal.

2.3.3 Formatos de Anotação

Os sistemas que realizam a anotação lingüística de textos, na maioria dos casos, disponibilizam as informações geradas através de padrões representados em XML. Nessa seção são descritos dois formatos de representação: TigerXML[3] e XCES[4].

Segundo (BRANTS et al., 2002), o formato TigerXML é tipicamente dividido nos elementos "*head*" e "*body*". O *head* possui meta-dados sobre o corpus (tais como nome do corpus, autor, etc.) e a declaração das tags utilizadas para codificar os dados armazenados. No *body* estão presentes todos os níveis lingüísticos anotados pelo parser, nesse caso, morfológico, sintático e semântico. Os dados disponíveis nele são separados entre os

[3]http://www.ims.uni-stuttgart.de/projekte/TIGER/
[4]http://www.cs.vassar.edu/XCES/

elementos *terminals* e *nonterminals*. No primeiro estão informações sobre as palavras como PoS, etiquetas morfológicas e *lemmata*. Nos segundo são armazenadas informações relacionadas às estruturas sintagmáticas da sentença.

Na Figura 4 é apresentado um exemplo de saída TigerXML para a frase "*O fumo é extremamente prejudicial à saúde.*". No exemplo é possível perceber que tanto os *terminals* quanto os *nonterminals* são netos do elemento 's', que representa a sentença indicada pelo atributo "*text*". Os *terminals* possuem um ou mais sub-elementos 't', conforme o número de tokens que constituem a frase. Os atributos de 't' disponibilizam as seguintes informações:

- *id*: contém um identificador para o token, podendo ser usado em qualquer ponto do arquivo.
- *word*: armazena a palavra cujas informações estão sendo disponibilizadas no elemento.
- *lemma*: disponibiliza o lemma de uma palavra. Por exemplo, as palavras "vai" e "foi" possuem lemma "ir" (*lemma*="ir"), já as palavras "artistas" e "artista" possuem lemma "artista" (*lemma*="artista").
- *pos*: indica a categoria gramatical das palavras. Por exemplo, o token "saúde" (*id*="s1_7") pertence à classe dos substantivos (*pos*="n").
- *morph*: armazena informações morfológicas sobre a palavra. Considerando o token do exemplo anterior (saúde), as seguintes informações estão armazenadas nesse atributo: *morph*="F S", onde, 'F'-feminino e 'S'-singular.
- *sem*: contém informações semânticas sobre o token. Por exemplo, a palavra "saúde" está relacionada com um estado humano (*sem*="state-h", onde, *"state-h"=human state*).
- *extra*: possui dados extras sobre um token, por exemplo, o elemento de "*id*=s1_4", onde a palavra "extremamente" é considerada um quantificador (*quant*).

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <body>
  - <corpus>
    - <s id="s1" ref="1" source="Running text" forest="1" text="O fumo é extremamente prejudicial a saúde.">
      - <graph root="s1_500">
        - <terminals>
            <t id="s1_1" word="O" lemma="o" pos="art" morph="M S" sem="--" extra="--" />
            <t id="s1_2" word="fumo" lemma="fumo" pos="n" morph="M S" sem="cm" extra="--" />
            <t id="s1_3" word="é" lemma="ser" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="mv" />
            <t id="s1_4" word="extremamente" lemma="extremamente" pos="adv" morph="--" sem="--" extra="quant" />
            <t id="s1_5" word="prejudicial" lemma="prejudicial" pos="adj" morph="M S" sem="--" extra="--" />
            <t id="s1_6" word="a" lemma="o" pos="art" morph="F S" sem="--" extra="--" />
            <t id="s1_7" word="saúde" lemma="saúde" pos="n" morph="F S" sem="state-h" extra="--" />
            <t id="s1_8" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--" />
          </terminals>
        - <nonterminals>
          - <nt id="s1_500" cat="s">
              <edge label="STA" idref="s1_501" />
            </nt>
          - <nt id="s1_501" cat="fcl">
              <edge label="S" idref="s1_502" />
              <edge label="P" idref="s1_3" />
              <edge label="Cs" idref="s1_503" />
              <edge label="fCs" idref="s1_504" />
            </nt>
          - <nt id="s1_502" cat="np">
              <edge label="DN" idref="s1_1" />
              <edge label="H" idref="s1_2" />
            </nt>
          - <nt id="s1_503" cat="adjp">
              <edge label="DA" idref="s1_4" />
              <edge label="H" idref="s1_5" />
            </nt>
          - <nt id="s1_504" cat="np">
              <edge label="DN" idref="s1_6" />
              <edge label="H" idref="s1_7" />
            </nt>
          </nonterminals>
        </graph>
      </s>
    </corpus>
  </body>
```

Figura 4: Exemplo escrito no formato TigerXML.

No TigerXML, as estruturas sintáticas que compõem uma sentença são representadas por elementos "nt", filhos dos *nonterminals* (conforme visto na Figura 4). Os nt's são identificados pelo atributo "id" e classificados pelo atributo "cat", que define o tipo de estrutura armazenada. Esse elemento possui um ou mais sub-elementos "*edge*", que indicam as estruturas internas que compõem a informação disponível no "nt". Os "nt's" e *terminals* podem ser representados através de um grafo (Figura 5), onde cada elemento é considerado um nodo, sendo que, nós folhas são constituídos por (*terminals*), enquanto os demais são representados por *nonterminals*. Para exemplificar isso, na Figura 5, os nós cinza são elementos *"nt"*, enquanto os brancos são *"terminals"*. Cada nó do tipo "nt" é ligado a outro qualquer através de uma referência ("idref"), atributo de "*edge*". Portanto, ao analisarmos a Figura, é possível perceber que a estrutura "O fumo" é um sin-

tagma nominal (<nt id="s1_502" cat="**np**"/>) constituído por um determinante ("O") e um núcleo ("fumo"), ambos *terminals*. Essa mesma estrutura, no nível superior, desempenha a função de sujeito (<edge label="**S**" idref="**s1_502**"/>) na cláusula finita <nt id="s1_501" cat="**fcl**"/>. A cláusula, por sua vez, é constituída por um predicador (<edge label="**P**" idref="s1"/>), formado pelo *terminal* "é", e por outras sub-estruturas presentes na Figura anterior.

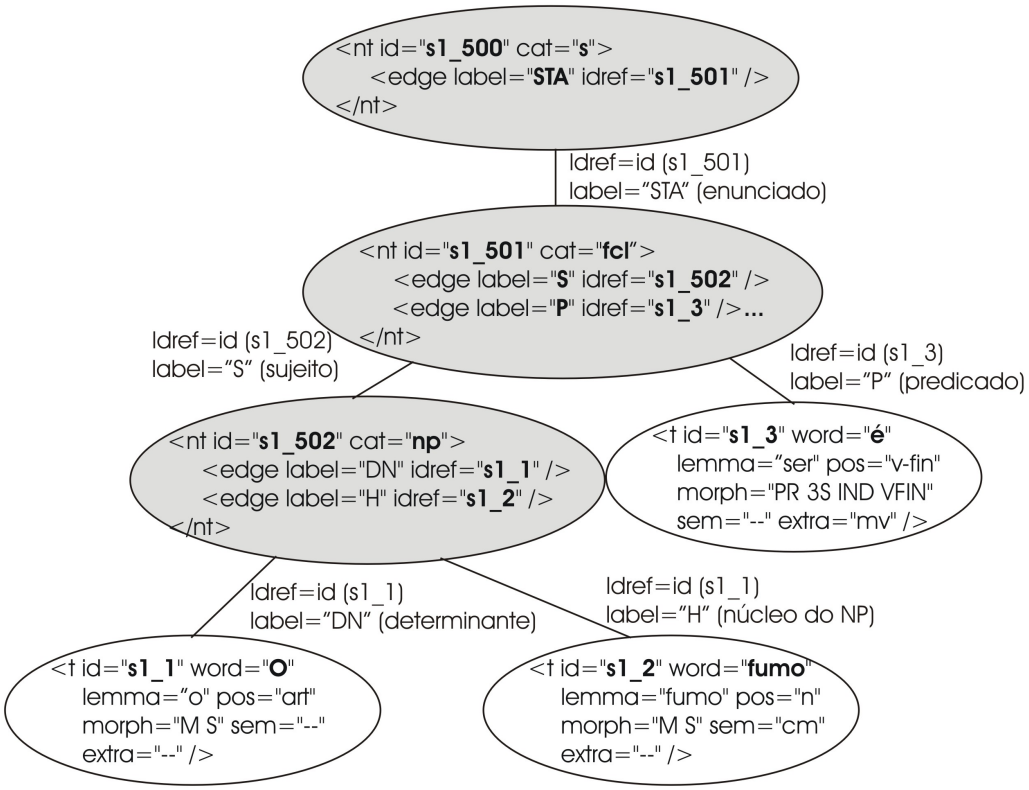


Figura 5: Representação da estrutura dos *nonterminals* através de um grafo.

Já o XCES é um formato XML para o chamado *Corpus Encoding Standard* (CES), conjunto de padrões para aplicações de PLN baseadas em corpus. A principal motivação para o desenvolvimento do XCES foi prover o estado da arte em representação e um framework para o *American National Corpus* (ANC)(IDE; BONHOMME; ROMARY, 2000).

O objetivo do XCES é especificar um formato que possibilite grande interoperabilidade entre anotações do mesmo fenômeno e entre tipos de anotação. O foco é propiciar um ambiente no qual elas possam ser facilmente definidas e validadas, em vez de ditar o uso de valores e elementos específicos (IDE; BONHOMME; ROMARY, 2000).

Para a codificação das informações do texto, o XCES baseia-se em dois elementos XML principais: "*struct*" e "*feat*". O primeiro está relacionado à estrutura textual que será representada, definida através do atributo "*type*". O segundo é um sub-elemento de *struct* que descreve as características dessa estrutura. *Feat* possui como atributos *name* e *value*, através dos quais são descritos o tipo de informação e seu valor respectivamente.

```
<?xml version="1.0" standalone="yes" ?>
- <cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
  - <struct type="token" from="0" to="1">
      <feat name="id" value="t1" />
      <feat name="base" value="O" />
    </struct>
  - <struct type="token" from="2" to="6">
      <feat name="id" value="t2" />
      <feat name="base" value="fumo" />
    </struct>
  - <struct type="token" from="7" to="8">
      <feat name="id" value="t3" />
      <feat name="base" value="é" />
    </struct>
  - <struct type="token" from="9" to="21">
      <feat name="id" value="t4" />
      <feat name="base" value="extremamente" />
    </struct>
  - <struct type="token" from="22" to="33">
      <feat name="id" value="t5" />
      <feat name="base" value="prejudicial" />
    </struct>
  - <struct type="token" from="34" to="35">
      <feat name="id" value="t6" />
      <feat name="base" value="a" />
    </struct>
  - <struct type="token" from="36" to="41">
      <feat name="id" value="t7" />
      <feat name="base" value="saúde" />
    </struct>
  - <struct type="token" from="41" to="42">
      <feat name="id" value="t8" />
      <feat name="base" value="." />
    </struct>
</cesAna>
```

Figura 6: Exemplo do formato XCES.

O formato utiliza arquivos separados para representar os diferentes níveis lingüísticos anotados pelo parser. Sendo assim, as informações de PoS (*terminals* no formato TigerXML) são armazenadas em um arquivo chamado de "*pos.xml*", enquanto as informações relacionadas as estruturas sintagmáticas (*nonterminals* no TigerXML) são descritas em um arquivo chamado "*phrases.xml*". Cabe salientar que ambos apontam para um arquivo base, chamado de "*token.xml*", onde estão presentes todas as palavras do texto. Um exemplo do arquivo, anotado para a mesma sentença do exemplo TigerXML,

é apresentado na Figura 6. Nele o elemento *struct* é do tipo token (*type*="token") e possui os atributos "*from*" e "*to*", que delimitam o início e fim das palavras no arquivo original. Nos sub-elementos *feat* são armazenados um identificador (<*feat name*="id" *value*="**t1**"/>) e a palavra em questão (<*feat name="base" value="**O**"*/>).

2.4 Construção Automática de Ontologias

A construção automática de ontologias (*Ontology learning*) é definida como o conjunto de técnicas e métodos usados para construir, enriquecer ou adaptar ontologias de uma maneira (semi-)automática, utilizando alguma fonte de conhecimento (WÄCHTER et al., 2006; PEREZ; MANCHO, 2003). Para essa tarefa, as abordagens propostas são variadas e, segundo (MAEDCHE; STAAB, 2001), podem ser diferenciadas pela fonte de conhecimento escolhida, sendo classificadas da seguinte forma:

- Construção de ontologias a partir de Dicionário: utilizam dicionários passíveis de processamento automático para extração de conceitos relevantes e seus relacionamentos.
- Construção de ontologias a partir de Base de Conhecimento: utilizam como fonte de informação bases de conhecimentos.
- Construção de ontologias a partir de esquemas semi-estruturados: extraem ontologias de fontes que têm alguma estrutura predefinida, como XML.
- Construção de ontologias a partir de esquemas relacionais: extraem conceitos e relações relevantes do conhecimento disponível em bases de dados.
- Construção de ontologias a partir de texto: criam ontologias através da aplicação das técnicas de análise de linguagem natural, sendo esse o foco deste trabalho.

Alguns pesquisadores defendem o uso de textos como fonte para a construção de ontologias (BUITELAAR; OLEJNIK; SINTEK, 2003; WÄCHTER et al., 2006; AUSSENAC-

GILLES; BIÉBOW; SZULMAN, 2000). Entretanto, existem algumas dificuldades que devem ser levadas em consideração quando do uso de aprendizado a partir de textos. Segundo (BASÉGIO, 2006), um dos problemas está em determinar uma linha divisória entre o que é explícito no texto e o que é assumido implicitamente. Dessa forma, torna-se difícil capturar computacionalmente o conhecimento que o autor de um texto assumiu estar compartilhado com seus leitores. Outra dificuldade é a de escolher uma fonte de conhecimento baseada em textos que seja apropriada e confiável para a construção de ontologias. As fontes mais conhecidas para a realização dessa tarefa são, segundo (BREWSTER; CIRAVEGNA; WILKS, 2003):

- Enciclopédia: parecem fontes ideais para o conhecimento ontológico, pois incluem definições e textos explicativos que podem ser explorados. Entretanto, esse tipo de fonte possui dois problemas principais: a dificuldade de acesso e sua provável desatualização.

- Livros didáticos e manuais: associados ao domínio, têm utilidade potencial. Seu principal problema é identificar os textos relevantes e obtê-los eletronicamente.

- Internet: esta é a fonte mais óbvia. Devido ao seu tamanho qualquer informação requerida é facilmente encontrada, por seu dinamismo é provável que conceitos novos estejam rapidamente disponíveis e é de fácil acesso. Seu problema principal é que o conceito de determinado termo pode aparecer em diversos textos de domínios diferentes. Além disso, é difícil determinar critérios para decidir se um site será confiável ou não.

A tarefa de construção de ontologias a partir de texto abrange basicamente as seguintes etapas baseadas em (BUITELAAR; CIMIANO; MAGNINI, 2005) e também adotadas em (BASÉGIO, 2006):

1. Seleção de Conceitos: são selecionados os termos candidatos a conceitos da ontologia.

2. Extração de relações taxonômicas: os termos selecionados na etapa anterior são organizados em uma estrutura hierárquica (taxonomia).

3. Extração de relações não taxonômicas: são identificadas as relações não hierárquicas entre os conceitos da ontologia.

4. Extração de Regras (axiomas): são extraídas regras que permitem a agentes de software realizar inferência sobre a ontologia.

5. Identificação de instâncias: é feita a extração do conhecimento extensional dos textos.

É importante ressaltar que a complexidade da tarefa cresce à medida que as etapas avançam, havendo uma relação de dependência entre elas. Dessa forma, as etapas iniciais são consideradas pré-requisito para as finais.

Devido à complexidade envolvida no processo de construção automática de ontologias, este trabalho visa abranger apenas as tarefas de extração de conceitos e sua organização hierárquica, respectivamente, as fases 1 e 2 descritas acima. Sendo assim, as seções 2.4.1 e 2.4.2 apresentam alguns dos principais métodos utilizados em cada etapa.

2.4.1 Métodos para Seleção de Conceitos

A extração de termos relevantes para um determinado domínio é uma tarefa central a qualquer abordagem para a construção automática de ontologias. Segundo (BUITELAAR; CIMIANO; MAGNINI, 2005), termos são realizações lingüísticas de conceitos de uma área específica, sendo vitais para a realização das etapas mais complexas.

Existem diversos métodos para a seleção de termos relevantes em um conjunto de textos, sendo que a maioria utiliza técnicas estatísticas para a tarefa. Algumas das métricas mais conhecidas são provenientes da área de Recuperação de Informação (RI). Cabe salientar que, quando aplicamos métodos puramente estatísticos, um documento é

tratado como um simples vetor de termos e, em alguns casos, suas freqüências. Portanto, é possível aplicá-los sem a necessidade da anotação dos textos.

Uma das abordagens mais simples para a tarefa é chamada de Freqüência Relativa (FR) (MANNING; SCHÜTZE, 1999) de um termo. A FR considera o número de vezes que o termo 't' aparece em um documento dividido pelo total de palavras do documento. Como resultado da aplicação do método, os termos são organizados em ordem decrescente, indicando sua relevância para o domínio. Em 2.4 é apresentada a equação para o cálculo da FR.

$$FR(t) = \frac{F(t)}{N} \tag{2.4}$$

Onde,

- $F(t)$ é o número de vezes que o termo aparece em um texto.
- 'N' é o número total de termos no documento.

Outra métrica bastante comum é a *term frequency-inverse document frequency* (tf-idf) (MANNING; SCHÜTZE, 1999). Essa medida considera que termos que possuem alta freqüência de ocorrência em um número limitado de documentos são relevantes para o domínio. Em 2.5 é apresentada a equação para o cálculo da *tf-idf*.

$$tfidf_{l,d} = lef_{l,d} * log\left(\frac{|D|}{df_l}\right) \tag{2.5}$$

Onde,

- lef_{ld} é a freqüência da entrada lexical 'l' em um documento 'd'.
- df_l é o número de documentos do corpus 'D' em que 'l' ocorre.

A equação apresentada acima retorna o *tf-idf* para um termo relacionado a um

único documento do corpus. Para obtermos um resultado que considere o corpus inteiro basta que façamos um somatório dos valores obtidos em cada documento (2.6).

$$tfidf_l = \sum_{d \in D} tfidf_{l,d} \tag{2.6}$$

Existem também medidas que calculam a relevância de um termo comparando sua freqüência em um corpus de domínio[5] e em um corpus de referência[6]. Sendo assim, um termo presente em um domínio particular é comparado com palavras de uso mais geral. A medida qui-quadrado (AGIRRE et al., 2001) é um dos principais exemplos dessa abordagem. A equação para seu cálculo é apresentada em 2.7.

$$w_{i,j} = \begin{cases} (freq_{i,j} - m_{i,j})m_{i,j} \quad if \quad freq_{i,j} > m_{i,j} \\ 0 \quad otherwise \end{cases} \tag{2.7}$$

Onde,

- $freq_{i,j}$ é a freqüência da palavra 'j' em um corpus 'i'.
- $m_{i,j}$ é a média esperada da palavra 'j' no documento 'i', definida como 2.8.

$$m_{i,j} = \frac{(\sum_k freq_{k,j})(\sum_k freq_{i,k})}{(\sum_{k,l} freq_{k,l})} \tag{2.8}$$

Outra medida que utiliza a abordagem anterior é a *log-likelihood* (MANNING; SCHÜTZE, 1999). Segundo (RAYSON; BERRIDGE; FRANCIS, 2004), essa métrica apresenta uma pequena melhora sobre a qui-quadrado para algumas situações. Em 2.9 é apresentada a equação da medida.

$$G2 = 2 * ((a * ln(\frac{a}{E1})) + (b * ln(\frac{b}{E2}))) \tag{2.9}$$

[5]Corpus de domínio é um conjunto de textos sobre determinada área de conhecimento, como por exemplo, Medicina.

[6]Um corpus de referência é constituído por documentos de diversas áreas, tornando-se independente de um domínio específico.

Onde,

- $E1 = c * \frac{(a+b)}{(c+d)}$.
- $E2 = d * \frac{(a+b)}{(c+d)}$.
- 'a' é a freqüência do termo no corpus de referência.
- 'b' corresponde à freqüência do termo no corpus de domínio.
- 'c' indica o total de palavras no corpus de referência.
- 'd' refere-se ao total de palavras no corpus de domínio.

Além das medidas apresentadas acima, existem também aquelas que procuram extrair exclusivamente termos complexos ou multi-palavras, ou seja, termos constituídos por duas ou mais palavras. Um método muito utilizado para essa tarefa é o cálculo de Informação Mútua (MI) (MANNING; SCHÜTZE, 1999), que indica o quanto uma palavra reduz a incerteza sobre outra. Por exemplo, dadas as palavras 'X' e 'Y', se 'X' for totalmente independente de 'Y', então o valor para a MI entre elas será igual a 0 (zero). O modo de cálculo para a MI é dado em 2.10.

$$MI(x, y) = log\frac{P(x, y)}{P(x)P(y)} \tag{2.10}$$

Onde,

- $P(x, y)$ é a probabilidade dos termos 'x' e 'y' ocorrerem juntos.
- $P(x)$ e $P(y)$ são, respectivamente, a probabilidade de ocorrência do termo 'x' e do termo 'y'.

Para a extração de termos complexos existe ainda o método chamado *n-grama* (INNISS et al., 2006), onde 'n' indica a quantidade de palavras que podem constituir um

termo. Essa técnica geralmente é aplicada em conjunto com alguma medida estatística, para avaliar a probabilidade de um *n-grama* ser realmente um termo. A idéia é percorrer um documento extraindo 'n' palavras de cada vez, calculando alguma medida estatística para cada n-grama extraído. Um exemplo visual do processo é apresentado na Figura 7.

Figura 7: Exemplo de seleção de n-gramas com $n = 2$.

Segundo (FRANTZI; ANANIADOU; TSUJII, 1998), o problema em abordagens puramente estatísticas é que ainda não existe uma com resultados em sua grande maioria úteis, ou seja, que retornem poucos termos indesejáveis. Por esse motivo, essas técnicas vêm sendo aplicadas em conjunto com outras metodologias, visando obter melhores resultados.

Existem ainda abordagens que extraem termos agrupando-os de acordo com sua similaridade, considerando os conjuntos resultantes como prováveis conceitos da ontologia. Esses métodos utilizam técnicas de clusterização e medidas de similaridade, como os coeficientes *Dice*, *Jacquard* e *cosine* (MANNING; SCHÜTZE, 1999). Nessas abordagens os termos são representados por vetores, sendo que as características utilizadas para descrevê-los são variadas, entre as principais estão:

- Freqüência de ocorrência do termo (MAEDCHE; STAAB, 2000).
- Informações lingüísticas como quais os modificadores, os verbos ou as preposições que acompanham o termo (FAURE; NEDELLEC, 1999).

Segundo (MANNING; SCHÜTZE, 1999), o coeficiente *Dice* é normalizado através do comprimento dos vetores, isso é feito pela divisão entre o número total de entradas diferentes de zero para ambos os vetores. Para o cálculo do coeficiente *Dice* é definida a equação 2.11.

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.11}$$

Onde,

- 'X' e 'Y' são os vetores de características que representam os termos a serem comparados.

De acordo com (MANNING; SCHÜTZE, 1999), o coeficiente *Jaccard* e o coeficiente *Dice* são semelhantes, entretanto, o *Jaccard* retorna valores menores quando há baixa sobreposição entre os vetores. Um exemplo pode ser visto considerando a equação de cálculo do coeficiente (2.12).

$$Jaccard = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.12}$$

Onde, dados dois vetores com 10 entradas não zero e uma entrada em comum, o coeficiente resulta em 0.05263:

$$\frac{1}{10 + 10 - 1} = 0.05263 \tag{2.13}$$

Enquanto o *Dice* resulta em 0.1:

$$\frac{2 * 1}{10 + 10} = 0.1 \tag{2.14}$$

Vale lembrar que ambos os coeficientes são normalizados entre 0 (zero) e 1 (um).

A última medida citada, o *cosine*, é idêntica ao coeficiente *Dice* para vetores com números iguais de entradas não zero, porém, aplica penas menores aos casos em que esses números diferem (MANNING; SCHÜTZE, 1999). Para o cálculo do *cosine* é utilizada a equação 2.15.

$$cosine = \frac{|X \cap Y|}{\sqrt{|X| * |y|}} \tag{2.15}$$

Além das técnicas descritas acima existem algumas provenientes das áreas de Ter-

minologia e PLN. Um exemplo é o trabalho apresentado em (PANTEL; LIN, 2001), que realiza a extração de bigramas e em seguida combina as medidas de MI e *Log-Likelihood* para auxiliar no processo de extração de termos multi-palavras do domínio. A execução ocorre da seguinte forma: (1) é realizada a extração dos bigramas a partir dos textos; (2) são coletadas características para representar esses bigramas; (3) são calculadas as medidas MI e *log-likelihood* para selecionar termos complexos a partir dos bigramas.

Em (FRANTZI; ANANIADOU; TSUJII, 1998) é apresentado um método de extração a partir de corpora anotados. A técnica divide-se em duas etapas: *C-value* e *NC-value*. Na primeira são selecionados os termos com base nos filtros lingüísticos apresentados na Tabela 1. Segundo (FRANTZI; ANANIADOU; TSUJII, 1998), esses padrões são constituídos por substantivos, adjetivos e preposições, pois geralmente são essas classes que compõem termos complexos.

Tabela 1: Filtros utilizados para a extração de termos multi-palavras.

	Filtros
1.	$Noun + Noun$
2.	$(Adj\|Noun)^{+}Noun$
3.	$((Adj\|Noun)^{+}\|((Adj\|Noun)^{*}(NounPrep)^{?})(Adj\|Noun)^{*})Noun$

Em seguida, os termos são ordenados através da medida *C-Value* (2.16).

$$Cvalue(a) = \begin{cases} log_2|a|.f(a) & se\ `a'\ está\ contido\ em\ outro\ termo, \\ log_2|a|f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) & , caso\ contrário. \end{cases} \quad (2.16)$$

Onde,

- 'a' corresponde ao termo candidato.
- $f(.)$ é a FR do termo no corpus.
- T_a indica conjunto de termos candidatos extraídos que contêm 'a'.
- $P(T_a)$ é a probabilidade de ocorrência do termo 'a' como parte de termos candidatos.

A segunda etapa utiliza a medida *NC-Value* para reordenar os termos extraídos. Para esse processo, o método recebe uma lista de termos/medida de relevância, nesse caso, a *C-Value*. Aos termos são incorporadas as chamadas "palavras de contexto", que são palavras que ocorrem em sua vizinhança, sendo constituídas por substantivos, adjetivos e verbos. Uma porcentagem dos termos mais relevantes, segundo a *C-Value*, é selecionada, e às "palavras de contexto" relacionadas a eles são atribuídos valores com base na equação 2.17.

$$weight(w) = \frac{t(w)}{n} \tag{2.17}$$

Onde,

- 'w' é a palavra a ser analisada.
- '$t(w)$' corresponde ao número de termos com os quais a palavra 'w' aparece.
- 'n' indica o total de termos presente no conjunto que está sendo considerado.

Finalmente, a lista é reorganizada com base na presença ou não das "palavras de contexto" com os demais termos (equação 2.18). O método considera que termos bem ranqueados são específicos do domínio, e que o contexto em que eles ocorrem (representado pelas "palavras de contexto") provavelmente auxilie na identificação de outros termos.

$$NC - value = 0.8Cvalue(a) + 0.2 \sum_{b_a} f_a(b) weight(b) \tag{2.18}$$

Onde,

- 'w' é o termo candidato.
- C_a representa o conjunto de palavras de contexto distintas de 'a'.
- 'b' é uma palavra de C_a.

- $f_a(b)$ indica a freqüência de 'b' como uma "palavra de contexto" de 'a'.

- $weight(b)$ retorna o peso de 'b' como uma "palavra de contexto".

Segundo os autores, os fatores aplicados ao *C-value* e *weight* (respectivamente, 0.8 e 0.2) foram obtidos com base em experimentos preliminares.

Diversas abordagens já foram propostas para a realização dessa etapa. Além das técnicas apresentadas aqui, existem outras descritas na seção 2.5.

2.4.2 Métodos para Construção de Taxonomias

Os métodos utilizados para a construção de taxonomias, em sua maioria, utilizam informações lingüísticas para a execução da tarefa. Nessa seção são apresentados alguns dos principais disponíveis atualmente.

(HEARST, 1992) propõe uma abordagem que utiliza um conjunto de padrões baseados em informações léxico-sintáticas. O objetivo do método proposto é a extração de relações de hiponímia (relação "*é-um*") dos textos. Cabe salientar que os padrões de Hearst visam extrair somente termos constituídos por SN's, conforme demonstrado na Tabela 2. Os padrões foram originalmente criados para a aplicação em textos da língua inglesa.

Tabela 2: Padrões de Hearst.

	Padrão de Hearst
1.	NP such as {(NP,)*(or—and)} NP
2.	such NP as {(NP,)*(or—and)} NP
3.	NP {,NP}* {,} or other NP
4.	NP {,NP}* {,} and other NP
5.	NP {,} including {NP,}*{or—and} NP
6.	NP {,} especially {NP,}*{or—and} NP
NP = Noun Phrase (Sintagma Nominal)	

Onde,

- NP = Noun Phrase (Sintagma Nominal).

Abaixo são apresentados trechos de sentenças retiradas de textos, ordenados conforme o padrão aplicado. Para cada trecho são demonstradas as relações extraídas (HEARST, 1992):

1. Sentença: "*Agar is a substance prepared from a mixture of red algae, such as Gelidium...*"
    - Relação: hiponímia("Gelidium", "red algae")
2. Sentença: "*...works by such authors as Herrick, Goldsmith, and Shakespeare*"
    - Relação: hiponímia("author", "Herrick")
    - Relação: hiponímia("author", "Goldsmith")
    - Relação: hiponímia("author", "Shakespeare")
3. Sentença: "*Bruises, wounds, broken bones or other injuries...*"
    - Relação: hiponímia("bruise", "injury")
    - Relação: hiponímia("wound", "injury")
    - Relação: hiponímia("broken bone", "injury")
4. Sentença: "*...temples, treasuries, and other important civic buildings*"
    - Relação: hiponímia("temple", "civic building")
    - Relação: hiponímia("treasury", "civic building")
5. Sentença: "*All common-law countries, including Canada and England...*"
    - Relação: hiponímia("Canada", "common-law country")
    - Relação: hiponímia("England", "common-law country")
6. Sentença: "*...most European countries, especially France, England and Spain.*"
    - Relação: hiponímia("France", "European country")

- Relação: hiponímia("England", "European country")
- Relação: hiponímia("Spain", "European country")

Outra abordagem que utiliza padrões baseados em informações lingüísticas é apresentada em (MORIN; JACQUEMIN, 2004), construídos para o Francês. Os padrões propostos por Morin/Jacquemin são apresentados na Tabela 3, na qual a numeração que acompanha os *NP*'s e *LIST*'s indica sua posição na relação, (1) hiperônimo e (2) hiponímia, respectivamente, super-classe e sub-classe.

Tabela 3: Padrões propostos por Morin/Jacquemin.

	Padrões de Morin/Jacquemin
1.	{deux—trois...—1—2—3—4...} NP1 (LIST2)
2.	{certain—quelque—de autre...} NP1 (LIST2)
3.	{deux—trois...—1—2—3—4...} NP1: LIST2
4.	certain—quelque—de autre...} NP1: LIST2
5.	de autre}? NP1 tel que LIST2
6.	NP1, particullièrement NP2
7.	{de autre}? NP1 comme LIST2
8.	NP1 tel LIST2
9.	NP2 {et—ou} de autre NP1
10.	NP1 et notamment NP2
NP = noun phrase (SN), LIST = lista de SN's separados por ',' (virgula).	

A aplicação desses padrões ocorre da mesma forma que os de Hearst, mudando apenas, em alguns casos, as estruturas extraídas. Abaixo são apresentados dois exemplos (padrão 1 e 2) aplicados a fragmentos de um texto (MORIN; JACQUEMIN, 2004):

1. Sentença: "*...analyse foliaire de quatre espèces ligneuses (chêne, frêne, lierre et cornouiller) dans l'ensemble des sites éstudiés.*"
   - Relação: hiperônimo("chêne", "espèces ligneuses")
   - Relação: hiperônimo("frêne", "espèces ligneuses")
   - Relação: hiperônimo("lierre", "espèces ligneuses")
   - Relação: hiperônimo("cornouiller", "espèces ligneuses")

2. Sentença: "*Aprés cinq années de résultats sur les principales cultures vivrières (sorgho, maïs, mil), il apparaît qu'il existe un grand nombre de combinaisons possibles.*"

    - Relação: hiperônimo("sorgho", "cultures vivrières")
    - Relação: hiperônimo("maïs", "cultures vivrières")
    - Relação: hiperônimo("mil", "cultures vivrières")

Apesar de serem largamente utilizados para a construção de taxonomias, tanto os padrões de Hearst quanto os de Morin/Jacquemin ocorrem raramente em textos, obtendo alta precisão, porém, baixa abrangência.

Outra abordagem de construção de taxonomias é a baseada em Termos Complexos, também chamados de termos compostos ou multi-palavras. Esse método analisa a estrutura interna dos termos, portanto, quando um termo ocorre dentro de outro (**Arquitetura**→ **Arquitetura** Colonial), ele é considerado seu hiperônimo. Três trabalhos que utilizam essa técnica são descritos em (BUITELAAR; OLEJNIK; SINTEK, 2003; RYU; CHOI, 2006; BASÉGIO, 2006), sendo detalhados na seção 2.5. Cabe ressaltar que esse método parte de uma lista de termos previamente extraída. Na Figura 8 é apresentado um exemplo para os conceitos "**Arquitetura**", "**Arquitetura** Colonial", "**Arquitetura** Eclesiastica" e "**Arquitetura** Moderna".

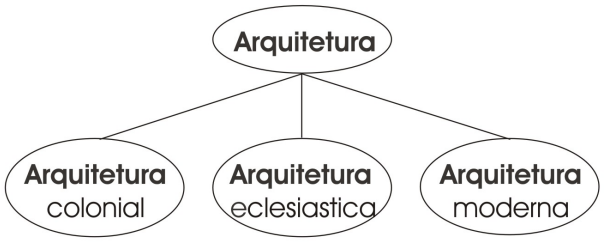


Figura 8: Sub-árvore gerada a partir do método baseado em Termos Complexos.

Finalmente, existem trabalhos que buscam construir estruturas taxonômicas a partir da clusterização hierárquica dos termos. Em (FAURE; NEDELLEC, 1999) é proposto um método que utiliza o relacionamento sintático entre os termos e os verbos como atributos

para representá-los. Os clusters são construídos através de uma abordagem *bottom-up*, onde cada termo é considerado um cluster individual e, a cada iteração do algoritmo, os termos mais similares são agrupados. Esse processo é repetido até restar somente um único cluster (raiz), onde todos os termos estão presentes.

2.5 Trabalhos Relacionados

2.5.1 Demais Abordagens

Nesta seção são apresentadas algumas metodologias para a construção automática de ontologias a partir de textos. Os processos descritos utilizam variações dos métodos discutidos nas seções anteriores.

(BUITELAAR; OLEJNIK; SINTEK, 2003) propõe uma abordagem que utiliza técnicas de processamento estatístico e informações lingüísticas para a construção de ontologias. Com base no processo proposto foi desenvolvida a ferramenta OntoLT, um plug-in para o editor Protégé (seção 4.1) que possibilita a construção e expansão de ontologias. Nessa metodologia a tarefa é realizada basicamente a partir de dois conceitos:

- Linguagem de pré-condição: é uma linguagem utilizada para definir restrições impostas às sentenças. Segundo (BUITELAAR; OLEJNIK; SINTEK(DIPL.-INFORMATIKER KM, 2004), as pré-condições são um conjunto de termos (funções e constantes) e predicados que selecionam entidades lingüísticas específicas. Elas são implementadas através da linguagem de consulta XPath[7].

- Regras de Mapeamento: são estruturas definidas através da linguagem de pré-condição e que possuem um conjunto de operadores vinculados. Através delas o OntoLT seleciona termos candidatos a conceitos e atributos da ontologia automaticamente. O plug-in possibilita que o usuário defina suas próprias regras,

[7]http://www.w3.org/TR/xpath

além de possuir duas pré-definidas: "HeadNounToClass_ModToSubClass" (núcleo do SN para Classe e o núcleo e seus modificadores para Sub-Classe) e "SubjTo-Class_PredToSlot_DObjToRange" (sujeito lingüístico para classe, predicado para slot da classe e o objeto direto para "imagem").

A abordagem possui uma etapa estatística que visa selecionar apenas informações lingüísticas relevantes ao domínio, utilizando a medida qui-quadrado (seção 2.4.1). A avaliação do plug-in foi feita com o auxílio de um especialista, embora (BUITELAAR; OLEJNIK; SINTEK(DIPL.-INFORMATIKER KM, 2004) defenda o uso de medidas de avaliação de ontologias, como Precisão e Abrangência.

(RYU; CHOI, 2006) propõem a construção automática de taxonomias através de duas medidas estatísticas que indicam a especificidade de um termo e a similaridade entre termos. A primeira calcula o quanto de informação um termo possui sobre determinado domínio, termos com alta especificidade estão mais próximos dos nós folha da taxonomia, enquanto termos de baixa especificidade estão mais próximos da raiz. A segunda medida calcula a similaridade entre os termos presentes na taxonomia e um novo termo a ser inserido, posicionando o novo termo próximo aos mais semelhantes semanticamente. O cálculo dessas métricas é feito a partir de algumas "características" baseadas em informações lingüísticas. Para a avaliação da abordagem foram utilizadas as medidas de precisão, abrangência e *f-measure*. A ontologia construída foi avaliada com base em uma ontologia de referência[8]. Dessa forma, segundo (RYU; CHOI, 2006), a avaliação é feita levando em consideração a semelhança entre a ontologia construída automaticamente e a construída com o auxílio do especialista. De acordo com os autores, apesar de a abordagem obter bons resultados para *f-measure*, a precisão é baixa. Para melhorar essas medidas devem ser estudados novos métodos para o cálculo de especificidade e similaridade.

(FREITAS, 2007) apresenta subsídios para a elaboração automática de ontologias a partir de textos, voltado para a Língua Portuguesa. O trabalho assume que determinadas

[8]Ontologias de referência são aquelas consideradas como sendo a correta descrição de uma área e são construídas com o auxílio de um especialista no domínio.

relações semânticas, como a hiperonímia, podem estar sistematicamente expressas em textos por meio de determinados padrões léxico-sintáticos. Entre os padrões avaliados no trabalho, três foram identificados através de uma análise sobre córpora em português e outros três adaptados do trabalho de Hearst. Os padrões considerados são apresentados na Tabela 4.

Tabela 4: Padrões utilizados por (FREITAS, 2007).

1	tipos de SN0: SN1 , SN2 ... , (e — ou) NPi
2	SN0 chamado/s/a/as SN1
3	SN conhecido/s/a/as como SN
4	NP0 such as NP1 , NP2 ... , (and — or) NPi
5	NP , NP* , or other NP0
6	NP , NP* , and other NP0

A partir da análise da aplicação desses padrões ao corpus, outro ponto importante é abordado pelo trabalho, a dificuldade de alguns padrões em extrair corretamente construções em que ocorrem sintagmas preposicionados. Para esses casos foram propostas duas regras para minimizar esses problemas:

- **SN HHiper**: que considera como hiperônimo apenas o primeiro substantivo a esquerda do padrão. Por exemplo, a sentença "[Infecções por bactérias] como [a Salmonella] e [a Shighella]" com a aplicação direta do padrão "tais como/como" de Hearst, extrai as seguintes relações:

  – relação: hiperônimo(Salmonella, infecções por bactérias);

  – relação: hiperônimo(Shighella, infecções por bactérias);

  No exemplo, o termo "infecções por bactérias" aparece de forma errada como hiperônimo de "Salmonella" e "Shighella". Com a aplicação da regra **SN HHiper** temos as relações:

  – relação: hiperônimo(Salmonella, bactérias);

  – relação: hiperônimo(Shighella, bactérias);

  Isto por que "bactérias" é o substantivo mais próximo de "tais como/como".

- **SN HHipo**: é considerado hipônimo o primeiro substantivo anterior à expressão "e/ou outros" e, no caso de uma coordenação de hipônimos, a estrutura **HHipo** se aplicará sempre ao sintagma mais à esquerda da relação. Sendo assim, no exemplo "... pode contribuir para [a maior ocorrência de [doenças cardiovasculares]], [cânceres] e outras [enfermidades] ..." as relações extraídas são:
  - hipônimo(enfermidades, doenças cardiovasculares);
  - hipônimo(enfermidades, cânceres).

Da definição dessas regras, os padrões estudados foram reestruturados para os apresentados na Tabela 5.

Tabela 5: Padrões resultantes da aplicação das regras propostas por (FREITAS, 2007).

1.a	SN HHiper (tais como — como) SN1 , SN2 ... , (e — ou) SNi
1.b	SN Hiper, (tais como — como) SN1 , SN2 ... , (e — ou) SNi
2	SN HHipo ,SN Hipoi * , e—ou outros SN Hiper
3	tipos de SN Hiper: SN1 , SN2 ... , (e — ou) SNi
4	SN HHiper chamado/s/a/as (de) SN Hipo
5	SN Hiper conhecido/s/a/as como SN Hipo

O trabalho propõem ainda a aplicação de uma regra de inferência, chamada **HiperN**. Essa regra se assemelha à identificação de relações hierárquicas baseada em termos complexos. Contudo, a identificação por termos complexos necessita que tanto os hipônimos quanto seu hiperônimo estejam presentes em uma lista de termos previamente extraídos, como em "sintomas" e "sintomas agudos". Na regra proposta por Freitas, com base em dois hipônimos de mesmo núcleo um hiperônimo é inferido. Por exemplo, para termos como "sintomas agudos" e "sintomas de gripe" a aplicação da regra possibilita a inferência de "sintoma" como hiperônimo.

2.5.2 Abordagem Proposta por Baségio

Em (BASÉGIO, 2006) é apresentada uma metodologia para a construção de ontologias a partir de textos em língua portuguesa. Segundo o autor, a abordagem tem como

foco a identificação de termos relevantes e suas relações taxonômicas. O método proposto utiliza como entrada um corpus anotado da área de Turismo. De acordo com Baségio, as seguintes informações estão disponíveis nesse corpus:

- A palavra no seu formato original.
- O lema da palavra original, ou seja, a palavra em sua forma singular e masculina.
- Informações de *PoS*.

Para a primeira parte do processo, a identificação de termos candidatos a conceitos, Baségio definiu cinco etapas:

1. Eliminação de termos que não representam conceitos de domínio: palavras que possuem significado semântico limitado são excluídas. Isto é feito através de uma lista de *stopwords*.
2. Pesagem dos termos: nessa etapa são utilizadas as medidas *log-likelihood* e *tf-idf*. A primeira medida é usada para selecionar apenas termos considerados relevantes para o domínio. A segunda para organizar os termos por ordem de relevância e posteriormente apresentar os termos ao especialista.
3. Definição do limiar mínimo para os termos: o engenheiro de ontologia define um limiar mínimo para a medida *tf-idf*, termos que estiverem abaixo do ponto de corte são desconsiderados.
4. Exclusão/Inclusão de termos: a lista resultante dos processos anteriores é apresentada ao engenheiro de ontologia, possibilitando que ele inclua ou exclua termos de acordo com seu julgamento.
5. Identificação de termos compostos: são selecionados termos compostos que tenham ao menos um termo considerado relevante em sua composição. Essa seleção é feita através de um conjunto de padrões que, quando encontrados no texto, extraem

termos considerados relevantes. Na Tabela 6 são apresentadas os padrões utilizados por Baségio.

Tabela 6: Regras para extração de termos compostos (BASÉGIO, 2006).

Nro.	**Regra**
1	_SU _AJ _PR _AD _SU _AJ
2	_SU _AJ _PR _AD _SU
3	_SU _PR _AD _SU _AJ
4	_SU _PR _AD _SU
5	_SU _AJ _PR _SU _AJ
6	_SU _AJ _PR _SU
7	_SU _PR _SU _AJ
8	_SU _PR _SU
9	_SU _AJ

Onde,

- _SU: substantivos.
- _AJ: adjetivos.
- _PR: preposições.
- _AD: advérbios.

Alguns exemplos de termos extraídos através da aplicação dessas regras são apresentados abaixo:

- _SU _AJ ⇒ "hotel excelente".
- _SU _PR _SU ⇒ "agencia de viagens".
- _SU _PR _SU _AJ ⇒ "agencia de turismo local".

A segunda parte do processo é responsável pela identificação de relações taxonômicas. Para essa tarefa são utilizados os termos extraídos e avaliados anteriormente pelo especialista. Com base nesses termos são aplicados três métodos diferentes, organizados no seguintes passos:

1. Identificar relações taxonômicas com base em termos complexos (seção 2.4.2): as relações são identificadas a partir do núcleo de um termo composto. Por exemplo, o termo "contrato" é considerado hiperônimo de "contrato de venda".

2. Identificar relações taxonômicas através dos padrões de Hearst: Baségio propõe uma adaptação dos padrões de Hearst para a língua portuguesa. Os padrões adaptados por Baségio estão descritos na Tabela 7.

3. Identificar relações taxonômicas através dos padrões de Morin/Jacquemin: assim como na etapa anterior, nessa foi feita uma adaptação dos padrões de Morin/Jacquemin para o Português, conforme apresentado na Tabela 8.

Tabela 7: Padrões de Hearst adaptados para o português por (BASÉGIO, 2006).

	Padrão Original	**Tradução/Adaptação**
1	NP such as {(NP,)*(or—and)} NP	SUB como {(SUB,)*(ou—e)} SUB
		SUB tal(is) como {(SUB,)*(ou—e)} SUB
2	such NP as {(NP,)*(or—and)} NP	tal(is) SUB como {(SUB,)*(ou—e)} SUB
3	NP {,NP}* {,} or other NP	SUB {,SUB}* {,} ou outro(s) SUB
4	NP {,NP}* {,} and other NP	SUB {,SUB}* {,} e outros SUB
5	NP {,} including {NP,}*{or—and} NP	SUB {,} incluindo {SUB,}*{ou—e} SUB
6	NP {,} especially {NP,}*{or—and} NP	SUB {,} especialmente {SUB,}*{ou—e} SUB
		SUB {,} principalmente {SUB,}*{ou—e} SUB
		SUB {,} particularmente {SUB,}*{ou—e} SUB
		SUB {,} em especial {SUB,}*{ou—e} SUB
		SUB {,} em particular {SUB,}*{ou—e} SUB
		SUB {,} de maneira especial {SUB,}*{ou—e} SUB
		SUB {,} sobretudo {SUB,}*{ou—e} SUB

Tabela 8: Padrões de Morin/Jacquemin adaptados para o português por (BASÉGIO, 2006)

	Padrão Original	**Tradução/Adaptação**
1	{deux—trois...—1—2—3—4...} NP1 (LIST2)	{dois—três...—1—2—3—4...} SUB1 (LIST_SUB2)
2	{certain—quelque—de autre...} NP1 (LIST2)	{certos—quaisquer—de outro(s)...} SUB1 (LIST_SUB2)
3	{deux—trois...—1—2—3—4...} NP1: LIST2	{dois—três...—1—2—3—4...} SUB1: LIST_SUB2
4	{certain—quelque—de autre...} NP1: LIST2	{certos—quaisquer—de outro(s)...} SUB1: LIST_SUB2
5	{de autre}? NP1 tel que LIST2	{de outro(s)}? SUB1 {tal(is)}* como LIST_SUB2
6	NP1, particullièrement NP2	{particularmente—especialmente} SUB2
7	{de autre}? NP1 comme LIST2	{de outro(s)}? SUB1 como LIST_SUB2
8	NP1 tel LIST2	SUB1 como LIST_SUB2
9	NP2 {et—ou} de autre NP1	SUB2 {e—ou} de outro(s) SUB1
10	NP1 et notamment NP2	SUB1 e (notadamente—em particular) SUB2

Baségio identificou a equivalência entre alguns padrões propostos em ambas as abordagens (Tabela 9), descartando os sugeridos por Morin/Jacquemin. Os restantes foram generalizados, com o objetivo de tentar aumentar sua abrangência (Tabela 10). Finalmente, dos padrões de Morin/Jacquemin foram considerados apenas as generalizações mais o padrão 10.

Tabela 9: Equivalência entre os padrões de Hearst e Morin/Jacquemin.

Morin/Jacquemin	**Hearst**
5- {de autre}? NP1 tel que LIST2	1- NP such as {(NP,)*(or—and)} NP
6- NP1, particullièrement NP2	6- NP {,} especially {NP,}*{or—and} NP
7- {de autre}? NP1 comme LIST2	1- NP such as {(NP,)*(or—and)} NP
8- NP1 tel LIST2	1- NP such as {(NP,)*(or—and)} NP
9- NP2 {et—ou} de autre NP1	3- NP {,NP}* {,} or other NP
	4- NP {,NP}* {,} and other NP

Tabela 10: Padrões de Morin/Jacquemin generalizados por Baségio.

Padrões Considerados	**Generalização**
1 e 2	1- SUB1 (LIST_SUB2)
3 e 4	2- SUB1: LIST_SUB2

Cabe salientar que, como Hearst e Morin/Jacquemin trabalham com SN's, em ambas as adaptações os sintagmas foram substituídos por substantivos (SUB) quando utilizados por Baségio. Nos padrões de Morin/Jacquemin adaptados, Baségio considera LIST_SUB como uma lista de substantivos separados por vírgula, enquanto a numeração utilizada (1 e 2) indica a posição do termo na relação (1-hiperônimo, 2-hipônimo), conforme descrito na seção 2.4.2.

Para a avaliação da metodologia foram feitos dois estudos de caso, realizados com o auxílio de um especialista. De acordo com Baségio, os resultados indicaram que a medida *log-likelihood* (seção 2.4.1), utilizada para exclusão de termos irrelevantes, se mostrou importante para o processo de poda. Com a aplicação da medida foram retornados 412 termos candidatos a conceitos, sem sua aplicação foram retornados 3.308. A medida *tf-idf* também obteve bons resultados na definição da relevância dos termos. Segundo o autor, mais de 50% dos termos identificados pelo especialista estavam entre os 100 mais relevantes, estando 80% deles entre os 50 primeiros. As regras para extração de termos

complexos obtiveram bons resultados, 57% dos extraídos foram indicados como relevantes pelo especialista. A regra que obteve melhores resultados foi a "_SU _AJ", responsável por mais da metade dos termos identificados pela ferramenta e selecionados pelo especialista (57,41% do total de termos selecionados).

2.5.3 Considerações

Conforme apresentado, as metodologias para a construção automática de ontologias são variadas. Contudo, é possível perceber uma tendência à utilização de informações lingüísticas para obtenção de melhores resultados na realização da tarefa.

O método apresentado por Buitelaar possui certas particularidades com relação aos demais. Em sua abordagem os critérios para a extração de termos candidatos a conceitos podem ser adaptados pelo engenheiro de ontologia. Isto é interessante, pois a extração pode ser adequada para diferentes contextos ou necessidades. Porém, para que o engenheiro realize esse processo é necessário que conheça a linguagem XPath, o que limita sua utilização. Essa característica vai de encontro ao proposto neste trabalho, que é disponibilizar um conjunto de técnicas de auxílio a construção de ontologias com uma configuração pré-definida.

A abordagem proposta por Ryu e Choi utiliza as medidas de especificidade e similaridade para a construção de taxonomias. Para o cálculo dessas métricas os termos são representados por diferentes características lingüísticas, provenientes da anotação de PoS. A avaliação dos seus resultados é feita com base nas medidas de Precisão, Abrangência e F-measure, adaptadas para levar em consideração a estrutura hierárquica dos termos. Essas métricas são utilizadas neste trabalho através de um módulo automático de avaliação de taxonomia.

Finalmente, Baségio apresenta um esforço inicial para a construção de ontologias a partir de textos da língua portuguesa. Devido a algumas dificuldades, como a falta de uma ferramenta de anotação e a indisponibilidade de recursos (termos e ontologia

de referência) para testes, são utilizadas técnicas simples de construção de ontologias e avaliação da metodologia. Apesar disso, a abordagem faz uma importante contribuição com a adaptação dos padrões de Hearst e Morin/Jacquemin para o português. Neste trabalho, utilizamos os padrões adaptados por Baségio para o português com algumas alterações.

Finalmente, os padrões avaliados por (FREITAS, 2007) não foram aplicados nesse trabalho, sendo citados como trabalho futuro (seção 6.2).

Capítulo 3

Metodologia para a Construção Semi-Automática de Ontologias

Este capítulo descreve a metodologia proposta para a construção de ontologias a partir de textos da Língua Portuguesa. A abordagem utiliza algumas das técnicas descritas nas seções 2.4.1, 2.4.2 e 2.5, para as quais foram feitas adaptações, conforme apresentado nas próximas seções.

A Figura 3 exibe uma visão geral da metodologia. É importante salientar que o trabalho abrange as duas primeiras etapas de construção de ontologias, extração de termos candidatos a conceito e sua organização hierárquica.

Conforme demonstrado na Figura 3, o processo divide-se em três etapas:

- Leitura do Corpus XCES: depois de anotado lingüísticamente e representado no formato XCES, o corpus é utilizado nesta etapa como fonte de conhecimento para a construção da ontologia.

- Extração de Termos: etapa na qual são aplicados diferentes níveis de conhecimento lingüístico e métodos, visando à extração de termos simples e complexos. Permite ao engenheiro de ontologia visualizar e editar a lista de termos extraída, gerando

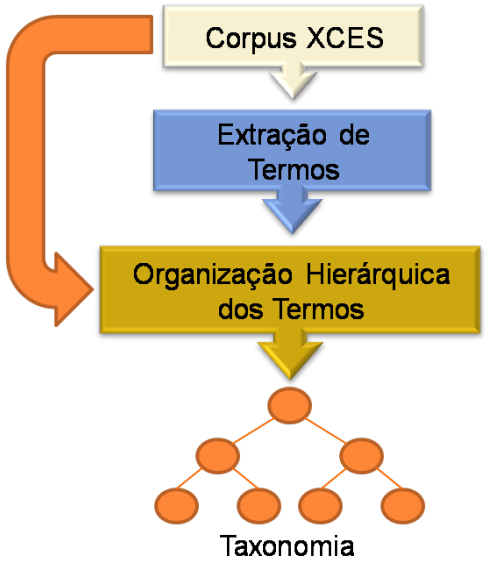


Figura 9: Metodologia proposta para a construção de ontologias a partir de textos.

uma lista final.

- Organização Hierárquica dos Termos: os termos extraídos na etapa anterior servem de entrada para os métodos de organização hierárquica. O engenheiro pode editar as taxonomias geradas, melhorando o resultado final.

Cada etapa da metodologia permite que o engenheiro de ontologia intervenha no processo, aprimorando a saída dos módulos. Esta estratégia foi adotada por que, como afirma (CIMIANO; VÖLKER; STUDER, 2006), não existe atualmente técnica de extração e/ou organização de termos que não necessite da intervenção do usuário para obtenção de bons resultados.

Essas etapas são detalhadas nas seções 3.1, 3.2 e 3.3, respectivamente, Leitura do Corpus XCES, Extração de Termos e Organização Hierárquica dos Termos.

3.1 Leitura do Corpus XCES

Os métodos de extração e organização hierárquica de termos aplicados neste trabalho utilizam conhecimento lingüístico. Sendo assim, é necessário que o corpus de entrada

esteja anotado com informações lingüísticas. Como apresentado na seção 2.3.1, para obtermos tais informações automaticamente faz-se necessário a utilização de um analisador sintático. Neste trabalho está sendo utilizado o analisador para o português, PALAVRAS, detalhado abaixo.

3.1.1 Analisador Sintático PALAVRAS

O PALAVRAS é um sistema de anotação lingüística de textos que realiza uma análise profunda sobre os documentos. Ele disponibiliza informações sobre os níveis lingüísticos: morfológico, sintático e semântico. Essas informações são dadas através de etiquetas (tags) que indicam características de uma determinada palavra ou estrutura frasal. O sistema fornece três tipos diferentes de saídas: um formato visual, um formato próprio do parser (VISL) e o formato TigerXML, esse último apresentado na seção 2.3.3.

Além das três saídas nativas do sistema, foi desenvolvida no Laboratório de Engenharia da Linguagem da UNISINOS (LEL) uma biblioteca de conversão do formato TigerXML para o formato XCES/PLN-BR, também descrito na seção 2.3.3. O formato XCES/PLN-BR foi adotado como padrão de entrada para o protótipo desenvolvido neste trabalho. A escolha desse formato envolve a facilidade de processamento e entendimento da estrutura de armazenamento dos dados, a divisão das informações lingüísticas em diferentes níveis, possibilitando carregar somente os arquivos necessários para a execução dos métodos, e a adoção do formato pelo projeto PLN-BR.

3.2 Extração de Termos

Esta seção explica em detalhes a etapa de extração de termos candidatos a conceitos. A Figura 10 apresenta as sub-tarefas do processo, bem como os métodos aplicados em cada uma delas. Para auxiliar a extração de termos simples e complexos foi adicionado um módulo baseado em informações semânticas, originalmente proposto neste trabalho.

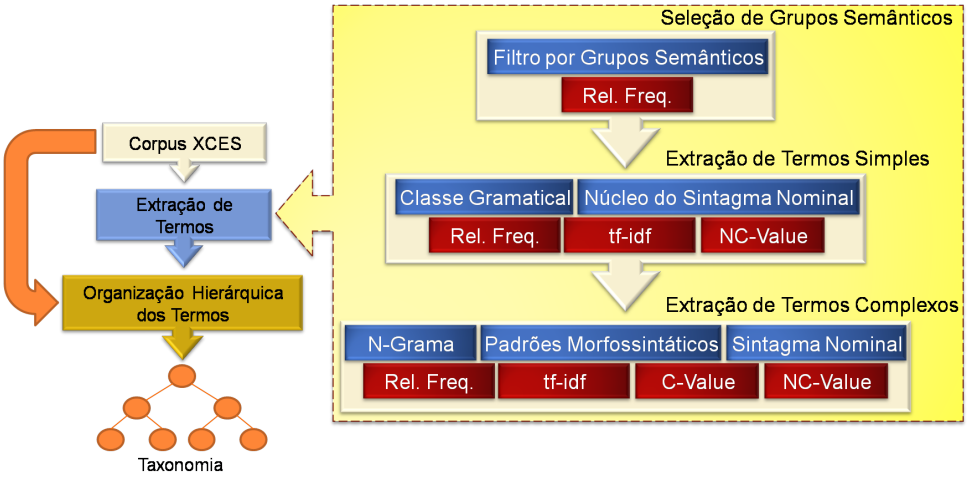


Figura 10: Abordagem Principal de extração de termos.

Conforme será visto no capítulo 4, o protótipo desenvolvido possibilita ao engenheiro de ontologia alterar a metodologia de extração de termos através de painéis de configuração. A Tabela 11 descreve resumidamente os passos definidos para a execução do processo, indicando quais deles são opcionais.

Tabela 11: Processo resumido de extração de termos.

Processo de Extração de Termos
Extração dos Grupos Semânticos; (opcional)
Filtragem dos Grupos Semânticos irrelevantes feita pelo engenheiro; (opcional)
Extração de Termos Simples considerando apenas aqueles pertencentes aos Grupos selecionados;
Exclusão dos termos simples irrelevantes, realizada pelo engenheiro; (opcional)
Extração dos termos complexos considerando apenas aqueles que possuem no mínimo uma palavra presente na lista final de termos simples e que pertençam a um Grupo selecionado;
Exclusão dos termos complexos irrelevantes, feita pelo engenheiro. (opcional)

Nas seções seguintes são apresentados cada um dos métodos disponíveis nas etapas de extração de termos.

3.2.1 Seleção de Grupos Semânticos

Conforme descrito na seção 2.4.1, existem três principais abordagens para a identificação de termos: estatística, lingüística e híbrida. Na primeira, cada documento per-

tencente ao corpus é considerado simplesmente como um vetor de termos e, em alguns casos é empregada a freqüência de ocorrência dos termos. Na segunda é necessário que os textos estejam anotados com informações lingüísticas, geralmente disponibilizadas por sistemas analisadores sintáticos. Tais métodos procuram, na maioria dos casos, extrair determinados padrões morfossintáticos dos textos. Na última abordagem é feito o casamento entre as duas metodologias anteriores. Nesse caso, os termos são extraídos de acordo com alguns padrões lingüísticos, em seguida, calcula-se a relevância dos mesmos através dos métodos estatísticos.

Apesar de cada uma das abordagens de extração de termos possuírem suas particularidades, existem alguns passos que podem ser empregados em qualquer uma delas, por exemplo, a remoção de *stopwords*. O processo de remoção de *stopwords* costuma apresentar melhoras significativas aos métodos de extração de termos. Entretanto, a construção de uma lista de *stopwords* não é algo trivial. Listas compostas por classes gramaticais como artigos, preposições, pronomes, advérbios e etc. são de fácil obtenção na Web. Contudo, uma boa lista inclui também substantivos pré-selecionados a partir do corpus e identificados como não sendo termos do domínio. Nesse caso, uma lista utilizada em um domínio 'A' não necessariamente poderá ser empregada em um domínio 'B'. Portanto, para a construção de uma lista de *stopwords* é necessária uma análise prévia do corpus, o que denota um passo manual extra no processo de extração de termos.

A abordagem empregada nesse trabalho não utiliza listas de *stopwords*, ao invés disso, é proposta uma etapa anterior à de extração de termos, chamada de Seleção de Grupos Semânticos (Figura 10). Essa etapa utiliza as informações semânticas disponibilizadas pelo PALAVRAS (seção 2.4.1). Conforme explicado anteriormente, essas informações são dadas através de marcações, chamadas aqui de tags semânticas. Segundo (BICK, 2006), o PALAVRAS possui cerca de 160 tags semânticas organizadas de tal forma que, dependendo do contexto, pode haver uma ou mais etiquetas associadas a um substantivo. As tags semânticas agregam significado a um determinado termo, por exemplo, a tag "<an>" atribuída ao substantivo "olho", indica que a palavra pertence à classe "Anatomia". Es-

pecificando ainda mais, o mesmo termo poderá receber também a tag "<anmov>", indicando que ele é membro da subclasse de "Anatomia", "Anatomia Móvel". Cabe salientar que as tags semânticas utilizadas no Palavras baseiam-se no trabalho (ROSCH, 1978).

Considerando que diversos substantivos podem ser etiquetados com uma mesma tag semântica, é possível perceber que quando anotados pelo PALAVRAS eles são agrupados em classes semânticas. Na Figura 11 são demonstrados três exemplos resumidos de grupos semânticos:

- Grupo <an> (Anatomia): {testa, ouvido, neurônio, cintura,
  - subgrupo <anmov> (Anatomia Móvel): {olho, pé, mão}

  }.
- Grupo <L> (Lugar): {território, campo, mundo, zona, área,
  - subgrupo <Lwater> (Lugar Aquático): {rio, oceano, mar}

  }.
- Grupo <H> (Humano): {pessoa, assassino, vítima, ser-humano, torcedor,
  - subgrupo <Hprof> (Profissão Humana): {pesquisador, arqueólogo}

  }.

A partir do estudo desses grupos foi criado o método Filtro por Grupo Semântico, que visa auxiliar a extração de termos simples e complexos. Quando habilitado o método, antes da extração de termos, o corpus de domínio é analisado através de um critério semântico, onde são executadas as etapas apresentadas na Figura 12 e descritas abaixo:

- Extração dos Grupos Semânticos: aqui são extraídas todas as tags semânticas presentes no corpus de entrada;

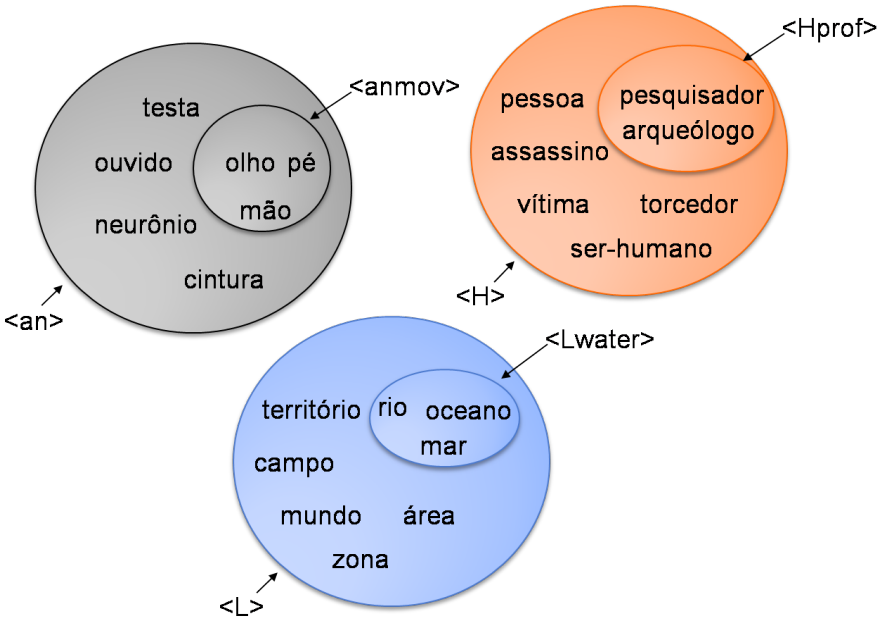


Figura 11: Exemplo de grupos e subgrupos semânticos.

- Cálculo de relevância dos Grupos Semânticos: nessa etapa é aplicado o cálculo de FR a lista de tags semânticas extraída anteriormente. Esse cálculo indica a importância do grupo para o domínio, ou seja, grupos que ocorrem com maior freqüência têm maior probabilidade de possuírem termos relevantes ao domínio. O vetor de tags é então apresentado ao engenheiro de ontologias ordenado pela freqüência.

- Exclusão dos Grupos Irrelevantes: nessa etapa o engenheiro pode excluir os grupos semânticos que considera não ter relação com o domínio representado pelo corpus de entrada.

Figura 12: Processo de Construção dos Grupos Semânticos.

Analisando melhor o método, a cada grupo semântico teremos então um conjunto de termos associados. Sendo assim, quando o engenheiro exclui um determinado grupo que considera não possuir relação semântica com o domínio, estará automaticamente desprezando todos os termos pertencentes somente àquele grupo.

Portanto, é possível dizer que o método de Filtro por Grupo Semântico possibilita, através de uma abordagem semi-automática, a construção de uma lista de *stopwords* específica para um domínio. Obviamente, a seleção dos Grupos corretos depende do conhecimento do engenheiro de ontologia sobre a área.

Cabe salientar que a etapa de Seleção de Grupos Semânticos é opcional e pode ser desabilitada no plug-in, possibilitando a extração de termos apenas por critérios morfossintáticos.

3.2.2 Extração de Termos Simples

O processo de extração de termos simples proposto neste trabalho é apresentado na Figura 13 e seus passos são descritos abaixo:

1. O módulo de Extração de Termos Simples recebe a lista de Grupos Semânticos gerada na etapa anterior (opcional). Os métodos de extração percorrem o corpus, selecionando os termos considerados aptos e que pertençam a pelo menos um grupo semântico presente na lista de entrada. Esse último, caso habilitado o método Filtro por Grupos Semânticos;

2. 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;

3. Os termos são apresentados ao engenheiro, que exclui o que considera irrelevante;

4. A lista final é utilizada como entrada para os métodos de extração de termos com-

plexos.

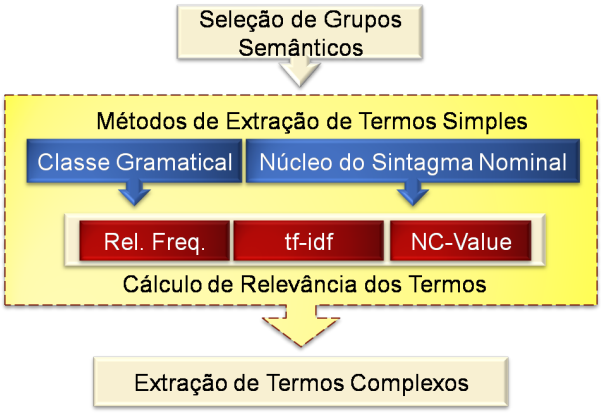


Figura 13: Processo de extração de termos simples.

Conforme a figura, esta etapa possui dois métodos de seleção de termos: Classe Gramatical e Núcleo do Sintagma Nominal. Ambos utilizam informações lingüísticas para executar a tarefa. O primeiro possibilita que o engenheiro selecione quais classes gramaticais deseja extrair do corpus. O segundo extrai apenas termos considerados núcleo de um sintagma nominal. Esse método parte da idéia que sintagmas nominais constituem uma unidade significativa dentro de uma oração, organizando-se ao redor de um núcleo (seção 2.3.2). Sendo assim, restringir a extração somente aos núcleos dos sintagmas pode diminuir a ocorrência de palavras indesejadas durante a tarefa e servir como indicativo da importância de um termo para o domínio. Ao contrário do método anterior, o Núcleo do Sintagma Nominal não dá a opção de selecionar quais classes gramaticais serão extraídas. Nesse caso são considerados apenas substantivos, visto que essa classe constitui a maioria dos conceitos de uma ontologia.

Algumas restrições foram impostas aos métodos de extração de termos simples:

- Forma canônica: os termos são extraídos sem flexões de gênero e número. Essa restrição minimiza a probabilidade das flexões interferirem na relevância de um conceito. Por exemplo, as palavras “professores”, “professora” e “professoras” são normalizadas para o termo “professor”. Sendo assim, a relevância estatística do

termo será a soma das ocorrências das palavras flexionadas. A mesma estratégia é utilizada em trabalhos como (BASÉGIO, 2006; FRANTZI; ANANIADOU; TSUJII, 1998).

- Tamanho e Tipo da palavra: essa restrição é composta por duas heurísticas: tamanho da palavra, que ignora palavras com tamanho menor que três caracteres, e tipo da palavra, que restringe as constituídas por caracteres não alfanuméricos. Sendo assim, formações como "tel.", "&", "1h", "2h", entre outras, são descartadas da lista de termos. Em experimentos preliminares utilizando corpora de Turismo e Ecologia (explicados no capítulo 5) os métodos extraíram formas como "%", "$" e "1º". Com a aplicação das heurísticas essas palavras foram excluídas sem haver perda no total de termos corretos (Anexo A, Tabela 31). Obviamente a quantidade de palavras excluídas foi pequena quando comparada ao total de termos extraídos, entretanto, algumas dessas formas possuíam freqüência alta, o que influenciou diretamente na precisão dos métodos. É importante salientar que essa restrição, apesar dos bons resultados em ambos os corpora, pode influenciar negativamente domínios caracterizados pela presença de termos pequenos, por exemplo, na Química, onde os elementos da tabela periódica podem ser considerados conceitos mesmo com comprimento igual a um.

Para o cálculo de relevância dos termos foram implementados três métodos: FR, *tf-idf* e *NC-Value*. O desenvolvimento dessas medidas seguiu as especificações definidas na seção 2.4.1.

3.2.3 Extração de Termos Complexos

Essa etapa visa extrair termos constituídos por duas ou mais palavras. A Figura 14 apresenta o processo utilizado para a execução da tarefa, que ocorre da seguinte forma:

1. O módulo recebe a lista de Grupos Semânticos (opcional) e a lista de termos simples das etapas anteriores;

2. Os métodos realizam a extração somente dos termos complexos que possuem: (1) uma palavra pertencente a um grupo semântico da lista de entrada (caso esteja habilitado o método) e/ou (2) uma palavra presente na lista de termos simples de entrada;

3. As listas de termos geradas são repassadas aos módulos responsáveis pelo cálculo das medidas de relevância;

4. Os termos são apresentados ao engenheiro em ordem decrescente de relevância para que ele exclua os não têm relação com o domínio.

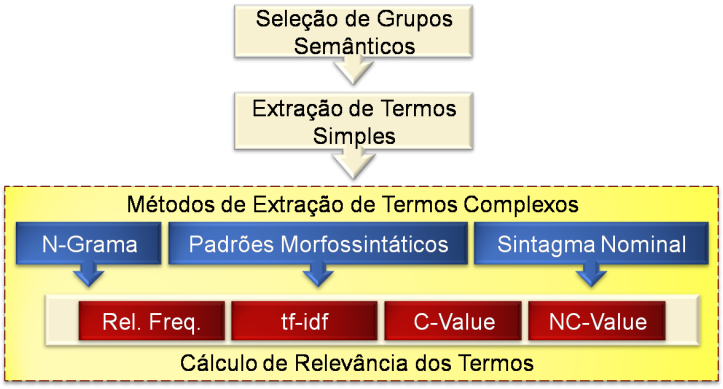


Figura 14: Processo de Extração de Termos Complexos

Para a extração de termos complexos foram implementados três métodos:

- N-Grama: baseia-se no método de mesmo nome apresentado na seção 2.4.1. Sua execução percorre o documento extraindo 'n' palavras por vez, onde n>1. Geralmente esse método é aplicado em texto cru, sem anotação lingüística, o que gera um grande número de termos irrelevantes. Para minimizar esse problema, neste trabalho o método foi adaptado para tirar proveito das informações lingüísticas do corpus. Dessa forma, são extraídos somente os termos pertencentes às classes gramaticais que usualmente constituem conceitos de uma ontologia. A implementação do método possibilita ao engenheiro definir quais classes deseja considerar. Caso o engenheiro execute o método sem alterações, serão considerados somente termos

formados por substantivo, adjetivo, preposição e/ou artigo. Cabe salientar que o método não restringe a ordem de ocorrência entre as classes. Por exemplo, para bigramas podem ser extraídas tanto formações como "adjetivo_substantivo", quanto "substantivo_adjetivo", além das outras possíveis combinações entre classes. Estruturas que dificilmente representam termos também podem ser extraídas, como "preposição_artigo". Contudo, elas são mais fáceis de serem filtradas através de heurísticas simples, como as apresentadas adiante.

- Padrões Morfossintáticos: esse método utiliza uma abordagem semelhante às expressões regulares. O documento é percorrido em busca da ocorrência de determinados padrões morfológicos. Na seção 2.4.1 foram apresentados os padrões de (FRANTZI; ANANIADOU; TSUJII, 1998), enquanto na seção 2.5.2 estão os padrões de (BASÉGIO, 2006). Em experimentos iniciais realizados neste trabalho foram utilizados ambos os padrões. Como resultado, tanto os de Frantiz quanto os de Baségio extraíram a mesma lista de termos. Portanto, optamos pelo segundo, por serem específicos para o Português.

- Sintagma Nominal: esse método extrai apenas termos que compõem todo ou parte de um sintagma nominal. Como sintagmas nominais podem possuir estruturas um tanto complexas, como no exemplo "Luiz Inácio Lula da Silva, Presidente da República", foram definidas algumas restrições para lidar com tais ocorrências. Cabe salientar que esse método está sendo originalmente proposto neste trabalho.

Para melhorar o processo de extração foram criadas algumas heurísticas que restringem os termos selecionados de acordo com as características de cada método:

- Forma canônica: visa auxiliar a definição de relevância do termo. Essa restrição é exatamente igual à utilizada nos métodos de extração de termos simples, desconsiderando as flexões de gênero e número. Entretanto, quando aplicada a termos complexos, torna-se necessária a revisão dos termos quando terminada a ontologia. Isto por que, termos como "pensão completa" e "companhia aérea" em sua

forma canônica são “pensão completo” e “companhia aéreo”, perdendo parte do seu sentido.

- Preposição e artigo: nesse caso são utilizadas duas heurísticas diferentes que trabalham de forma semelhante: (1) preposição e (2) artigo. A primeira é aplicada somente ao método N-Grama, e serve para exclusão de termos irrelevantes. A restrição avalia se o candidato inicia ou termina com uma preposição, em caso verdadeiro, ele é descartado. Em experimentos realizados com um corpus de Turismo sem a utilização da heurística, termos como “com café” (preposição_substantivo) e “apartamento com” (substantivo_preposição) foram ranqueados entre os dez mais relevantes, prejudicando consideravelmente a precisão. A segunda heurística é aplicada aos métodos N-Grama e Sintagma Nominal. No primeiro ela tem a mesma finalidade da anterior. Em testes aplicados aos corpora de Ecologia e Turismo a quantidade de termos extraídos diminuiu consideravelmente, sem haver perda na quantidade de termos corretos, ou seja, foram excluídos apenas termos “mal formados” (Anexo A, Tabela 32). No segundo, visa melhorar a qualidade dos termos extraídos. Essa restrição, quando aplicada ao Sintagma Nominal, exclui apenas o artigo que aparecer no início do termo. Por exemplo, o termo “o passeio de barco” (artigo_substantivo_preposição_substantivo) resulta em “passeio de barco”.

- Preposição+artigo: essa restrição é aplicada a todos os métodos. Ela admite que uma preposição seguida de um artigo dentro do termo deve ser considerada como uma única palavra. Por exemplo, o termo “escola_de_a_comunidade” (substantivo_preposição_artigo_substantivo) é aceito como o trigrama “escola_**da**_comunidade”. Essa heurística foi necessária devido à anotação gerada pelo PALAVRAS, que realiza a separação dessas contrações.

- Pontuação: essa restrição é aplicada ao método Sintagma Nominal. Como em alguns casos os SNs podem possuir internamente caracteres de pontuação, a restrição descarta aquilo que estiver depois do caractere, adicionando somente a parte ante-

rior à lista de termos. Por exemplo, o sintagma “Baile Finlandês, que acontece toda a noite no clube Finlândia” é adicionado à lista de termos candidatos apenas como “Baile Finlandês”.

- Tamanho e Tipo da palavra: essa restrição é aplicada a todos os métodos. Assim como para os termos simples, ela é utilizada para a exclusão de termos irrelevantes ao domínio, descartando os que possuem palavras constituídas por caracteres não alfanuméricos ou de tamanho menor que três, por exemplo, “% de arroz”, “1h da manhã”, entre outros. Em experimentos realizados nos domínios de Turismo e Ecologia (Anexo A, Tabela 33) a aplicação dessas heurísticas obteve bons resultados, mantendo a quantidade total de termos corretos e diminuindo a total de termos extraídos pelos métodos.

A Tabela 12 apresenta uma visão geral das restrições, suas funções e em quais métodos elas são aplicadas.

Tabela 12: Visão geral das restrições aplicadas aos métodos de extração de termos complexos.

Método	**Restrição**	**Função**
N-Grama	Forma Canônica	Auxiliar na definição de relevância do termo
	Preposição e Artigo	Exclusão de termos “mal formados”
	Preposição+Artigo	Melhorar a qualidade dos termos extraídos
	Tamanho e tipo da palavra	Exclusão de termos “mal formados”
Sintagma Nominal	Forma Canônica	Auxiliar na definição de relevância do termo
	Artigo	Melhorar a qualidade dos termos extraídos
	Preposição+Artigo	Melhorar a qualidade dos termos extraídos
	Pontuação	Melhorar a qualidade dos termos extraídos
	Tamanho e tipo da palavra	Exclusão de termos “mal formados”
Padrões Morfossintáticos	Forma Canônica	Auxiliar na definição de relevância do termo
	Preposição+Artigo	Melhorar a qualidade dos termos extraídos
	Tamanho e tipo da palavra	Exclusão de termos “mal formados”

Para o cálculo de relevância dos termos complexos foram utilizadas quatro medidas estatísticas: FR, *tf-idf*, *C-Value* e *NC-Value*. Assim como na etapa de extração de termos simples, esses métodos seguem as definições explicadas na seção 2.4.1.

3.3 Organização Hierárquica dos Termos

Conforme destacado na seção 2.4, à medida que avançamos nas etapas de construção de ontologias a complexidade de automatizar a tarefa cresce consideravelmente. Para auxiliar os métodos das fases posteriores, devemos focalizar o esforço na qualidade da saída das etapas anteriores. A maioria das abordagens, como já destacado, utiliza o *feedback* do engenheiro de ontologia ao final de cada etapa, sendo essa a estratégia adotada neste trabalho.

O processo de construção das estruturas taxonômicas é apresentado na Figura 15.

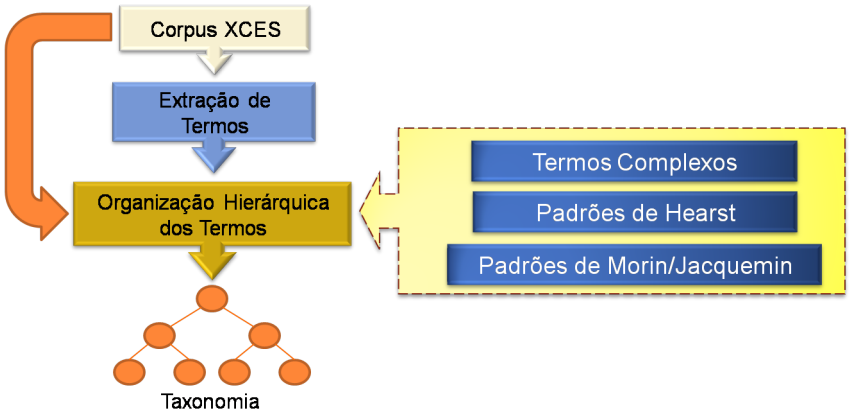


Figura 15: Processo de Organização Hierárquica dos Termos.

Para a realização da tarefa foram implementados três métodos: Termos Complexos, Padrões de Hearst e Padrões de Morin/Jacquemin, descritos nas seções abaixo. Cabe salientar que no plug-in desenvolvido as ontologias extraídas por esses métodos podem ser editadas pelo engenheiro, tornando-as mais consistentes.

3.3.1 Organização Baseada em Termos Complexos

Esse método recebe como entrada uma lista de termos simples e uma lista de termos complexos. Sua execução busca ocorrências de um termo simples dentro dos complexos. Quando alguma é encontrada, o termo complexo selecionado é organizado como hipônimo do termo simples, conforme descrito na seção 2.4.2 e apresentado na Figura 16.

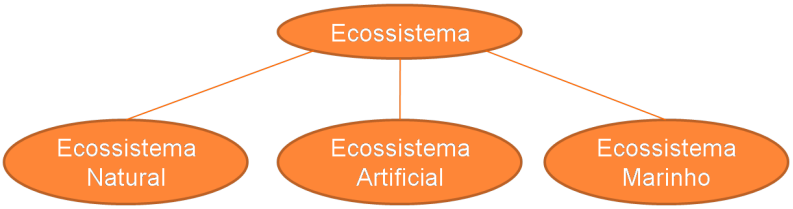


Figura 16: Exemplo de taxonomia gerada pelo método Termos Complexos.

Para obter melhores resultados durante a identificação das relações taxonômicas, no protótipo implementado, foram disponibilizadas algumas configurações que serão explicadas no capítulo 4.

3.3.2 Organização baseada nos Padrões de Hearst e Morin/Jacquemin

O método apresentado anteriormente extrai relações taxonômicas simples, identificadas através das palavras que compõem o termo. Os métodos baseados em padrões, conforme descrito na seção 2.4.2, utilizam estruturas frasais que servem de pistas para a identificação de relações hierárquicas não triviais.

Os padrões utilizados neste trabalho foram os propostos por Hearst e Morin/Jacquemin, ambos adaptados por (BASÉGIO, 2006) (seção 2.4.2) para o Português. Em seu trabalho, Baségio substituiu os Sintagmas Nominais por substantivos, isto por que o corpus utilizado não possuia informações sobre essas estruturas. Neste trabalho, apesar do parser PALAVRAS disponibilizar tais informações, também foi desconsiderado o uso de SN. Em experimentos preliminares a sua aplicação prejudicou a identificação das relações hierárquicas. Com base nos resultados constatou-se que alguns sintagmas não estavam devidamente delimitados, dificultando a execução dos métodos. Para possibilitar a extração de relações hierárquicas constituídas por termos complexos, os substantivos, nos padrões adaptados, foram substituídos por substantivos e/ou o método Padrões Morfossintáticos. Utilizando essa estratégia, qualquer termo (simples ou complexo) pode ser identificado como conceito de uma relação. Por exemplo, a relação 'hiponímia("habitat_terrestre",

"campo")', caso pertencente a uma estrutura frasal equivalente a algum padrão, seria selecionada pelas estruturas aplicadas neste trabalho.

A Tabela 13, na coluna "Hearst/Baségio", apresenta os padrões de Hearst adaptados por Baségio, enquanto na coluna "Baségio/Termos Complexos" estão os adaptados para extraírem termos simples e complexos (T). A mesma estrutura é utilizada na Tabela 14 para os padrões de Morin/Jacquemin adaptados por Baségio. Cabe salientar que os padrões de Morin/Jacquemin equivalentes aos de Hearst (seção 2.5.2) foram desconsiderados, sendo utilizados apenas as generalizações propostas por Baségio, mais o padrão 10.

Tabela 13: Padrões de Hearst adaptados por Baségio e alterados para extrair termos complexos.

Hearst/Baségio	**Baségio/Termos Complexos**
SUB como {(SUB,)*(ou—e)} SUB	T como {(T,)*(ou—e)} T
SUB tal(is) como {(SUB,)*(ou—e)} SUB	T tal(is) como {(T,)*(ou—e)} T
tal(is) SUB como {(SUB,)*(ou—e)} SUB	tal(is) T como {(T,)*(ou—e)} T
SUB {,SUB}* {,} ou outro(s) SUB	T {,T}* {,} ou outro(s) T
SUB {,SUB}* {,} e outros SUB	T {,T}* {,} e outros T
SUB {,} incluindo {SUB,}*{ou—e} SUB	T {,} incluindo {T,}*{ou—e} T
SUB {,} especialmente {SUB,}*{ou—e} SUB	T {,} especialmente {T,}*{ou—e} T
SUB {,} principalmente {SUB,}*{ou—e} SUB	T {,} principalmente {T,}*{ou—e} T
SUB {,} particularmente {SUB,}*{ou—e} SUB	T {,} particularmente {T,}*{ou—e} T
SUB {,} em especial {SUB,}*{ou—e} SUB	T {,} em especial {T,}*{ou—e} T
SUB {,} em particular {SUB,}*{ou—e} SUB	T {,} em particular {T,}*{ou—e} T
SUB {,} de maneira especial {SUB,}*{ou—e} SUB	T {,} de maneira especial {T,}*{ou—e} T
SUB {,} sobretudo {SUB,}*{ou—e} SUB	T {,} sobretudo {T,}*{ou—e} T

Tabela 14: Padrões de Morin/Jacquemin adaptados por Baségio e alterados para extrair termos complexos.

Padrões de Morin/Jacquemin adaptados	**Padrões Adaptados para Termos Complexos**
1- SUB1 (LIST_SUB2)	1- T1 (LIST_T2)
2- SUB1: LIST_SUB2	2- T1: LIST_T2
10- SUB1 e (notadamente-em particular) SUB2	10- T1 e (notadamente-em particular) T2

No protótipo desenvolvido foram definidas duas configurações possíveis para a organização hierárquica através de padrões:

- Restrição por termos: quando habilitada são extraídas apenas relações onde no mínimo um dos conceitos está presente nas listas de termos simples e complexos

selecionadas pelos métodos de extração de termos.

- Restrição por Grupos Semânticos: extrai somente as relações onde os conceitos são pertencentes a um mesmo grupo semântico. Apesar dos métodos baseados em padrões possuírem boa precisão, em alguns casos são extraídas relações que não são relevantes para o domínio. Sendo assim, essa restrição tenta minimizar essas ocorrências. Cabe salientar que essa heurística só é aplicada caso a restrição por termos esteja desabilitada.

Ao final da execução dessa última etapa o engenheiro pode exportar as taxonomias inferidas para a interface principal de construção de ontologias do Protégé. Essas e outras funcionalidades do protótipo desenvolvido são apresentadas no próximo capítulo.

Capítulo 4

Plug-in OntoLP

Este capítulo apresenta o sistema OntoLP, construído no contexto deste trabalho e desenvolvido com base na metodologia descrita no capítulo 3. O sistema foi implementado como um plug-in para o ambiente de auxílio a criação de ontologias Protégé, exposto na seção abaixo.

4.1 Protégé

O Protégé (Figura 17) é um editor de ontologias que vem sendo aprimorado por mais de duas décadas, a primeira versão do sistema data de 1987. Apesar de todas as atualizações o objetivo principal do Protégé mantém-se o mesmo, possibilitar o uso de meta-conhecimento para a criação de ferramentas de aquisição de conhecimento utilizáveis.

As principais funções do sistema são:

- Modelar classes: possui uma interface gráfica que possibilita a definição de classes, seus atributos e relacionamentos;
- Criar instâncias: instanciação de entidades a partir das classes definidas;

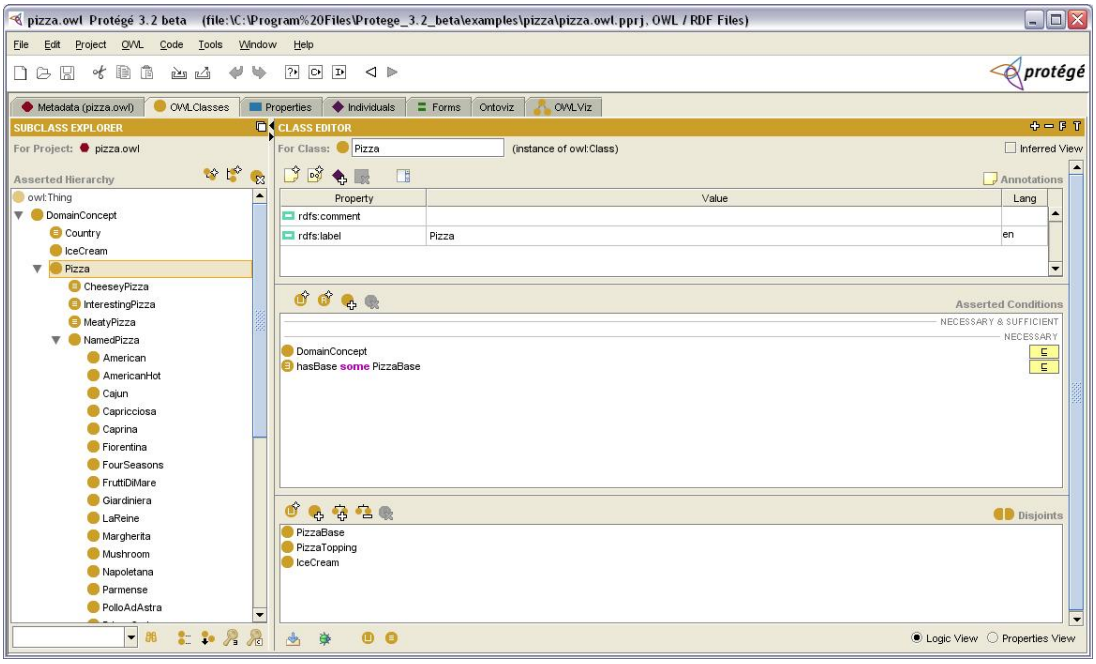


Figura 17: Interface do ambiente Protégé.

- Realizar processamento sobre a ontologia: possui uma biblioteca de plug-ins que permite criar consultas e comportamentos através de definições lógicas;

- Exportar a ontologia em diferentes formatos: os modelos resultantes podem ser exportados em vários formatos, possibilitando a integração com diversas aplicações.

Existem outros editores de ontologias disponíveis atualmente, entre os mais conhecidos estão o Kaon[1] e o OntoEdit[2]. O primeiro é uma ferramenta *open-source* que possibilita a integração de plug-ins para acrescentar funcionalidades ao ambiente. O segundo tem recursos semelhantes aos do Protégé, entretanto, está vinculado ao ambiente OntoStudio[3], um pacote de softwares pago para desenvolvimento e manutenção de ontologias.

A escolha pelo ambiente Protégé ocorreu por diferentes motivos, entre eles: por ser uma ferramenta *open-source*, por sua grande aceitação no meio acadêmico, pela crescente comunidade que contribui com o seu desenvolvimento e pelas constantes atualizações.

[1]http://kaon.semanticweb.org/frontpage
[2]http://www.ontoknowledge.org/tools/ontoedit.shtml
[3]http://ontoedit.com/

Além disso, o sistema possui suporte completo à linguagem de construção de ontologias Web (OWL). O que possibilita que as ontologias construídas com o auxílio do protótipo sejam exportadas diretamente nesse formato, estando de acordo com os padrões definidos pela W3C para a Web Semântica.

4.2 OntoLP

Assumindo o princípio de que o conhecimento sobre um domínio pode estar representado em bases textuais. O OntoLP implementa um conjunto de métodos que auxiliam o engenheiro durante as etapas iniciais de construção de ontologias: extração e organização hierárquica de termos. O plug-in emprega a metodologia descrita no capítulo 3, somando a ela um conjunto de funcionalidades que visam auxiliar o engenheiro nas fases que necessitam de interação. A estrutura de execução do plug-in é apresentada na Figura 18.

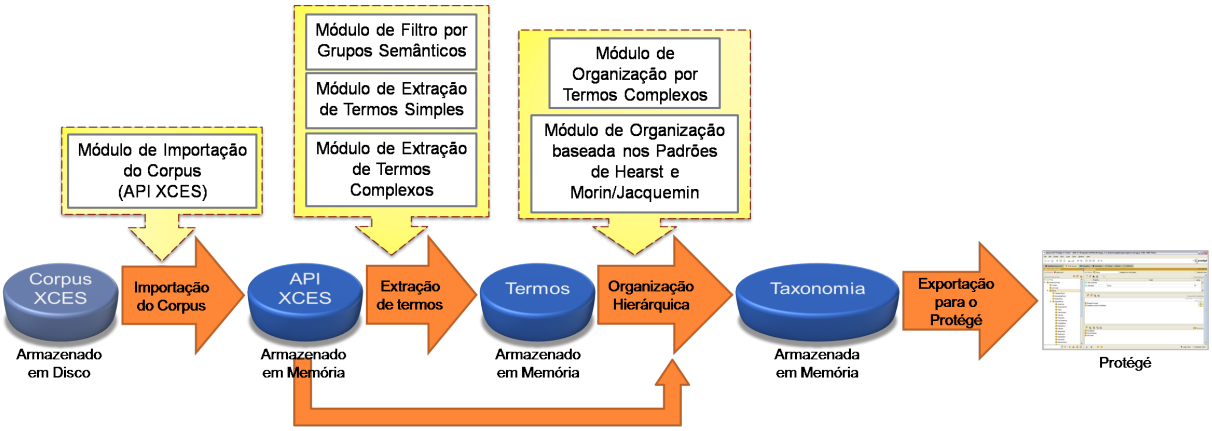


Figura 18: Processo de execução do plug-in OntoLP.

De uma forma geral, a execução do plug-in ocorre segundo os passos descritos abaixo:

- O engenheiro seleciona o corpus previamente anotado como entrada para o sistema;
- O sistema carrega o corpus através do "Módulo de Importação do Corpus". Esse

módulo utiliza uma API XCES desenvolvida no LEL que extrai as informações dos textos e as armazena em uma estrutura de dados para serem consultadas.

- O engenheiro executa a extração dos termos candidatos a conceito. Nessa etapa estão disponíveis três módulos: Filtro por Grupo Semântico, Extração de Termos Simples e Extração de Termos Complexos. Os passos de execução da tarefa são descritos na Tabela 15.

Tabela 15: Seqüência de execução da extração de termos.

Processo de Extração
Extração automática dos Grupos Semânticos e filtragem manual deles. (opcional)
Extração automática dos Termos Simples, utilizando os Grupos Semânticos como filtro (caso habilitados), e edição manual da lista gerada.
Extração automática dos Termos Complexos, utilizando os Grupos Semânticos (caso habilitados) e a Lista de Termos Simples como filtro, e edição manual da lista gerada.

- O sistema armazena as listas de termos extraídas em uma estrutura de dados específica. Para cada termo é registrado também um conjunto de informações que auxiliam o engenheiro durante o processo de edição das listas, descritas na seção 4.2.1.

- O engenheiro executa os métodos de Organização Hierárquica dos termos. Esta etapa é constituída por dois módulos: Organização por Termos Complexos e Organização baseada nos Padrões de Hearst e Morin/Jacquemin. A execução dos módulos de organização hierárquica não possui ordem predefinida, podendo ser feita conforme descrito abaixo:

  - Módulo de Organização por Termos Complexos: o engenheiro seleciona uma lista de termos simples e uma lista de termos compostos. Essas listas são utilizadas como entrada para o módulo, que executa a organização hierárquica dos termos conforme o método Termos Complexos (seção 3.3.1);

  - Módulo de Organização baseada em Padrões de Hearst e Morin/Jacquemin: esse módulo corresponde aos padrões de organização hierárquica de termos apresentados na seção 3.3.2. Durante a execução a estrutura de dados da API XCES é consultada em busca de estruturas frasais e informações lingüísticas que combinem com os padrões.

- O engenheiro edita as taxonomias construídas automaticamente, caso considere necessário;

- O engenheiro exporta as taxonomias geradas para a interface principal do sistema Protégé.

Nas seções 4.2.1, 4.2.2 são descritas as interfaces de Extração de Termos e Organização Hierárquica respectivamente. A Interface do Módulo de Importação do Corpus é apresentada no Anexo B.

4.2.1 Interface do Módulo de Extração de Termos

A Interface desenvolvida para esse módulo divide sua execução em três partes. Cada parte está relacionada a uma etapa de extração de termos presente na metodologia proposta (capítulo 3). A relação entre o processo e o módulo é apresentada na Figura 19.

A aba de extração de termos adapta-se à tarefa em execução. Quando o engenheiro está aplicando a Seleção de Grupos Semânticos são apresentadas informações para auxiliá-lo no processo (Figura 20).

Na Figura, a coluna 1 contém uma tabela com os grupos semânticos extraídos do corpus, as definições de cada grupo e a relevância deles. Esse componente é responsável pela parte visual do método Filtro por Grupos Semânticos. Sendo assim, após a extração dos grupos o engenheiro deve excluir o que considera irrelevante para o domínio. Nas primeiras versões do plug-in a única informação de auxílio à tarefa de exclusão eram as

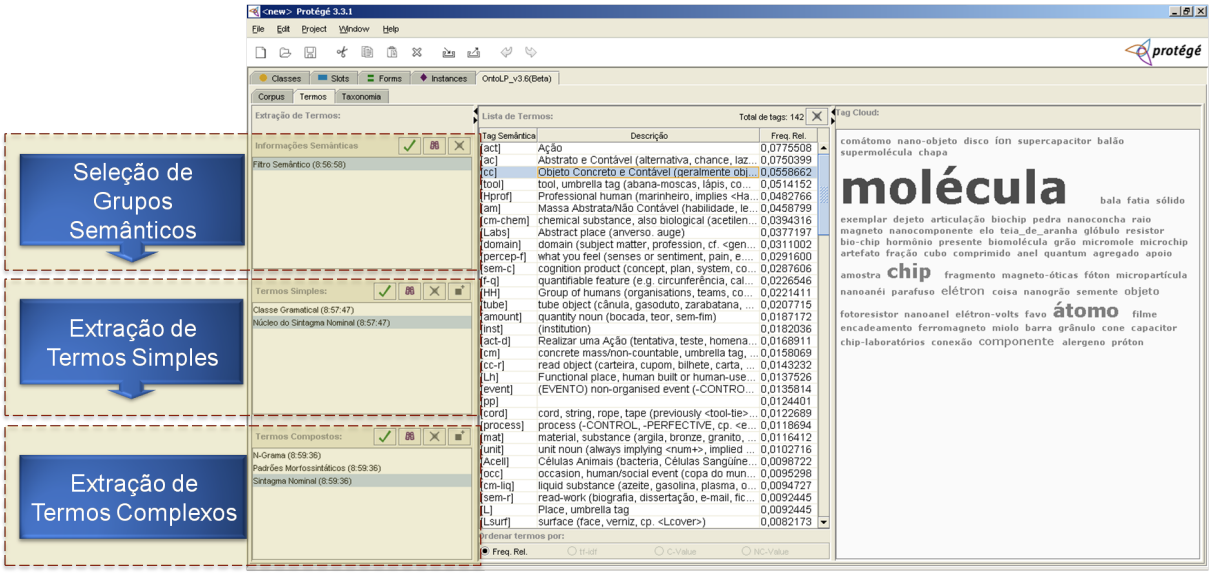


Figura 19: Mapeamento da metodologia proposta com as etapas da interface.

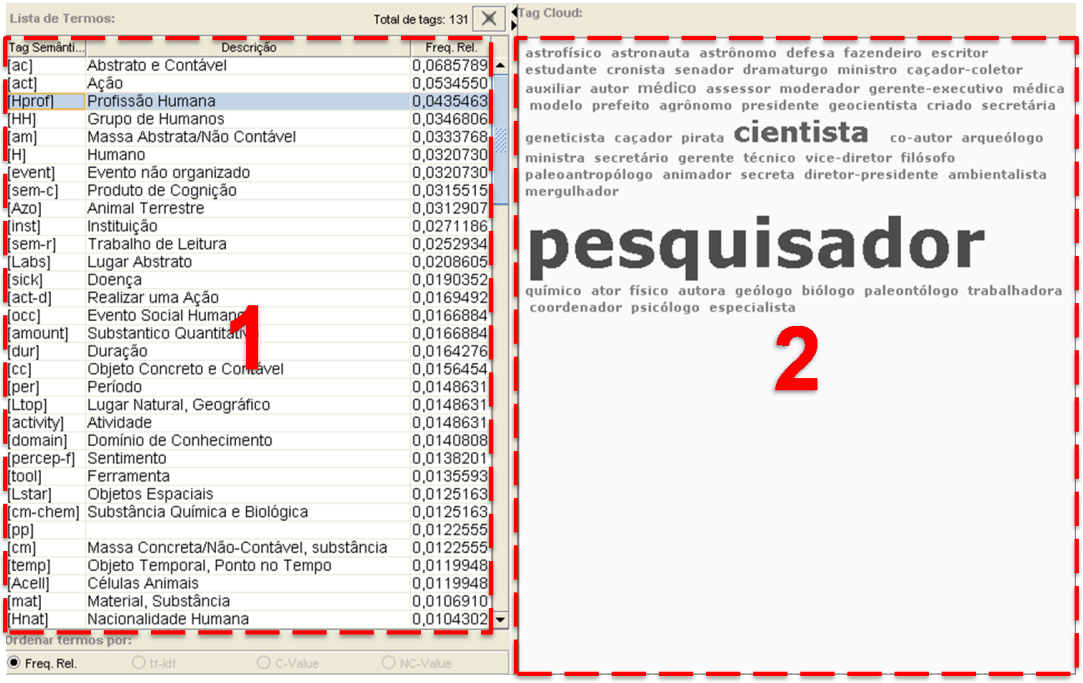


Figura 20: Interface de visualização das informações semânticas.

definições dos grupos[4]. Entretanto, foi constatado que em alguns casos elas não indicavam claramente seu significado, como para os grupos "<ac>" e "<cc>", definidos como "Abstract Countable" (Contável Abstrato) e "Concrete Countable Object" (Objeto Contável Concreto) respectivamente.

Para complementar as definições foram agregados dois métodos de auxílio ao en-

[4]http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags_nouns

genheiro:

- Ordem de relevância: os grupos são apresentados em ordem decrescente de relevância (FR). Isso define quais tiveram maior número de palavras ocorridas no corpus, servindo de indicação para importância do grupo no domínio.

- *Tag Clouds*: é um conceito que vem sendo aplicado em algumas páginas Web. A idéia é auxiliar a navegação do usuário no site, disponibilizando os termos mais relevantes presentes em seus documentos. Cada termo possui um link para as páginas que tratam sobre o assunto e sua relevância é definida pelo tamanho da fonte aplicada. A Figura 21 apresenta um exemplo de *tag cloud* para o domínio Web 2.0.

Figura 21: *Tag cloud* com termos relacionados à Web 2.0 (http://en.wikipedia.org/wiki/Tag_cloud).

Além disso, as *tag clouds* também servem de indício dos assuntos abordados em um site. Na Figura de exemplo, o termo "Web 2.0" é o de maior destaque justamente por ser o foco principal da página. Com essa mesma finalidade o conceito está sendo aplicadas aos grupos semânticos. Quando um determinado grupo é selecionado os termos pertencentes a ele são apresentados ao engenheiro utilizando esse método. A coluna 2 da Figura 20 apresenta a *tag cloud* referente ao grupo "<Hprof>" (Profissional Humano), no qual, para o domínio de Ecologia, o termo "pesquisador" foi o mais relevante.

Além da Seleção de Grupos Semânticos, a interface adapta-se também aos métodos de extração de termos (Figura 22).

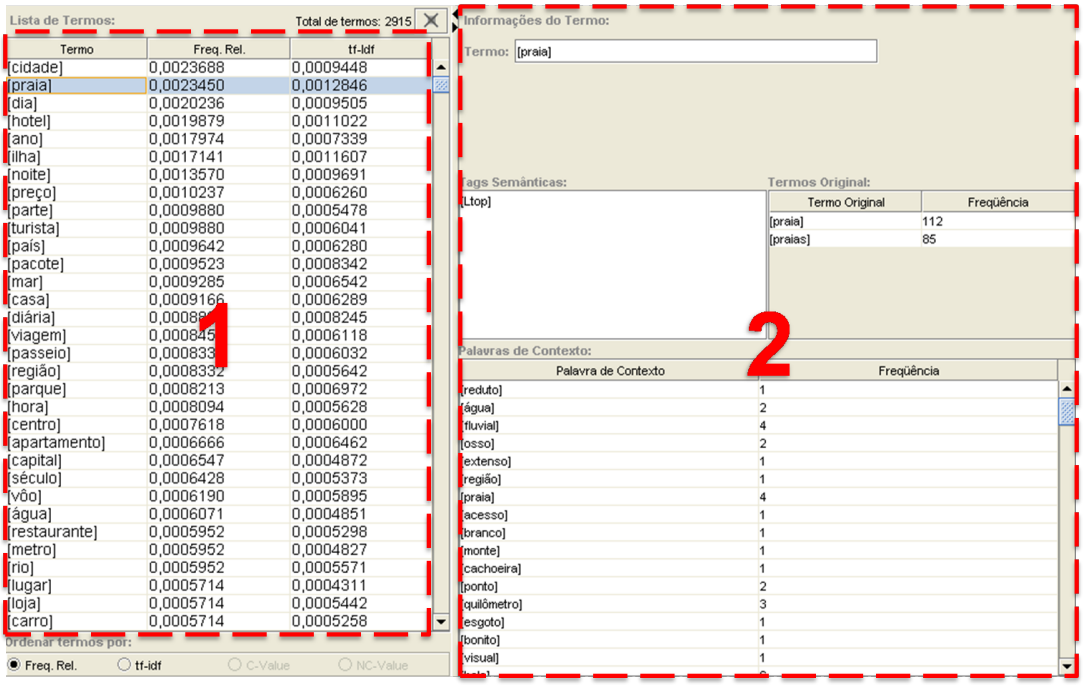


Figura 22: Interface de visualização da extração de termos.

Na coluna 1 da Figura é apresentada a lista de termos extraídos para o método Núcleo do Sintagma Nominal. Já a coluna 2 disponibiliza informações sobre o termo "praia", selecionado a partir da lista. Cada uma delas é explicada abaixo:

- Tags Semânticas: apresenta o(s) grupo(s) semântico(s) ao(s) qual(is) pertence o termo.

- Termos Originais: o OntoLP realiza a extração dos termos na forma canônica. Essa estratégia causa, como citado anteriormente, a perda de significado em alguns casos ("companhia aérea" → "companhia aéreo"). Para que o engenheiro conheça as flexões com que o termo ocorreu, o plug-in apresenta as formas originais extraídas do corpus e sua freqüência de ocorrência. Para o termo "praia", no exemplo da Figura, foram encontradas {praia, 112} e {praias, 85}. Isso auxilia o engenheiro a definir a forma mais apropriada a ser utilizada na ontologia.

- Palavras de Contexto: essas palavras são originalmente utilizadas para o cálculo *NC-Value*. Entretanto, foram disponibilizadas na interface com dois objetivos: auxiliar na resolução de ambigüidades dos termos e na identificação de termos complexos. No primeiro, palavras de contexto como "financeiro", "Brasil", "real" etc., quando vizinhas de "banco", indicam que no corpus o termo foi abordado como instituição financeira. Apesar das informações semânticas teoricamente auxiliarem nesse processo, o PALAVRAS não trata ambigüidades, atribuindo em alguns casos mais de uma tag semântica. No segundo, um valor alto de co-ocorrência entre o termo e uma determinada palavra de contexto, pode indicar a formação de um termo complexo. Por exemplo, o termo "diária" no domínio de Turismo teve "casal" como palavra de contexto de maior freqüência. Isto sugere a existência do termo "diária_para_casal" neste domínio.

Conforme descrito anteriormente, o plug-in permite ao engenheiro configurar a estratégia de extração de termos de diferentes formas. Na Figura 23 são apresentados os painéis de configuração da Seleção de Grupos Semânticos e de Termos Complexos. Através deles é possível desabilitar o uso de informações semânticas e a extração de termos complexos com base na lista de termos simples. Na Tabela 23 são destacadas as opções responsáveis por essas alterações.

Tabela 16: Configurações necessárias de acordo com a abordagem utilizada.

Opção	**Função**
Habilitar Filtro Semântico	Habilita a execução dos Métodos Semânticos, servindo de entrada para as etapas de extração de termos simples e complexos.
Restrição por unigrama	Obriga a definição de uma lista de termos simples como entrada para a extração de termos complexos.

O plug-in disponibiliza também uma interface de configuração para os termos simples. A lista completa das configurações possíveis para a interface de extração de termos é apresentada no Anexo B.

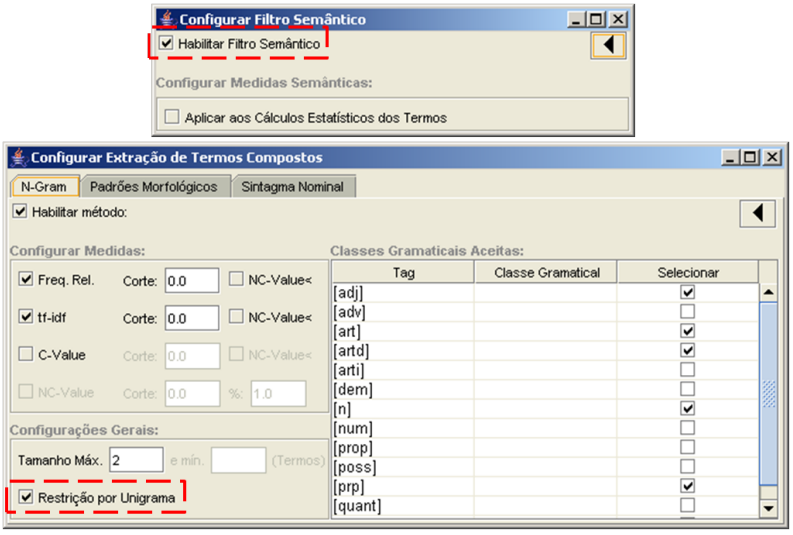


Figura 23: Interface de Configuração dos Métodos Semânticos e Extração de Termos Complexos.

4.2.2 Interface do Módulo de Organização Hierárquica

Cada uma das etapas de execução do Módulo de Organização Hierárquica dos termos está relacionada a um dos métodos descritos no capítulo 3. Essa relação é apresentada na Figura 24.

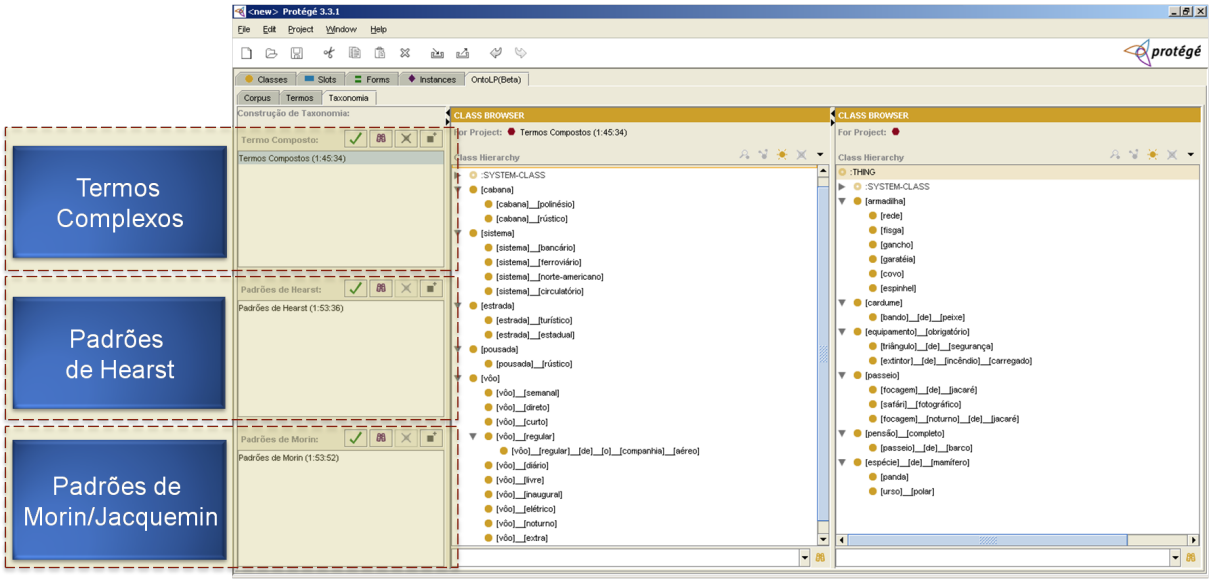


Figura 24: Mapeamento entre os métodos e as etapas de organização hierárquica dos termos.

A Figura 25 demonstra a divisão da Interface de Organização Hierárquica. A primeira coluna é responsável pela execução e configuração dos métodos, e pela exportação

dos resultados obtidos para a interface principal do Protégé. A segunda coluna exibe a ontologia extraída pela etapa selecionada, no exemplo, Termos Complexos. Essa coluna permite a edição, inserção e exclusão de conceitos presentes nas taxonomias geradas automaticamente. A última coluna instancia a interface principal de edição do Protégé (aba "Classes"). Portanto, o usuário pode visualizar as alterações na ontologia principal sem sair da aba do plug-in.

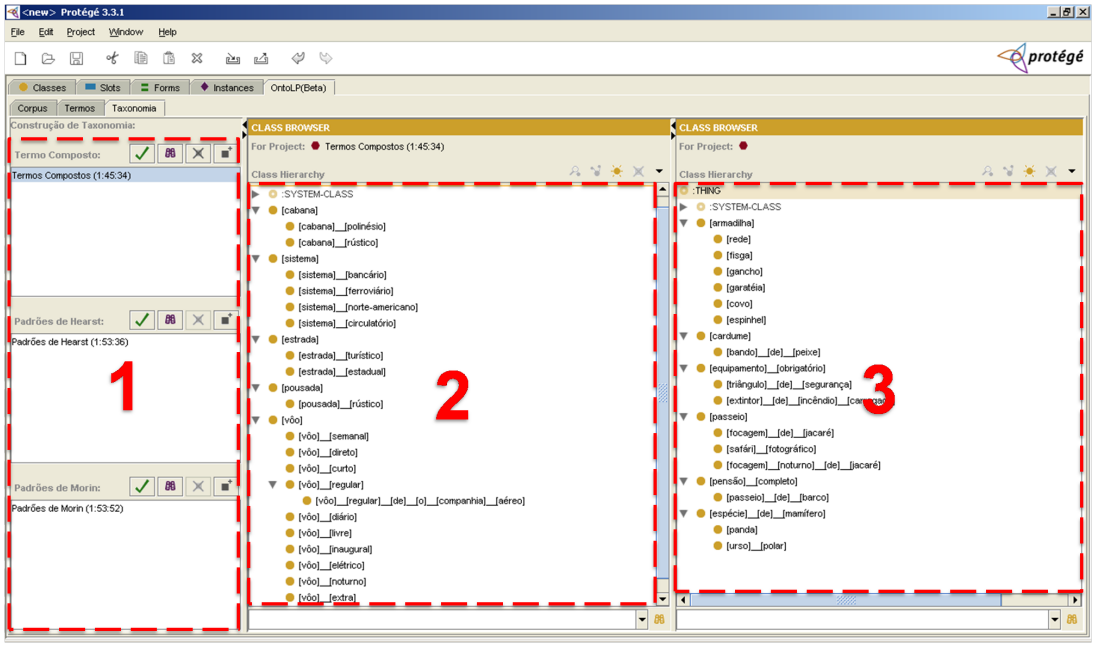


Figura 25: Divisão da Interface do Módulo de Organização Hierárquica dos Termos.

Cada método de construção de taxonomia disponível no OntoLP gera uma ontologia própria. Essas ontologias são integradas somente quando exportadas para a interface principal do Protégé (coluna 3). A exportação é realizada individualmente, um método por vez, na ordem que o engenheiro desejar. Como as diferentes taxonomias podem ter conceitos em comum, o plug-in trata essa duplicidade durante o processo. Quando ocorrem, o algoritmo de exportação considera a estrutura da Ontologia Principal (coluna 3) como sendo a correta, mantendo-a inalterada. A execução da tarefa é demonstrada nas Figuras 26, 27 e 28. Na primeira são apresentadas duas ontologias: a) Ontologia extraída pelo método Padrões de Hearst; b) Ontologia Principal. Como a ontologia 'b' encontrasse vazia, o algoritmo simplesmente copia os conceitos da estrutura 'a' para a 'b', gerando o resultado apresentado na Figura.

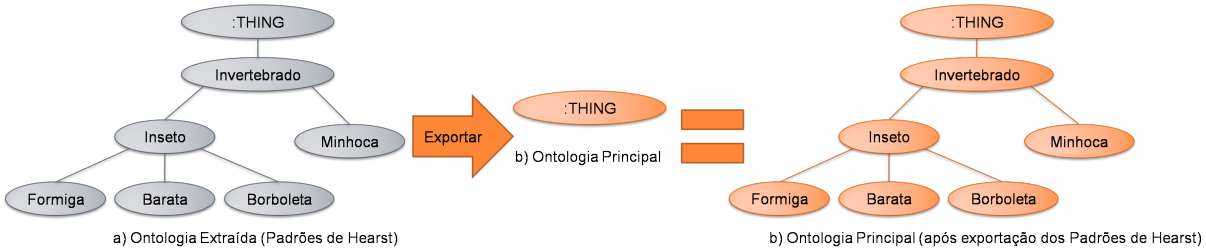


Figura 26: Primeira etapa do processo de exportação das taxonomias extraídas.

A Figura 27 contém a estrutura extraída pelo método Termos Complexos (c) e a Ontologia Principal (b) após a exportação. Como o algoritmo considera a estrutura de 'b' inalterável e ela contém conceitos em comum com 'c', ela recebe somente as classes exclusivas da ontologia 'c'. O resultado do processo é apresentado na Figura 28, onde os conceitos em cinza foram os adicionados a partir de 'c'.

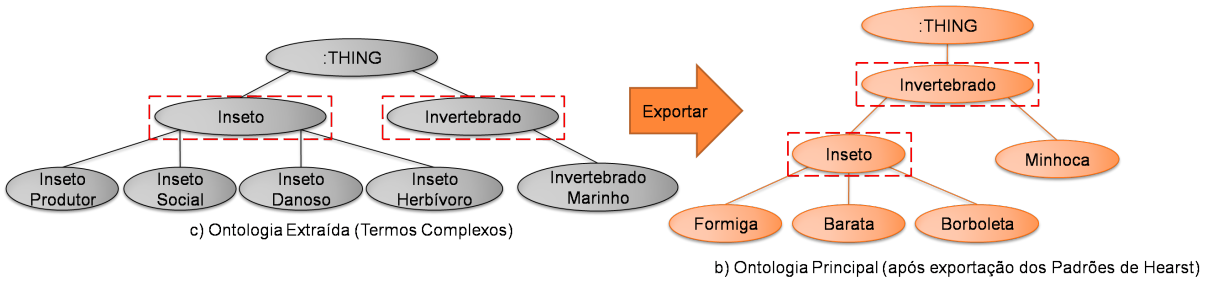


Figura 27: Segunda etapa do processo de exportação das taxonomias extraídas.

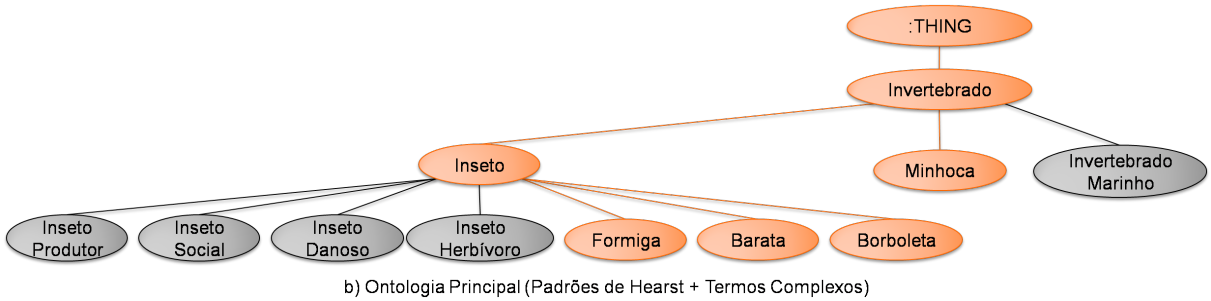


Figura 28: Taxonomia final gerada pela exportação dos Padrões de Hearst seguido dos Termos Compostos.

A opção de considerar a Ontologia Principal como uma estrutura correta (inalterável), faz com que a ordem na qual as taxonomias são exportadas influencie no resultado final. Por exemplo, a ontologia final do processo anterior (Figura 28) possuiria a estrutura apresentada na Figura 29 caso fosse exportada a ontologia 'c' antes da 'a'.

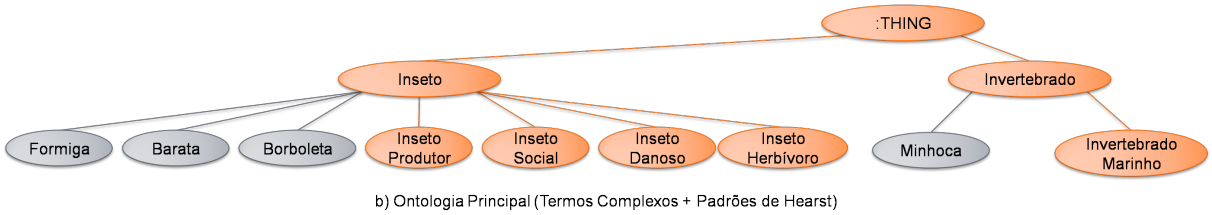


Figura 29: Taxonomia final gerada pela exportação dos Termos Compostos seguido dos Padrões de Hearst.

Essa característica não é considerada um problema e a estratégia foi adotada pensando na construção de ontologias com base em uma estrutura inicial. Nesses casos, o sistema mantém essa estrutura, apenas expandido-a. Além disso, obviamente o engenheiro tem a opção de editar manualmente essa taxonomia caso necessário.

As demais configurações do Módulo de Organização Hierárquica são descritas no Anexo B.

4.3 Questões de Desenvolvimento

O OntoLP foi desenvolvido na linguagem Java utilizando a plataforma JSE versão 1.5.0. Os testes iniciais para correção de problemas e avaliação de desempenho foram executados em um computador de processador Intel Centrino 1.73GHz com 2GB de memória. Durante as avaliações feitas pelos usuários (Capítulo 5) o plug-in foi submetido às configurações de hardware destacadas na Tabela 17.

Tabela 17: Configurações de hardware utilizadas nos experimentos.

Processador	**Memória**
Intel Pentium 4 1.2GHz	256MB
Intel Pentium 4 1.8GHz	256MB
Athlon 64 2.19GHz	1GB

Quanto à compatibilidade com o ambiente Protégé, todos os testes foram executados nas versões 3.3.1 com Sistema Operacional Windows XP.

Capítulo 5

Experimentos

Neste capítulo são descritos os experimentos de avaliação dos métodos e da ferramenta desenvolvida. Os experimentos foram divididos em três partes: (1) avaliação dos métodos de extração de termos; (2) avaliação dos métodos de organização hierárquica dos termos e (3) avaliação do plug-in feita por usuários. As duas primeiras dependem da disponibilidade de um conjunto de recursos constituído por um corpus de domínio e uma ontologia de referência. Para essas duas etapas foi utilizado um corpus e uma ontologia de Ecologia (seção 5.1.1). Os métodos de extração de termos avaliados em (ZAVAGLIA et al., 2007), serviram como *baseline* para a avaliação. Esses recursos são necessários devido às métricas de avaliação empregadas (seções 5.2.1 e 5.2.2).

A última etapa de avaliação foi feita por usuários. Para o experimento foram convidados dois grupos de pesquisa com experiência em extração e organização hierárquica de termos. Cada grupo utilizou um corpus representando um domínio de seu conhecimento, nesse caso, N&N e Pediatria, explicados nas seções 5.1.2 e 5.1.3 respectivamente.

5.1 Corpora de Avaliação

Nesta seção são apresentados os corpora utilizados nos experimentos de avaliação do trabalho.

5.1.1 Corpus de Ecologia (*CórpusEco*)

Neste trabalho, o *CórpusEco* (ZAVAGLIA, 2004), por disponibilizar uma ontologia construída a partir dele, foi utilizado nos experimentos de avaliação dos métodos de extração e organização hierárquica. O corpus é constituído por textos extraídos de parte dos livros "A Economia da Natureza" e "Ecologia", além de revistas presentes no projeto LácioWeb[1]. O *CórpusEco* conta com um total de 260.921 palavras.

Como descreve (ZAVAGLIA et al., 2007), 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[2]. Finalmente, foram eliminados os termos duplicados nas listas e feita a intersecção com o *CórpusEco*, restando um total de 520 termos divididos em 322 unigramas, 136 bigramas e 62 trigramas.

Quanto à construção da taxonomia, foram selecionados três subdomínios de Ecologia: Ecologia de Ecossistemas, Ecologia de Populações e Ecologia de Comunidades. Para cada subdomínio foi construída uma taxonomia, todas baseadas na lista de termos extraídos anteriormente. Conforme citado em (ZAVAGLIA, 2004), a ontologia foi construída para ser aplicada em tarefas como Tradução Automática, Recuperação de Informação, Refinamento de Queries, entre outras.

Além dos recursos disponíveis, o *CórpusEco* foi utilizado ainda em experimentos de extração automática de termos (EAT). Esses experimentos foram realizados com um conjunto de métodos de extração, conforme descrito em (ZAVAGLIA et al., 2007), e seus

[1] http://www.nilc.icmc.usp.br/lacioweb/
[2] http://www.jornaldomeioambiente.com.br/

resultados serviram de *baseline* na avaliação dos métodos utilizados neste trabalho.

5.1.2 Corpus de Nanotecnologia & Nanociência (*NanoTerm*)

O corpus de Nanotecnologia & Nanociência foi utilizado na etapa de avaliação feita pelos usuários. Esse corpus foi desenvolvido pelo Grupo de Estudos e Pesquisas em Terminologia da Universidade Federal de São Carlos (GETerm/UFSCar), convidado a avaliar o OntoLP. O uso do corpus nos experimentos deve-se ao conhecimento da equipe sobre o domínio, possibilitando que os avaliadores identifiquem termos do domínio e avaliem as estruturas hierárquicas geradas pelo plug-in.

O *NanoTerm* é constituído por documentos extraídos de diversas fontes, conforme descrito abaixo:

- Informativos: constituídos por jornais, revistas, portais, entre outros. Para esse gênero foram totalizadas 361.307 palavras.
- Científico: formada por textos extraídos de revistas científicas, do Banco de Teses Capes, doadas por CD-ROM, entre outros. Nesse caso, o total de palavras foi de 1.846.763.
- Científico de Divulgação: composto por documentos extraídos de sites especializados, revistas, Fundação de Desenvolvimento da Pesquisa (FUNDEP), entre outros. Aqui foram reunidos um conjunto de documentos constituídos por 310.018 palavras.
- Técnico Administrativo: textos retirados do portal do Ministério da Ciência e da Tecnologia[3]. Dessas fontes foram selecionados um total de 26.877 palavras.
- Outros: formado por textos presentes em *slides* de apresentações convertidas para ".txt". Total de 20.525 palavras.

[3]http://www.mct.gov.br

Portanto, o corpus possui um total geral de 2.565.490 palavras, distribuídas em 1057 textos, extraídos de 57 fontes diferentes.

5.1.3 Corpus de Pediatria (*JPED*)

Assim como o *NanoTerm*, o *JPED* foi utilizado durante a avaliação do plug-in feita pelos usuários. A escolha pelo corpus foi feita pelos mesmos motivos do *NanoTerm*, contar com a experiência do grupo que realizou a avaliação. Nesse caso, foi convidado o Grupo TERMISUL da Universidade Federal do Rio Grande do Sul (TERMISUL/UFRGS), que vem realizando um trabalho de extração dos termos presentes no corpus para a elaboração de um glossário para o domínio.

O corpus foi construído em um trabalho de mestrado (COULTHARD, 2005). Segundo o autor, o *JPED* é constituído por 283 textos em português extraídos do Jornal de Pediatria, totalizando 785.448 palavras.

5.2 Métricas de Avaliação

Uma questão importante para este trabalho é a avaliação dos métodos desenvolvidos através de métricas utilizadas na área. Consideramos duas formas adotadas em outros trabalhos:

- (BUITELAAR; OLEJNIK; SINTEK, 2003; BASÉGIO, 2006) avaliam seus métodos de construção com o auxílio de um especialista na área de domínio. Nessa abordagem, cada uma das etapas de construção de ontologias é avaliada por ele. Por exemplo, para os termos, o especialista avalia a lista extraída automaticamente, indicando quais devem ser considerados conceitos do domínio. Para a etapa de taxonomia, o especialista avalia as relações inferidas pelos métodos, apontando as corretas. Através dessa metodologia é possível realizar o cálculo de Precisão para as etapas

avaliadas.

- (ZAVAGLIA et al., 2007) utiliza as medidas de Precisão, Abrangência e F-measure para a avaliação de seus métodos de extração de termos, enquanto (CIMIANO; HOTHO; STAAB, 2005; RYU; CHOI, 2006) as utilizam para avaliar seus métodos de organização hierárquica. A aplicação dessas métricas depende da obtenção de uma lista de termos e uma taxonomia de referência, geradas manualmente a partir de um corpus e validadas por um especialista.

A primeira forma de avaliação foi aplicada neste trabalho nos experimentos realizados pelos usuários. A segunda foi utilizada na avaliação dos métodos de extração e organização hierárquica, com base no *CórpusEco* e sua ontologia, conforme detalhados a seguir.

5.2.1 Métricas de Avaliação dos Termos

As métricas de avaliação são adaptadas da área de RI. As equações utilizadas nos cálculos não diferem das originais, apenas altera-se documentos por termos. O cálculo das medidas parte de uma lista referência, comparando-a à lista extraída automaticamente. As métricas aplicadas são apresentadas abaixo:

- Precisão (P): indica a capacidade do método de identificar os termos corretos, considerando a lista de referência. É dada pela equação 5.1, onde o número de termos corretos extraídos é dividido pelo número total de termos extraídos.

$$P(t) = \frac{\left|TermosRelevantes \bigcap TermosExtraídos\right|}{TermosExtraídos} \tag{5.1}$$

- Abrangência (A): avalia a quantidade de termos corretos extraídos pelo método. É calculado através da equação 5.2, onde o número de termos extraídos corretamente é dividido pelo total de termos relevantes.

$$A(t) = \frac{\left|TermosRelevantes \bigcap TermosExtraídos\right|}{TermosRelevantes} \quad (5.2)$$

- F-measure (F): é considerada uma média harmônica entre a Precisão e Abrangência (equação 5.3).

$$F(t) = \frac{2 * (P * A)}{P + A} \quad (5.3)$$

Para a realização do cálculo das medidas, foi desenvolvido e integrado ao plug-in um módulo de avaliação automática, que recebe como entrada:

- As listas de termos geradas por cada método de extração disponível no OntoLP, ranqueadas pelas medidas estatísticas configuradas.
- A lista de referência.

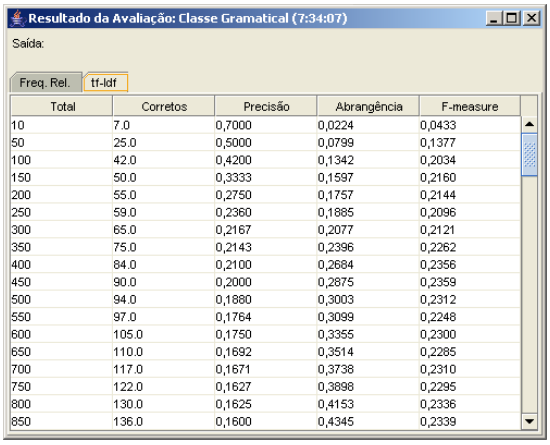
No módulo, as medidas são calculadas considerando os termos ranqueados como mais relevantes em diferentes faixas. O resultado final é apresentado ao usuário através de uma tabela, como na Figura 30. No exemplo, foi avaliado o método Classe Gramatical ordenado pelas medidas FR e *tf-idf*. A coluna "Total" indica qual faixa de termos está sendo considerada, enquanto a coluna "Corretos" indica o número de termos nessa faixa presentes na lista de referência[4].

Cabe salientar que esse módulo não está na versão do OntoLP disponível para os usuários.

5.2.2 Métricas de Avaliação das Taxonomias

Assim como para os termos, os métodos de organização hierárquica também são avaliados através das métricas de Precisão, Abrangência e F-measure. Nesse caso, a

[4]Lista de referência é uma lista de termos extraídos a partir de um corpus e validada por um especialista da área, podendo ser considerada como um gabarito na avaliação de sistemas de extração automática de termos



Total	Corretos	Precisão	Abrangência	F-measure
10	7.0	0,7000	0,0224	0,0433
50	25.0	0,5000	0,0799	0,1377
100	42.0	0,4200	0,1342	0,2034
150	50.0	0,3333	0,1597	0,2160
200	55.0	0,2750	0,1757	0,2144
250	59.0	0,2360	0,1885	0,2096
300	65.0	0,2167	0,2077	0,2121
350	75.0	0,2143	0,2396	0,2262
400	84.0	0,2100	0,2684	0,2356
450	90.0	0,2000	0,2875	0,2359
500	94.0	0,1880	0,3003	0,2312
550	97.0	0,1764	0,3099	0,2248
600	105.0	0,1750	0,3355	0,2300
650	110.0	0,1692	0,3514	0,2285
700	117.0	0,1671	0,3738	0,2310
750	122.0	0,1627	0,3898	0,2295
800	130.0	0,1625	0,4153	0,2336
850	136.0	0,1600	0,4345	0,2339

Figura 30: Interface do módulo de avaliação dos termos.

adaptação de tais medidas envolve um conjunto maior de conceitos e foi feita por (CIMIANO; HOTHO; STAAB, 2005).

Quando aplicados a estruturas hierárquicas os princípios básicos das medidas se mantêm, ou seja, o cálculo parte de uma taxonomia de referência construída com o auxílio de um especialista. Entretanto, elas utilizam uma medida de similaridade como base, avaliando a semelhança entre a ontologia construída automaticamente e a ontologia de referência.

As medidas baseiam-se nos conceitos: (1) *Common Semantic Cotopy* (SC'') e (2) *Taxonomy Overlap* ($\overline{TO}$). O primeiro é definido como o conjunto de todos os sub e super-conceitos de uma determinada classe. Entretanto, para a comparação entre ontologias, foram definidas algumas restrições em sua aplicação:

- O SC'' de um conceito é constituído apenas por sub-classes e super-classes presentes em ambas ontologias, de referência e automática.
- A própria classe não é incluída em seu SC''.
- SC'' de conceitos folhas são desconsiderados da avaliação das taxonomias.

Considerando as taxonomias O_{auto} e O_{ref} da Figura 31, ambas representam o mesmo domínio e possuem a mesma estrutura hierárquica. Contudo, alguns conceitos

são representados por termos diferentes, o que não interfere na SC''. Por exemplo, para os conceitos "*Rentable*" (O_{auto}) e "*Object To Rent*" (O_{ref}) teríamos:

- $SC''\,(Rentable, O_{auto}, O_{ref}) = \{Apartment, Car, Bike\}$;
- $SC''\,(ObjectToRent, O_{ref}, O_{auto}) = \{Apartment, Car, Bike\}$.

Conforme o exemplo, o SC'' para ambos os conceitos são o mesmo. Portanto, apesar de lexicalmente diferentes, suas relações com outras classes indicam que representam o mesmo conceito.

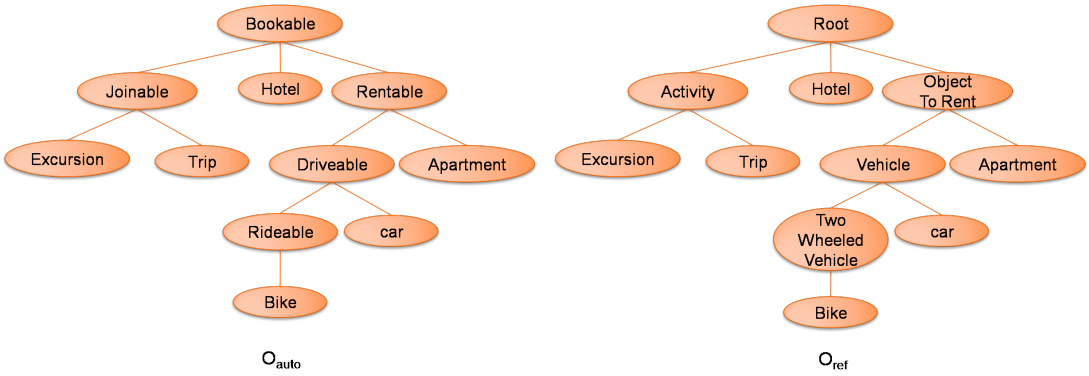


Figura 31: Ontologias de exemplo para SC'' (CIMIANO; HOTHO; STAAB, 2005).

A Taxonomy Overlap é uma medida de similaridade que se baseia na comparação dos SC'' entre as ontologias confrontadas. Segundo Cimiano, a $\overline{TO}(O_1, O_2)$ é calculada através da média considerando todas as sobreposições taxonômicas dos conceitos presentes na ontologia O_1.

$$\overline{TO}(O_1, O_2) = \frac{1}{|C_1C_2|} \sum_{c \in C_1C_2} max_{c' \in C_2 \bigcup root} \frac{|SC''(c, O_1, O_2) \bigcap SC''(c', O_2, O_1)|}{|SC''(c, O_1, O_2) \bigcup SC''(c', O_2, O_1)|} \tag{5.4}$$

Onde,

- $\overline{TO}(O_1, O_2)$ é a própria medida,
- O_1 e O_2 são as ontologias comparadas,

- C_1 e C_2 representam o conjunto de todos os conceitos presentes em cada ontologia,
- 'c' é o conceito atual, do qual está sendo extraído o SC'' ($c \in C_1$),
- 'c'' é o conceito que maximiza a sobreposição entre seu SC'' e o do conceito 'c' ($c' \in C_2$), ou seja, o conceito mais similar ao 'c' considerando seus SC''.

Finalmente, o cálculo das medidas de Precisão (P), Abrangência (A) e F-measure (F) são definidos pelas equações 5.5, 5.6 e 5.7:

$$P(O_1, O_2) = \overline{TO'}(O_1, O_2) \tag{5.5}$$

$$A(O_1, O_2) = \overline{TO'}(O_2, O_1) \tag{5.6}$$

$$F(O_1, O_2) = \frac{2 * (P(O_1, O_2) * A(O_1, O_2))}{P(O_1, O_2) + A(O_1, O_2)} \tag{5.7}$$

Para tornar mais clara a aplicação das medidas, Cimiano apresenta alguns exemplos, sendo três deles expostos abaixo:

- Para as ontologias apresentadas na Figura 31, temos $P(O_1, O_2) = R(O_1, O_2) = F(O_1, O_2) = 100\%$. Isto por que suas estruturas hierárquicas são exatamente iguais.
- Na Figura 32 são apresentadas as mesmas estruturas, exceto pelo conceito "*Rideable*", excluído de O_{auto}. O valor da Precisão mantém-se em 100%, pois todas as relações inferidas em O_{auto} estão presentes na O_{ref}. Entretanto, ocorre perda na abrangência, isto por que a relação correspondente ao conceito "*Two Wheeled Vehicle*" não foi identificada em O_{auto}. No cálculo exemplificado abaixo, os valores somados são os "SC''", enquanto entre parênteses, estão as classes das quais eles foram extraídos. O mesmo serve para os próximos exemplos.

– $A(O_{auto}, O_{ref}) = 1(Activity, Joinable) + 1(ObjectToRent, Rentable) + 1(Vehicle, Driveable) + 0.5(TwoWheeledVehicle, Driveable)/4 = 87,5\%$.

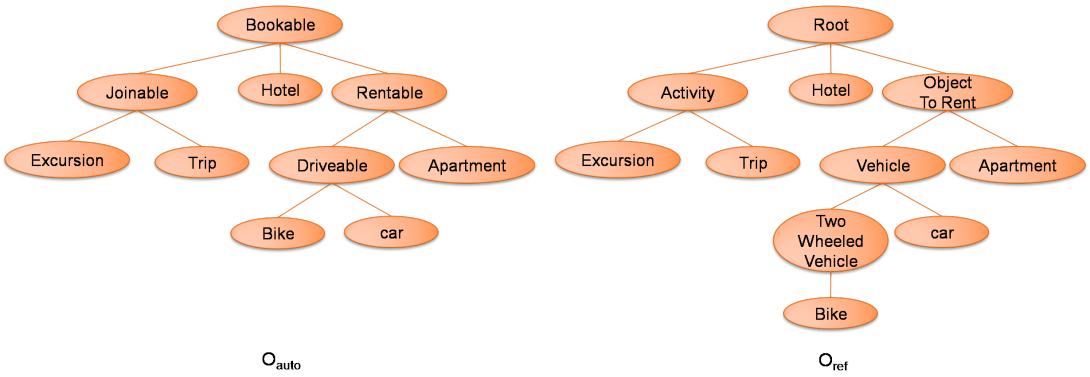


Figura 32: Exemplo de estrutura taxonômica com perda na Abrangência (CIMIANO; HOTHO; STAAB, 2005).

- Na Figura 33 é mantido a O_{ref} e adicionado o conceito "*Planable*" a estrutura original de O_{auto}. Nesse caso, a Abrangência volta a ser 100%, pois todas as relações presentes na O_{ref} estão presentes na O_{auto}. Já na Precisão ocorre perda, pois não existe nenhuma relação em O_{ref} com o mesmo SC'' de "*Planable*". O cálculo é descrito abaixo:

  – $P(O_{auto}, O_{ref}) = 1(Joinable, Activity) + 1(Rentable, ObjectToRent) + 1(Driveable, Vehicle) + 1(Rideable, TwoWheeledVehicle) + 0.5(Planable, Activity)/5 = 90\%$.

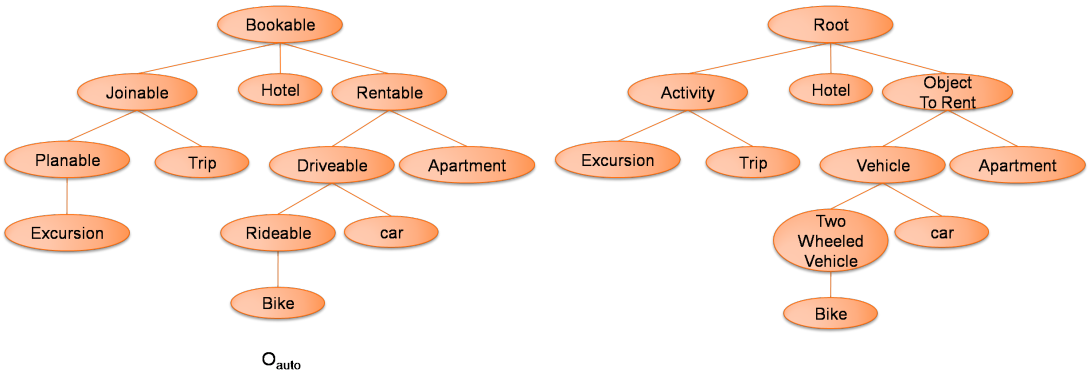


Figura 33: Exemplo de estrutura taxonômica com perda na Precisão (CIMIANO; HOTHO; STAAB, 2005).

A F-measure permanece como medida harmônica entre a Precisão e Abrangência, conforme a equação 5.7.

Além das medidas explicadas acima, (CIMIANO; HOTHO; STAAB, 2005) definiu ainda as métricas Abrangência Lexical e F-measure Lexical, como demonstradas a seguir:

- Abrangência Lexical (AL): dada pela divisão entre o número de conceitos corretos na taxonomia extraída pelo número de conceitos presentes na taxonomia de referência (equação 5.8).

$$AL(O_1, O_2) = \frac{|C_1 \bigcap C_2|}{|C_2|} \tag{5.8}$$

  Onde,

  - O_n são as ontologias consideradas no cálculo,
  - C_n são os conceitos que compõem cada ontologia.

- F-measure Lexical (F'): é a mesma medida de F-measure, entretanto, calculada com base na Abrangência Lexical (equação 5.9).

$$F' = \frac{2 * (P * AL)}{P + AL} \tag{5.9}$$

Assim como na extração de termos, para o avaliação das estruturas hierárquicas inferidas foi desenvolvido um módulo que executa a tarefa automaticamente. O módulo recebe como entrada:

- A taxonomia gerada pelos métodos de organização hierárquica;
- A taxonomia de referência;

Como resultado são retornados os métodos avaliados e os valores obtidos por cada um deles, como apresentado na Figura 34.

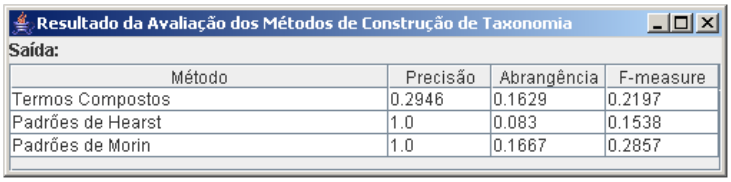


Figura 34: Interface do módulo automático de avaliação dos métodos de construção de Taxonomias.

Cabe salientar que as taxonomias de entrada para o módulo devem estar representadas por uma estrutura de dados do Protégé. Vale ressaltar ainda que as métricas AL e F' apresentadas, apesar de serem utilizadas durante a avaliação, foram calculadas manualmente.

5.3 Avaliação da Etapa de Extração dos Termos

Nesta seção são descritos os experimentos de avaliação da etapa de extração de termos. A tarefa foi dividida em duas partes, na primeira a avaliação foi feita sobre os pares método/medida, considerando todas as combinações possíveis entre as técnicas disponíveis no plug-in. Na segunda, foram selecionados os melhores pares na extração de uni, bi e trigramas, baseado nos resultados do experimento anterior. Esses métodos foram comparados aos utilizados no trabalho (ZAVAGLIA et al., 2007) (*baseline*).

Na primeira etapa as métricas utilizadas para a avaliação dos pares métodos/medidas foram Precisão (P), Abrangência (A) e F-measure (F). Para a apresentação dos resultados foi utilizado um corte que seleciona apenas os mil primeiros termos. Esse limiar foi escolhido, pois os valores de F-measure nessa faixa são próximos aos obtidos quando calculada a média geral para essa mesma medida. Os resultados desse experimento são demonstrados nas Tabelas 18, 19 e 20, respectivamente, uni, bi e trigramas. Nelas, 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.

Tabela 18: Resultados para a extração de unigramas considerando uma faixa de 1000 termos.

Método (total/corretos)	**Medida**	**Corretos**	**P**	**A**	**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%**		

Para a extração de unigramas houve um empate entre quatro pares de método/medida: 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 35 apresenta a F-measure obtida por cada um em diferentes faixas de termos. Conforme é possível perceber, o par Classe Gramatical/*tf-idf* foi o que obteve os melhores resultados, exceto para as faixas de 300 e 350 termos, onde se destacou o método Classe Gramatical/*NC-Value(tf-idf)*.

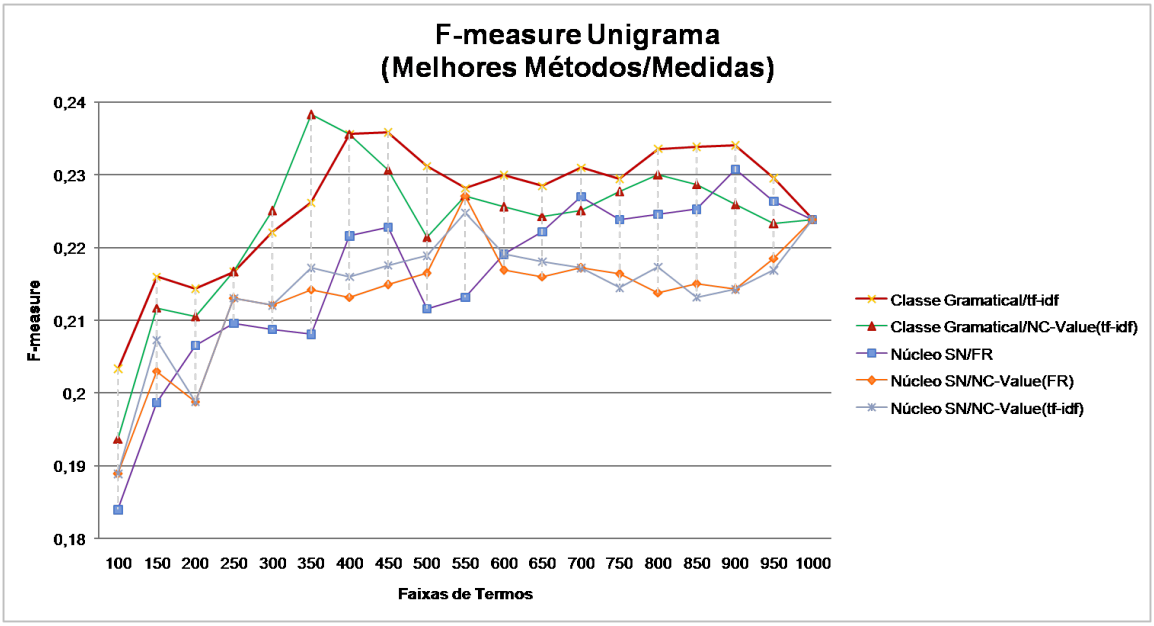


Figura 35: Resultados obtidos pelos métodos para diferentes faixas de termos durante a extração de unigramas (escala para F-measure entre 0,18 e 0,24).

Com relação aos bigramas, os melhores pares foram: Padrões Morfossintáticos/FR

Tabela 19: Resultados para a extração de Bigramas considerando uma faixa de 1000 termos.

Método (total/corretos)	**Medida**	**Corretos**	**P**	**A**	**F**	**Corte**	**Ref.**
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%*		

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 isto não ocorre, a medida se comporta como o cálculo da FR. Nesses experimentos, esse fenômeno acontece somente para os termos complexos bigramas, já que nos trigramas aparecem termos como "tamanho_de_o_população" (número de palavras=4) e "taxa_de_crescimento" (número de palavras=3). 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 36 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 Fi-

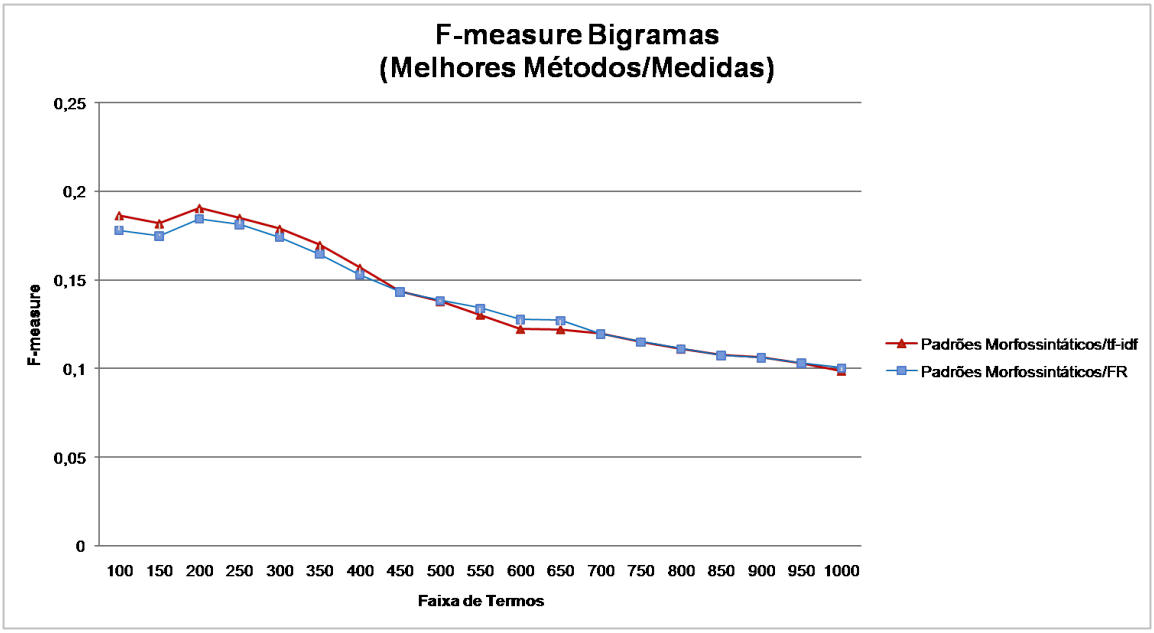


Figura 36: Resultados obtidos pelos métodos para diferentes faixas de termos durante a extração de bigramas (escala para F-measure entre 0 e 0,25).

Tabela 20: Resultados para a extração de Trigramas considerando uma faixa de 1000 termos.

Método (total/corretos)	**Medida**	**Corretos**	**P**	**A**	**F**	**Corte**	**Ref.**
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%		

gura 37 apresenta a comparação entre eles. O resultado obtido por cada método/medida para as diferentes faixas de termos confirmou o melhor desempenho do Padrões Morfossintáticos/*tf-idf*, sendo inferior somente para os 150 e 200 termos.

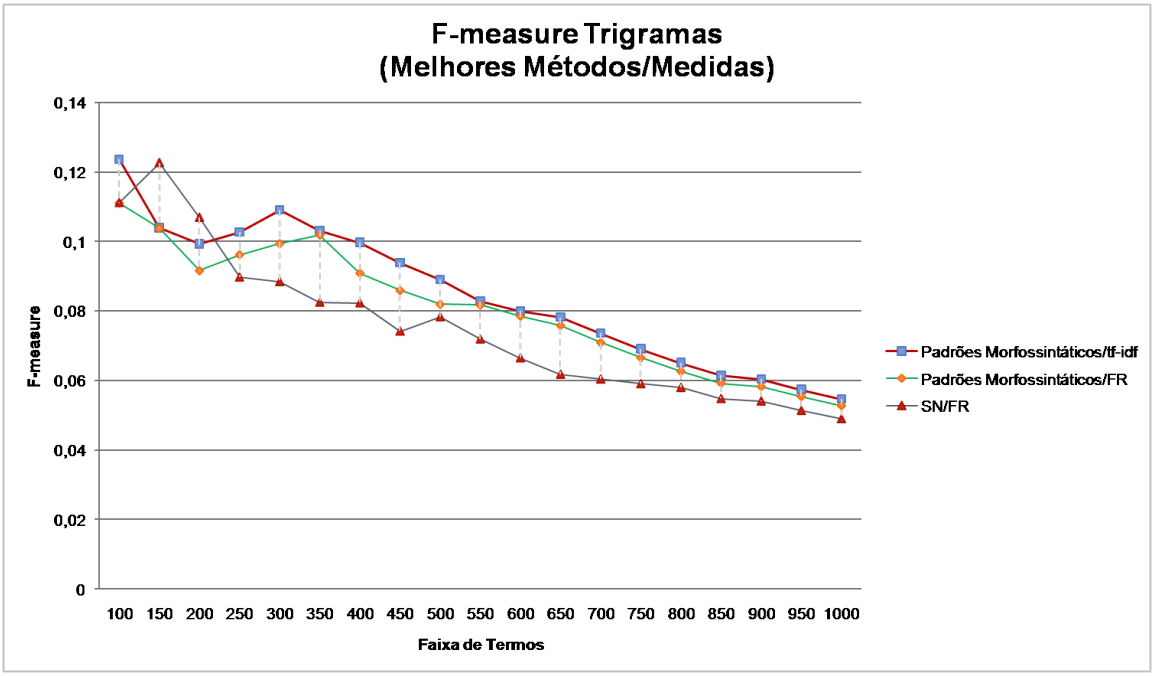


Figura 37: Resultados obtidos pelos métodos para diferentes faixas de termos durante a extração de trigramas (escala para F-measure entre 0 e 0,14).

Com base nos resultados descritos acima, foram selecionados os pares: Classe Gramatical/*tf-idf* (unigramas), Padrões Morfossintáticos/*tf-idf* (bi e trigramas) para serem comparados aos métodos utilizados no *baseline*. Nos experimentos realizados no *baseline* foram utilizadas as três técnicas de extração: estatística, lingüística e híbrida. No total foram avaliados 15 métodos diferentes, considerando apenas os 150 primeiros termos extraídos por cada método. Na Tabela 21 são apresentados os melhores resultados obtidos pelo *baseline* e pelo OntoLP. Cabe salientar que, para os bigramas e trigramas, nos quais o *baseline* obteve seus melhores resultados utilizando abordagem estatística, os valores alcançados com a aplicação da abordagem híbrida são demonstrados entre parênteses.

5.3.1 Considerações

Conforme os resultados obtidos na primeira seqüência de experimentos, as combinações Classe Gramatical/*tf-idf* (unigramas) e Padrões Morfossintáticos/*tf-idf* (bigramas e trigramas) foram as que obtiveram melhores resultados na avaliação geral. Para o caso dos unigramas, apesar do método Núcleo do Sintagma Nominal extrair uma quan-

Tabela 21: Comparação entre os resultados obtidos pelo *baseline* e o OntoLP.

Unigrama			
Método	**Termos Corretos Extraídos**	**Precisão**	**Corte**
Classe Gramatical/*tf-idf*	50	33%	150
baseline (Abordagem Híbrida)	45	30%	
Bigrama			
Método	**Termos Corretos Extraídos**	**Precisão**	**Corte**
Padrões Morfossintáticos/*tf-idf*	26	17%	150
baseline (Abordagem Estatística)	51 (19)	34%	
Trigrama			
Método	**Termos Corretos Extraídos**	**Precisão**	**Corte**
Sintagma Nominal/FR	13	8%	150
baseline (Abordagem Estatística)	10 (3)	6%	

tidade menor de termos (Núcleo do SN=4527 e Classe Gramatical=5742), isso não influenciou nos resultados de precisão. Além disso, quando considerado o total geral de termos corretos, o método continua em desvantagem (Núcleo do SN=258 e Classe Gramatical=272). Para os termos complexos os resultados obtidos foram bem inferiores aos dos unigramas, o que era esperado. Nesta tarefa, os métodos SN e N-Grama foram equivalentes em termos de resultados. Contudo, nenhum deles obteve valores tão bons quanto os Padrões Morfossintáticos, que apresentou os melhores resultados para ambas as listas (bi e trigramas).

Quanto às medidas estatísticas utilizadas, apesar do esforço de desenvolvimento e avaliação da *C-Value* e *NC-Value*, os melhores resultados foram obtidos pelas mais conhecidas, FR e *tf-idf*. De uma forma geral, a medida *tf-idf* foi superior as demais, incluindo a FR. Para todos os pares em que apareceu obteve resultados, quando não superiores, semelhantes aos das outras métricas. Além disso, cabe salientar que o custo computacional dessa medida é baixo, ao qual somente a FR se equivale.

Por fim, quanto a comparação entre o *baseline* e OntoLP, cabe salientar que cada trabalho utiliza entradas anotadas por analisadores diferentes. Essa questão influencia diretamente nos resultados, visto que os erros gerados pelos parsers são repassados para os métodos de extração. Tendo feitas essas considerações, conforme os resultados, o plug-

in foi superior em todos os casos, exceto quando avaliados os bigramas. Uma provável justificativa é o uso de medidas que levam em consideração a probabilidade de duas ou mais palavras ocorrerem juntas em um corpus, como a Informação Mútua. Contudo, para esses termos, quando comparadas somente as abordagens híbridas de ambos os trabalhos, o OntoLP obteve resultados superiores.

Vale lembrar que nessa avaliação não foi utilizada a Seleção de Grupos Semânticos, por ser um processo que depende obrigatoriamente da intervenção do usuário. Portanto, uma avaliação completa sobre a extração de termos com a aplicação dessa etapa é apresentada na seção 5.5.

5.4 Avaliação da Etapa de Organização Hierárquica dos Termos

Como apresentado na seção 2.4, existe uma relação de dependência entre as etapas de construção de ontologias. Normalmente, a lista de termos extraída é utilizada como entrada para os métodos de organização hierárquica. Sendo assim, se avaliarmos esses métodos utilizando esse processo, os resultados seriam influenciados pelos erros provenientes da etapa anterior. Para amenizar esse problema, neste trabalho foi utilizada a metodologia proposta por (RYU; CHOI, 2006). Nela as relações hierárquicas entre os conceitos presentes na taxonomia de referência são excluídas, restando uma lista de termos. Essa lista é utilizada como entrada para os métodos de construção de ontologias, que tentam reconstruir a estrutura original. O resultado final do processo é comparado com a taxonomia de referência.

Para a avaliação foi utilizada como referência a ontologia de "Ecologia de Comunidades", extraída a partir do *CórpusEco*, com uma lista de 59 termos. As medidas de avaliação utilizadas foram a Precisão (P), Abrangência (A) e F-measure (F), conforme o processo descrito na seção 5.2.2. Os resultados obtidos por cada método são descritos na

Tabela 22.

Tabela 22: Resultados da construção automática de Taxonomias.

Método	**P**	**A**	**F**
Termos Complexos	29,46%	16,29%	21,97%
Padrões de Hearst	100%	8,3%	15,38%
Padrões de Morin	100%	16,67%	28,57%

Analisando os resultados obtidos e as taxonomias geradas (Anexo C), é possível perceber que, mesmo com poucos conceitos e relações hierárquicas, o método baseado nos padrões de Morin/Jacquemin alcançou os melhores resultados. Contudo, numa análise qualitativa percebe-se que a estrutura extraída pelo método baseado em Termos Complexos auxiliaria mais o usuário durante a construção da ontologia (como será demonstrado na seção 5.5, Tabela 26). Partindo desse princípio, as taxonomias foram avaliadas também segundo as métricas AL e F'. Na Tabela 23 são apresentados os resultados obtidos por cada método com base nessas medidas.

Tabela 23: Resultados da construção automática de Taxonomias, considerando a AL e F'.

Método	**AL**	**F'**
Termos Complexos	100%	45,51%
Padrões de Hearst	3,4%	6,6%
Padrões de Morin	5,1%	9,7%

5.4.1 Consideracões

As medidas utilizadas na primeira avaliação punem taxonomias que tenham muitos conceitos e poucas relações comparadas à de referência. Sendo assim, apesar do método baseado em Termos Complexos apresentar todos os conceitos presentes na ontologia, ele possui baixa precisão e abrangência, já que poucos estão organizados hierarquicamente.

As medidas AL e F' parecem representar melhor o resultado obtido por cada método. Através dessas métricas, a ontologia baseada em Termos Complexos foi melhor

avaliada do que as demais, obtendo 100% para *AL*. Isto se deve a utilização da lista de conceitos extraída da própria ontologia de referência como entrada para o método. Esse processo é semelhante ao proposto neste trabalho, através do qual o especialista edita a lista de termos extraída na etapa anterior, restando apenas conceitos válidos. A aplicação dessas medidas em conjunto com as métricas P, A e F parece tornar os resultados mais fáceis de serem compreendidos.

No Anexo D são apresentados os resultados obtidos por (CIMIANO; HOTHO; STAAB, 2005), avaliando diferentes métodos de clusterização durante a realização da tarefa. O objetivo não é compará-los aos resultados deste trabalho, pois existem diferenças nos domínios, língua e analisador sintático, que influenciam os resultados. A idéia é apresentar um parâmetro para o que está sendo alcançado na área.

5.5 Avaliação do OntoLP feita por Usuários

Como mencionado anteriormente, o objetivo do OntoLP como ferramenta é ajudar o usuário durante as primeiras etapas de construção de ontologias. Nesse sentido, algumas informações foram integradas à interface do plug-in, como por exemplo, *tag clouds*, palavras de contexto, entre outras (seções 4.2.1 e 4.2.2). Essas informações não influenciam nos resultados dos métodos, somente auxiliam o engenheiro durante as etapas que necessitam de interação com o sistema.

Outra questão importante é a estratégia utilizada na Seleção de Grupos Semânticos (Capítulo 3), que necessita da intervenção do usuário durante a execução. Isto torna os resultados obtidos dependentes do nível de conhecimento do usuário sobre o domínio.

Nenhuma das funcionalidades discutidas acima pode ser avaliada automaticamente, como nas seções 5.3 e 5.4. Conseqüentemente, optou-se por uma avaliação qualitativa da ferramenta feita por usuários experientes nas tarefas de extração de termos e construção de taxonomias. Esse experimento teve como objetivos:

- Avaliar a interface do plug-in em termos de quais funcionalidades auxiliam mais o usuário durante a construção da taxonomia. A idéia é, com base nos resultados, caminhar em direção a uma interface mais simples, porém, com informações mais relevantes para o processo.

- Avaliar a extração de termos através da métrica de Precisão (P), utilizando a etapa de Seleção de Grupos Semânticos.

Para cumprir o primeiro objetivo foi necessário que usuários experientes realizassem a tarefa. Nesse caso, foi dada preferência por pesquisadores com experiência em construção de ontologias e que já utilizaram outras ferramentas de auxílio ao processo. Para o segundo objetivo, os mesmos usuários receberam listas de termos compostas 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* (seção 5.2.1), cujo corpus foi utilizado nos experimentos.

- 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* (seção 5.2.2), utilizado nos experimentos.

5.5.1 Metodologia

Inicialmente, cada avaliador recebeu:

- O manual de funcionamento do OntoLP (Anexo B);
- O corpus de entrada para o plug-in anotado pelo PALAVRAS e representado em XCES. Cada avaliador recebeu o corpus de acordo com o grupo de pesquisa ao qual está vinculado;
- O executável do plug-in OntoLP para ser instalado no Protégé;
- O questionário de avaliação (Anexo E).
- Um arquivo "*readme.txt*" indicando os passos a serem seguidos;

Os primeiros passos para a realização da avaliação foram a leitura do manual e a instalação do plug-in em conjunto com o sistema Protégé. Finalizadas essas tarefas, a avaliação ocorreu em duas etapas:

1. Preenchimento e execução das tarefas definidas no questionário, subdividido em três partes:
    - Avaliação da experiência do usuário: aqui os avaliadores responderam a uma série de perguntas, indicando qual sua experiência sobre as etapas de construção de ontologias;
    - Extração de Termos: nesse caso foram definidas uma série de tarefas de extração de termos. Ao final os avaliadores responderam a questões indicando o quanto cada informação presente na interface do plug-in o auxiliou durante a tarefa. Além disso, nessa etapa, cada avaliador gerou uma lista de Grupos Semânticos, utilizando o método Filtro por Grupos Semânticos (seção 3.2.1). Essas listas foram usadas na segunda parte da avaliação;

- Organização Hierárquica dos Termos: assim como a etapa anterior, os avaliadores foram instruídos a realizar algumas tarefas. Com base na interação com o sistema responderam a questões relacionadas às funcionalidades disponibilizadas pelo plug-in para a construção de taxonomias.

2. Na etapa de extração de termos, cada avaliador selecionou um conjunto de Grupos Semânticos. Para cada conjunto foram geradas três listas de termos, constituídas por uni, bi e trigramas. Os avaliadores receberam as três listas de acordo com os Grupos Semânticos que haviam selecionado. 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.

É importante destacar que os corpora utilizados durante as tarefas do questionário foram uma pequena parte dos apresentados nas seções 5.1.2 e 5.1.3, aproximadamente 256.549 palavras para o *NanoTerm* e 44.356 para o *JPED*. Os objetivos foram agilizar a realização dos experimentos, diminuindo o tempo de execução dos métodos, e minimizar a probabilidade de problemas por limitação de hardware, como memória insuficiente. Para a geração das listas de termos a partir dos Grupos Semânticos, porém, foram utilizados os corpora completos, já que a tarefa foi realizada pelo autor deste trabalho.

5.5.2 Resultados da Avaliação da Experiência dos Usuários

A primeira parte da avaliação visou definir a experiência dos avaliadores. Segundo as respostas, todos os consultados já haviam realizado as etapas de extração e organização hierárquica dos termos, e apenas o Avaliador1 indicou que já havia feito as demais tarefas. Com relação aos critérios que já utilizaram para a extração de termos, todos haviam usado alguma ferramenta para auxiliar na tarefa, e somente o Avaliador3 nunca realizou o processo através de critério semântico (manualmente). Os sistemas citados são apresentados na Tabela 24.

Tabela 24: Sistemas de EAT citados pelos avaliadores.

Gerador de n-gramas disponível em (http://www2.lael.pucsp.br/ tony/tony/Home.html)
Corpógrafo da Linguateca (http://www.linguateca.pt/Corpografo/)
(TELINE; MANFRIN; ALUíSIO, 2003)

Para a organização hierárquica, todos os avaliadores indicaram já ter realizado o processo, entretanto, nunca utilizando um sistema de auxílio. Finalmente, para a questão autocrítica sobre a experiência na construção de ontologias, o Avaliador 3 se considerou inexperiente, enquanto os demais se consideraram razoavelmente experientes.

5.5.3 Resultados da Avaliação das Funcionalidades

Para a avaliação das funcionalidades, foram definidas as seguintes opções: 0-Não auxiliou e aumentou o trabalho; 1-Não auxiliou; 2-Auxiliou e 3-Auxiliou muito. Na Tabela 25 são apresentados os resultados obtidos para a etapa de Extração de Termos.

Conforme demonstrado, no primeiro item (Auxílio a Seleção dos Grupos Semânticos), todos os avaliadores indicaram a "Definição dos Grupos Semânticos" como o pior critério durante a seleção dos relevantes. Entre as duas funcionalidades restante, as *tag clouds* obtiveram o melhor resultado geral. Para o segundo e terceiro item (Auxílio a Extração de Termos Simples e Termos Complexos), o método Filtro por Grupos Semânticos se destaca, sendo indicado pelos três avaliadores como a principal informação de auxílio as tarefas, seguida da Organização por Relevância e do Termo Original. O último item questionou qual a abordagem que mais agradou aos avaliadores, com ou sem a etapa de Seleção de Grupos Semânticos, sendo unânime a escolha pela primeira.

A segunda etapa das questões respondidas pelos avaliadores é apresentada na Tabela 26 e diz respeito a Organização Hierárquica dos Termos. Com base nos resultados é possível perceber que o melhor método de auxílio segundo os avaliadores foi, unanimemente, a abordagem baseada em Termos Complexos. Isto provavelmente se deve ao fato do método gerar uma taxonomia com todos os termos selecionados na etapa anterior,

Tabela 25: Resultados da avaliação feita pelos usuários para as funcionalidades da Extração de Termos.

Extração de Termos				
Auxílio à Seleção dos Grupos Semânticos				
Item Avaliado	**Avaliador 1**	**Avaliador 2**	**Avaliador 3**	**Média**
Organização por relevância	3	2	3	2,67
Tag clouds	3	3	3	3
A definição do Grupo	2	1	1	1,33
Auxílio à Extração de Termos Simples				
Organização por relevância	3	1	3	2,33
Tags Semânticas	1	1	1	1
Termo Original	2	3	2	2,33
Palavras de Contexto	1	3	1	1,67
Filtro por Grupos Semânticos (etapa anterior)	3	3	3	3
Auxílio à Extração de Termos Complexos				
Organização por relevância	3	2	3	2,67
Tags Semânticas	1	1	1	1
Termo Original	2	3	2	2,33
Palavras de Contexto	1	3	1	1,67
Filtro por Grupos Semânticos (etapa anterior)	3	3	3	3
Restrição por Unigramas	3	3	2	2,67
Melhor Abordagem para a tarefa				
Abordagem 1 (Seleção de Grupos Semânticos)	X	X	X	
Abordagem 2				

mesmo que com alguns erros estruturais. Os padrões de Hearst e Morin/Jacquemin, como são pouco freqüentes, extraem pouquíssimas relações, o que deve ter influenciado nos baixos resultados obtidos. A última funcionalidade (Visualização da Taxonomia Extraída ao lado da Taxonomia em Construção) foi dita auxiliar a tarefa por todos avaliadores.

5.5.4 Resultados da Avaliação da Extração de Termos

Como destacado anteriormente, para esse experimento os avaliadores realizaram duas etapas:

1. Seleção dos Grupos Semânticos relevantes para o domínio que estavam utilizando.

2. Seleção dos Termos Simples e Complexos (uni, bi e trigrama) com base nas listas geradas a partir dos Grupos Semânticos escolhidos.

Tabela 26: Resultados da avaliação feita pelos usuários para as funcionalidades da Organização Hierárquica.

Organização hierárquica dos Termos				
Item Avaliado	**Avaliador 1**	**Avaliador 2**	**Avaliador 3**	**Média**
Método baseado em Termos Complexos	3	3	3	3
Método baseado nos Padrões de Hearst	1	1	1	1
Método baseado nos Padrões de Morin/Jacquemin	1	1	1	1
Visualização da Taxonomia Inferida ao lado da Taxonomia em Construção	3	2	2	2,33

Nesta seção são apresentados os resultados obtidos em termos de Precisão para as duas partes. No primeiro caso, a lista de Grupos Semânticos selecionada por cada avaliador foi considerada referência, sendo comparada a lista ranqueada pelo método Filtro por Grupos Semânticos. No segundo, as listas de termos extraídos com base nos grupos semânticos selecionados foram editadas pelos avaliadores, servindo como referência para os termos identificados pelos métodos de extração.

Os resultados para a primeira etapa são apresentados na Tabela 27. Nela a Precisão é dada considerando a lista completa de grupos semânticos extraídos e ranqueados (P) e considerando uma faixa dos 30 primeiros grupos (P'). Nas colunas "GS Extraídos" e "GS Selecionados" são apresentados, respectivamente, o total de grupos extraídos pelo método e a quantidade selecionada pelos avaliadores. Nessa avaliação o ranqueamento de Filtro por Grupos Semânticos apresentou resultados satisfatórios para a realização da tarefa.

Tabela 27: Resultado da avaliação da precisão obtida pelos Filtros Semânticos.

NanoTerm				
Avaliador	**GS Extraídos**	**GS Selecionados**	**P**	**P'**
Avaliador1	131	52	40%	73%
Avaliador2		78	60%	80%
JPED				
Avaliador3	137	39	28,5%	66%

Para a segunda etapa foram considerados apenas os pares de método e medida que obtiveram melhores resultados nos experimentos da seção 5.3: Classe Gramatical/*tf-idf* (unigrama) e Padrões Morfossintáticos/*tf-idf* (bigrama e trigrama). Os resultados são

apresentados nas Tabelas 28, 29 e 30, respectivamente, unigramas, bigramas e trigramas. Nelas são demonstrados: o método acompanhado da medida (Método/Medida); a quantidade de termos presentes na lista de referência de acordo com o avaliador (Ref.); 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 era de se esperar, no domínio de Nanotecnologia & Nanociência os resultados foram inferiores aos obtidos no domínio de Pediatria, por ser muito específico. Essa característica influenciou 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 os resultados foram de 6,66% para Pediatria e 0,165% (N&N).

Tabela 28: Resultado da extração de unigramas com e sem a Seleção de Grupos Semânticos.

	Unigramas						
	Método/Medida	**Avaliador**	**Ref.**	**Corretos**	**P**	**GS**	**Corte**
NanoTerm	Classe Gramatical/ *tf-idf*	Av1	52	26	17,33%	S	150
				19	12,67%	N	
		Av2	75	59	39,33%	S	
				55	36,67%	N	
JPED		Av3	223	108	72%	S	
				82	54,67%	N	

Tabela 29: Resultado da extração de bigramas com e sem a Seleção de Grupos Semânticos.

	Bigramas						
	Método/Medida	**Avaliador**	**Ref.**	**Corretos**	**P**	**GS**	**Corte**
NanoTerm	Padrães Morfossintáticos/ *tf-idf*	Av1	31	19	12,67%	S	150
				17	11,33%	N	
		Av2	112	57	38%	S	
				51	34%	N	
JPED	Padrões Morfossintáticos/ *tf-idf*	Av3	163	139	92,67%	S	
				108	72%	N	

Tabela 30: Resultado da extração de trigramas com e sem a Seleção de Grupos Semânticos.

	Trigramas						
	Método/Medida	**Avaliador**	**Ref.**	**Corretos**	**P**	**GS**	**Corte**
NanoTerm	Padrões Morfossintáticos/ *tf-idf*	Av1	41	17	11,33%	S	150
				17	11,33%	N	
		Av2	131	45	30%	S	
				43	29,67%	N	
JPED	Padrões Morfossintáticos/ *tf-idf*	Av3	166	83	55,33%	S	
				73	48,67%	N	

5.5.5 Considerações

Nesta seção foram apresentados os resultados da avaliação feita com os usuários. Esses experimentos são relevantes, pois possibilitam analisar itens subjetivos, como a informatividade de alguns componentes disponibilizados na interface do plug-in e a aplicação das informações semânticas durante a extração de termos.

O experimento de avaliação das funcionalidades do plug-in foi dividido em extração de termos e organização hierárquica. Para a primeira etapa os avaliadores indicaram de forma unânime o método Filtro por Grupos Semânticos como o que mais auxilio a extração de termos simples e complexos. Quanto ao "Auxílio à Seleção dos Grupos Semânticos", as "*Tag clouds*" e a "Organização por Relevância" foram os melhores avaliados, confirmando a falta de informatividade das "Definições dos Grupos Semânticos". Finalmente, na definição da melhor abordagem para a extração de termos, os avaliadores escolheram de forma unânime a que utiliza as informações semânticas.

Para a segunda etapa, relacionada à taxonomia, os avaliadores reportaram que a aplicação dos padrões de Hearst e Morin/Jacquemin não auxiliaram na organização hierárquica de termos. Um provável motivo para o resultado é que, se o usuário executar tais métodos sem levar em consideração as listas de termos extraídas anteriormente, o método irá retornar relações entre termos considerados irrelevantes para o domínio. No entanto, quando são aplicadas heurísticas que consideram apenas termos presentes na lista, poucas relações são extraídas. Por outro lado, o método de construção hierárquica baseado em Termos Complexos obteve bons resultados.

Finalmente, quando avaliada em termos de precisão, à aplicação de Filtros Semânticos para a extração de termos foi melhor tanto para o domínio de N&N, quanto para Pediatria. Os resultados demonstraram que no domínio de N&N os métodos apresentaram uma precisão menor em relação a Pediatria, o que era esperado devido a especificidade do domínio. No entanto, houve acréscimo nos resultados para ambos os domínios, principalmente para os unigramas.

Capítulo 6

Conclusão

Este trabalho objetivou o estudo e desenvolvimento de métodos para a construção de ontologias a partir de textos da língua portuguesa. Nele foi proposta uma metodologia com base em informações lingüísticas, utilizando os níveis morfológicos, sintáticos e semânticos. A metodologia criada foi implementada em um plug-in para o ambiente Protégé, que será disponibilizado para a comunidade científica. As etapas de construção de ontologias executadas pelo plug-in são a extração de termos e sua organização hierárquica.

O trabalho foi bastante abrangente com relação às questões da área de construção de ontologias a partir de textos. Nele foram apresentados os conceitos relacionados à tarefa e às principais técnicas empregadas. Além disso, foram propostos novos métodos e adaptações a métodos já existentes. No trabalho foram consideradas ainda, questões de interface e funcionalidades para auxiliar o usuário durante a execução das tarefas. Cabe ressaltar também a avaliação da metodologia, que considerou tanto a avaliação com base em uma ontologia de referência como a opinião de especialistas.

A primeira etapa de experimentos demonstrou que os métodos Classe Gramatical e Padrões Morfossintáticos apresentaram os melhores resultados para o corpus estudado. Os métodos Núcleo do SN e SN obtiveram resultados razoáveis, mas sendo inferiores aos outros dois métodos. Quanto às medidas de relevância, os experimentos indicaram que a

tf-idf tem melhor resultado geral.

Na segunda parte dos experimentos o trabalho avaliou as funcionalidades e informações disponibilizadas na interface do plug-in e a metodologia completa, incluindo o uso de informações semânticas. Para a execução dos testes foram utilizados dois corpora de áreas distintas, Nanotecnologia & Nanociência e Pediatria. Nesse experimentos, as expectativas quanto as funcionalidades foram confirmadas. Segundo os avaliadores, para a etapa de extração de termos o conceito de *tag clouds* foi o que mais auxiliou o processo de seleção dos grupos semânticos, seguido da ordenação por relevância. Na mesma etapa, a aplicação da técnica de Filtro por Grupos Semânticos foi indicada como a que mais ajudou na extração de termos simples e complexos. Por fim, para todos os avaliadores o uso de informações semânticas durante o processo de extração melhorou a execução da tarefa. Quanto à interface de organização hierárquica, a visualização da estrutura que está sendo construída ao lado das geradas automaticamente foi bem avaliada.

É interessante destacar que os experimentos considerando a metodologia completa confirmaram a opinião dos avaliadores. Quanto à comparação entre o uso ou não da etapa de informações semânticas para auxiliar a extração de termos, em praticamente todos os casos foram obtidas melhoras consideráveis nos resultados. Já em relação a organização hierárquica dos termos, o método com melhor resultado geral foi o baseado em Termos Complexos, confirmando a opinião dos avaliadores, que indicaram o método como o único que auxiliou no processo.

Para os experimentos que utilizaram as medidas de Precisão, Abrangência e F-measure, foram desenvolvidos dois módulos que automatizam o processo. Esses módulos, além de agilizar a tarefa de avaliação, são importantes para o desenvolvimento de novos trabalhos na área. Atualmente é inviável a realização de pesquisas sem a comprovação de resultados por meio de métricas de consenso entre os pesquisadores.

O trabalho enfrentou algumas dificuldades, a começar pela escassez de recursos para avaliação dos métodos. Conforme mencionado, para a aplicação das métricas de

Precisão, Abrangência e F-measure são necessários corpora e ontologias de referência. A construção desses recursos é trabalhosa, para que uma lista de termos e/ou uma taxonomia possam ser utilizadas como referência elas devem ser criadas por especialistas (manualmente). Além disso, o processo deve ser minucioso, esgotando os termos presentes no corpus de domínio. Caso a extração de termos não seja feita dessa forma, termos corretos extraídos por métodos automáticos serão avaliados como incorretos, prejudicando os resultados obtidos.

Ainda relacionado à construção desses recursos, outro problema é a necessidade de validação da lista de termos feita por um especialista no domínio considerado. A construção de ontologia é um processo subjetivo, por exemplo, se dois ou mais especialistas realizarem a tarefa, existe uma grande probabilidade de discordarem em alguns pontos. Portanto, o ideal é que sejam consultados mais de um especialista, para que haja consenso sobre a estrutura de representação do domínio. Pensando em termos de construção automática, se tivermos diferentes ontologias para uma mesma área, um método provavelmente terá resultados diferentes quando comparada a cada uma delas. Nesses casos, o consenso é algo essencial, pois garante a comparação com o que seria a ontologia "consensual" extraída do corpus.

A dificuldade de obtenção desses recursos limita a realização de experimentos com diferentes domínios. Dessa forma, apesar dos resultados obtidos nas avaliações indicarem que alguns métodos se sobressaíram em relação a outros, isso pode estar relacionado a alguma particularidade da área de conhecimento. Portanto, para que sejam feitas afirmações mais concretas quanto ao desempenho dos métodos é importante uma avaliação robusta utilizando recursos de diferentes domínios.

É importante destacar ainda a dificuldade de comparação entre métodos desenvolvidos em diferentes trabalhos. Por exemplo, como citado na seção 5.3, na comparação entre o *baseline* e o plug-in OntoLP deve ser levado em consideração o nível de acertos dos analisadores sintáticos utilizados na anotação das informações lingüísticas, pois eles

influenciam os resultados obtidos pelos métodos.

Para tratar esses problemas foi submetido e aprovado um projeto de pesquisa para o EDITAL CT-INFO 2007 - "Grandes Desafios da Computação", coordenado pela Prof. Renata Vieira e contando com o autor deste trabalho como colaborador. O projeto foi intitulado "Pesquisa em desenvolvimento de ontologias para a língua portuguesa - OntoLP". Um dos objetivos principais desse projeto é desenvolver um portal de informações sobre ontologias em língua portuguesa, centralizando nele os recursos, métodos e ferramentas.

A idéia é impulsionar as pesquisas na área de ontologias para a língua portuguesa. Através do portal, pesquisadores poderão consultar o estado da arte na tarefa concernente a língua portuguesa, interagir com outros pesquisadores, fazer o download de corpora anotados e com recursos completos de avaliação, incluindo benchmarks com resultados obtidos por outros sistemas, possibilitando comparações.

Finalmente, cabe salientar que o autor desta dissertação realizou um intercâmbio de 3 meses na Universidade de São Paulo-São Carlos (ICMC/USP). O intercâmbio foi feito através do projeto "FORTALECIMENTO E INTEGRAÇÃO NAS COMPETÊNCIAS DO PROCESSAMENTO DA LÍNGUA" (FAROL PROCAD-CAPES).

6.1 Publicações

Este trabalho foi publicado no III Workshop de Teses e Dissertações em Inteligência Artificial (WTDIA 2006), realizado em Ribeirão Preto, com o título"Geração de ontologias para a web semântica a partir de textos da Língua Portuguesa". Além disso, o autor desta dissertação, em conjunto com o grupo de pesquisa do Laboratório de Engenharia da Linguagem, aprovou um artigo sobre o estudo de informações semânticas aplicadas a resolução de anáforas, intitulado "Uso de Informações Semânticas na Identificação de Anáforas Indiretas e Associativas", publicado no 5º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007).

6.2 Trabalhos Futuros

A partir desta dissertação, é possível identificar novas pesquisas para possíveis trabalhos futuros:

- Estudar métodos que realizem as demais etapas de construção de ontologias e integrá-los ao plug-in desenvolvido, como por exemplo, a extração de definições conceituais automaticamente.

- Estudar o uso de buscas na Web como forma de aprimorar os métodos, melhorando os resultados de extração de termos e a abrangência dos padrões de Hearst e Morin/Jacquemin.

- Avaliar a utilização dos padrões propostos por (FREITAS, 2007) para identificação de Hiperônimos e Hipônimos.

- Estudar a viabilidade de construção de taxonomias utilizando os grupos semânticos como super-conceitos. Por exemplo, o grupo "Profissão Humana" teria como filhos "Pesquisador", "Astrônomo", "Físico", entre outros.

- Possibilitar que o plug-in receba entradas disponibilizadas por analisadores sintáticos gratuitos.

- Criar interfaces distintas, considerando usuários leigos e avançados na tarefa de construção de ontologias.

Referências

AGIRRE, E. et al. *Enriching WordNet concepts with topic signatures*. 2001. Disponível em: <citeseer.ist.psu.edu/agirre01enriching.html>.

AUSSENAC-GILLES, N.; BIÉBOW, B.; SZULMAN, S. Corpus analysis for conceptual modelling. In: *Proc. of the EKAW'2000 workshop 'Ontologies and texts' , Juan-Les-Pins, 02/10/00-02/10/00*. Toulouse (F): Université Paul Sabatier, 2000. p. 13–20. Disponível em: <http://www.irit.fr/wsontologies2000>.

BASÉGIO, T. *Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, 2006.

BASILI, R.; PAZIENZA, M. T.; VELARDI, P. An empirical symbolic approach to natural language processing. *Artif. Intell.*, Elsevier Science Publishers Ltd., Essex, UK, v. 85, n. 1-2, p. 59–99, 1996. ISSN 0004-3702.

BERNERS-LEE, T.; HENDLER, J. A.; LASSILA, O. The semantic web. v. 284, n. 5, p. 34–43, maio 2001.

BICK, E. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese (Doutorado) — Arhus University, 2000.

BICK, E. Noun sense tagging: Semantic prototype annotation of a portuguese treebank. In: HAJIc, J.; NIVRE, J. (Ed.). *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006)*. Prague, Czech Republic: [s.n.], 2006.

BRANTS, S. et al. The tiger treebank. In: *Workshop on Treebanks and Linguistic Theories*. Sozopol: [s.n.], 2002. p. 24–42.

BREWSTER, C.; CIRAVEGNA, F.; WILKS, Y. Background and foreground knowledge in dynamic ontology construction. In: SIGIR. *Proceedings of the Semantic Web Workshop, Toronto, August 2003*. [S.l.], 2003.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: P-BUITELAAR; CIMIANO, P.; MAGNINI, B. (Ed.). *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005, (Frontiers in Artificial Intelligence and Applications, v. 123). Disponível em: <http://www.aifb.uni-karlsruhe.de/WBS/pci/OL-Book-Intro.pdf>.

BUITELAAR, P.; OLEJNIK, D.; SINTEK, M. Ontolt: A protégé plug-in for ontology extraction from text. In: *Proceedings of the Demo Session of the International Semantic Web Conference (ISWC)*. Sanibel Island, Florida,: [s.n.], 2003. Disponível em: <iswc03-demo.pdf>.

BUITELAAR, P.; OLEJNIK, D.; SINTEK(DIPL.-INFORMATIKER KM, m. a. M. A protégé plug-in for ontology extraction from text based on linguistic analysis. In: *Proceedings of the 1st European Semantic Web Symposium*. [s.n.], 2004. Disponível em: <http://dfki.de/ paulb/esws04.pdf>.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 14, n. 1, p. 20–26, 1999. ISSN 1541-1672.

CHEN, H. A textual database knowledge-base coupling approach to creating computer supported organizational memory. In: . MIS Departament, University of Arizona: [s.n.], 1994.

CIMIANO, P.; HOTHO, A.; STAAB, S. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR)*, AAAI Press, v. 24, p. 305–339, 2005. ISSN 1076-9757. Disponível em: <http://www.jair.org/abstracts/cimiano05a.html>.

CIMIANO, P.; VöLKER, J.; STUDER, R. Ontologies on demand? - a description of the state-of-the-art, applications, challenges and trends for ontology learning from text. *Information, Wissenschaft und Praxis*, v. 57, n. 6-7, p. 315–320, OCT 2006. See the special issue for more contributions related to the Semantic Web. Disponível em: <http://www.aifb.uni-karlsruhe.de/WBS/pci/Publications/iwp06.pdf>.

COULTHARD, R. J. *The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus*. Dissertação (Mestrado) — Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, 2005.

DECLERCK, T. A set of tools for integrating linguistic and non-linguistic information. In: *Proceedings of SAAKM (ECAI Workshop)*. [s.n.], 2002. Disponível em: <schug$_s$*aakm.pdf*¿.

FALBO, R. de A.; MENEZES, C. S. de; ROCHA, A. R. A systematic approach for building ontologies. In: *IBERAMIA '98: Proceedings of the 6th Ibero-American Conference on AI*. London, UK: Springer-Verlag, 1998. p. 349–360. ISBN 3-540-64992-1.

FAURE, D.; NEDELLEC, C. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In: *EKAW '99: Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*. London, UK: Springer-Verlag, 1999. p. 329–334. ISBN 3-540-66044-5.

FENSEL, D. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. 2001. Disponível em: <citeseer.ist.psu.edu/413498.html>.

FERNANDEZ, M.; GOMEZ-PEREZ, A.; JURISTO, N. Methontology: from ontological art towards ontological engineering. In: *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*. Stanford, USA: [s.n.], 1997. p. 33–40.

FERNEDA, E. *Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. Tese (Doutorado) — Universidade de São Paulo, 2003.

FRANTZI, K. T.; ANANIADOU, S.; TSUJII, J. ichi. 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*. London, UK: Springer-Verlag, 1998. p. 585–604. ISBN 3-540-65101-2.

FREITAS, M. C. de. *Elaboração automática de ontologias de domínio: discussão e resultados*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, 2007.

GRAU, B. C. A possible simplification of the semantic web architecture. In: *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2004. p. 704–713. ISBN 1-58113-844-X.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowl. Acquis.*, Academic Press Ltd., London, UK, UK, v. 5, n. 2, p. 199–220, 1993. ISSN 1042-8143.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, Academic Press, Inc., Duluth, MN, USA, v. 43, n. 5-6, p. 907–928, 1995. ISSN 1071-5819.

GUARINO, N. Formal ontology and information systems. In: GUARINO, N. (Ed.). *Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998*. Amsterdam: IOS Press, 1998. p. 3–15.

GUARINO, N.; GIARETTA, P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, p. 25–32, 1995. Disponível em: <http://www.csee.umbc.edu/771/papers/KBKS95.pdf.Z>.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1992. p. 539–545. Disponível em: <http://portal.acm.org/citation.cfm?id=992154>.

HENDLER, J. Agents and the semantic web. *IEEE Intelligent Systems*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 2, p. 30–37, 2001. ISSN 1541-1672.

HOLSAPPLE, C. W.; JOSHI, K. D. A collaborative approach to ontology design. *Commun. ACM*, ACM Press, New York, NY, USA, v. 45, n. 2, p. 42–47, 2002. ISSN 0001-0782.

IDE, N.; BONHOMME, P.; ROMARY, L. Xces: An xml-based encoding standard for linguistic corpora. In: *Proceedings of the Second International Language Resources and Evaluation Conference*. [S.l.: s.n.], 2000.

INNISS, T. R. et al. Towards applying text mining and natural language processing for biomedical ontology acquisition. In: *TMBIO '06: Proceedings of the 1st international workshop on Text mining in bioinformatics*. New York, NY, USA: ACM Press, 2006. p. 7–14. ISBN 1-59593-526-6.

KISHORE, R.; ZHANG, H.; RAMESH, R. A helix-spindle model for ontological engineering. *Commun. ACM*, ACM Press, New York, NY, USA, v. 47, n. 2, p. 69–75, 2004. ISSN 0001-0782.

MAEDCHE, A.; STAAB, S. Semi-automatic Engineering of Ontologies from Text. In: *Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering*. [S.l.: s.n.], 2000.

MAEDCHE, A.; STAAB, S. Ontology learning for the semantic web. *IEEE Intelligent Systems*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 2, p. 72–79, 2001. ISSN 1541-1672.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999. Disponível em: <citeseer.ist.psu.edu/635422.html>.

MORIN, E.; JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, v. 38, n. 4, p. 363–396, 2004. Disponível em: <http://www.lina.sciences.univ-nantes.fr/Publications/2004/MJ04>.

PANTEL, P.; LIN, D. A statistical corpus-based term extractor. In: *AI '01: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*. London, UK: Springer-Verlag, 2001. p. 36–46. ISBN 3-540-42144-0.

PEREZ, G. A.; MANCHO, M. D. *A Survey of Ontology Learning Methods and Techniques*. May 2003. OntoWeb Delieverable 1.5.

PERINI, M. A gramática gerativa. In: _____. [S.l.]: Vigília, 1985.

RAYSON, P.; BERRIDGE, D.; FRANCIS, B. Extending the cochran rule for the comparison of word frequencies between corpora. In: *In Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*. [S.l.: s.n.], 2004.

ROSCH, E. Principles of categorization. In: ROSCH, E.; LLOYD, B. (Ed.). *Cognition and Categorization*. [S.l.]: Lawrence Erlbaum Associate, 1978. p. 27–48.

RYU, P.-M.; CHOI, K.-S. 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*. Sydney, Australia: Association for Computational Linguistics, 2006. p. 41–48. Disponível em: <http://www.aclweb.org/anthology/W/W06/W06-0506>.

SILVA, M. C. P. d. S.; KOCH, I. G. V. *Lingüística aplicada ao português: sintaxe*. São Paulo: Cortez, 2001. ISBN 85-249-0204-3. Disponível em: <citeseer.ist.psu.edu/635422.html>.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, v. 25, n. 1-2, p. 161–197, 1998. ISSN 0169-023X.

TELINE, M. F.; MANFRIN, A. M. P.; ALUíSIO, S. M. *Extração Automática de Termos e Textos em Português: Aplicação e Avaliação de Medidas Estatísticas de Associação de Palavras*. [S.l.], 10 2003.

USCHOLD, M.; KING, M. Towards a methodology for building ontologies. In: *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conduction with IJCAI-95*. Montreal, Canada: [s.n.], 1995. Disponível em: <http://www.aiai.ed.ac.uk/project/pub/documents/1995/95-ont-ijcai95-ont-method.ps>.

W3C. *W3C Semantic Web Activity*. 2001. Disponível em: http://www.w3.org/2001/12/semweb-fin/w3csw Acessado em Março de 2007.

WÄCHTER, T. et al. A corpus-driven approach for design, evolution and alignment of ontologies. In: *WSC '06: Proceedings of the 37th conference on Winter simulation*. [S.l.]: Winter Simulation Conference, 2006. p. 1595–1602. ISBN 1-4244-0501-7.

ZAVAGLIA, C. Relatório de pesquisa de Pós-Doutorado, *Elaboração de uma Base Léxico-Ontológica Computacional (Português) do Subdomínio da Ecologia Bloc-Eco*. 2004.

ZAVAGLIA, C. et al. 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*. Rio de Janeiro, Brasil: [s.n.], 2007. p. 1575–1584. Disponível em: <http://www.nilc.icmc.usp.br/til/til2007/index.htm>.

Anexo A - Resultados obtidos em experimentos preliminares aplicando restrições aos termos

Tabela 31: Resultados dos experimentos preliminares utilizando heurísticas para excluir termos que não representam conceitos. Essa avaliação foi feita para as restrições de tamanho e tipo da palavra em unigramas. Elas excluem formações do tipo "hs.", "1º", "%" etc.. Os resultados indicam quantos termos foram excluídos, e se entre eles havia algum relevante.

Método	**Restrição**	**Termos Extraídos**	**Termos Corretos**	**Ref.**	**Domínio**
Classe Gramatical	Não	3630	41	50	Turismo
	Sim	3568	41		
Núcleo do Sintagma Nominal	Não	2962	41		
	Sim	2919	41		
Classe Gramatical	Não	5926	272	313	Ecologia
	Sim	5750	272		
Núcleo do Sintagma Nominal	Não	4630	258		
	Sim	4526	258		

Tabela 32: Resultado do experimento com a mesma finalidade do anterior, avaliar o impacto das heurísticas preposição e artigo no método N-Grama. Essas restrições excluem termos que comecem ou terminem com preposições e artigos, por exemplo, "com_café", "apartamento_com" etc..

Método	**Restrição**	**Termos Extraído**	**Termos Corretos**	**Ref.**	**Domínio**
N-Grama	Não	7360	70	88	Turismo
	Sim	2514	70		
	Não	20835	100	136	Ecologia
	Sim	10045	100		

Tabela 33: Resultados dos experimentos com a restrição de tamanho e tipo da palavra na extração de termos complexos. Nesse caso, essas heurísticas restringem formações como "%_de_arroz", "1h_da_manhã", entre outras.

Método	**Restrição**	**Termos Extraídos**	**Termos Corretos**	**Corretos**	**Domínio**
N-Grama	Não	2514	70	88	Turismo
	Sim	2456	70		
Sintagma Nominal	Não	1882	41		
	Sim	1521	41		
Padrões Morfossintáticos	Não	1609	70		
	Sim	1593	70		
N-Grama	Não	1045	100	136	Ecologia
	Sim	9538	100		
Sintagma Nominal	Não	5304	75		
	Sim	4600	75		
Padrões Morfossintáticos	Não	6518	98		
	Sim	6443	98		

Anexo B - Manual do OntoLP

1- Introdução ao OntoLP

O sistema OntoLP é um plug-in para o ambiente de construção de ontologias Protégé. O plug-in visa auxiliar o engenheiro de ontologias durante a execução das etapas iniciais de construção de ontologias: extração dos termos candidatos a conceitos e organização hierárquica desses termos. O sistema utiliza métodos de construção de ontologias a partir de textos baseados em medidas estatísticas e informações lingüísticas. Portanto, para que seja executado, é necessário anotar os textos com o analisador sintático PALAVRAS e representá-los no formato XCES (padrão de representação de informações lingüísticas adotado no projeto PLNBR).

O plug-in é organizado em três etapas: (1) aba de carga do corpus; (2) aba de extração dos termos e (3) aba de organização hierárquica dos termos, explicadas nas seções 5, 6 e 7 respectivamente.

2- Instalação do OntoLP

Para instalar o OntoLP é preciso possuir o ambiente Protégé (versão 3.3.1) instalado. A partir disso, são necessários os seguintes passos:

1.Entrar na pasta "plugins" no diretório de instalação do Protégé.

2.Descompactar o arquivo OntoLP.zip dentro dessa pasta.

Na Figura 38 é demonstrado como deve ficar a pasta "plugins".

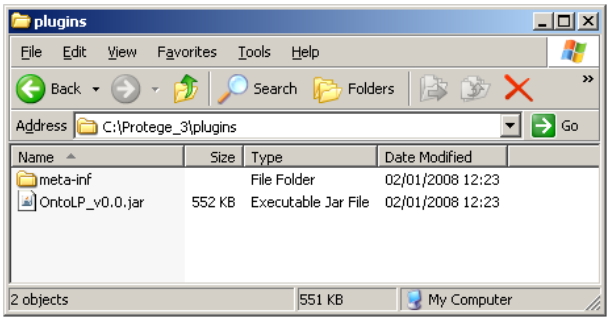


Figura 38: Arquivos de instalação do OntoLP.

3- Executando o OntoLP

Para a execução do OntoLP são necessários os seguintes passos:

1.Inicie o Protégé.

2.Na tela "*Welcome to Protégé*" (Figura 39) selecione "*new Project*" ou o projeto em uso.

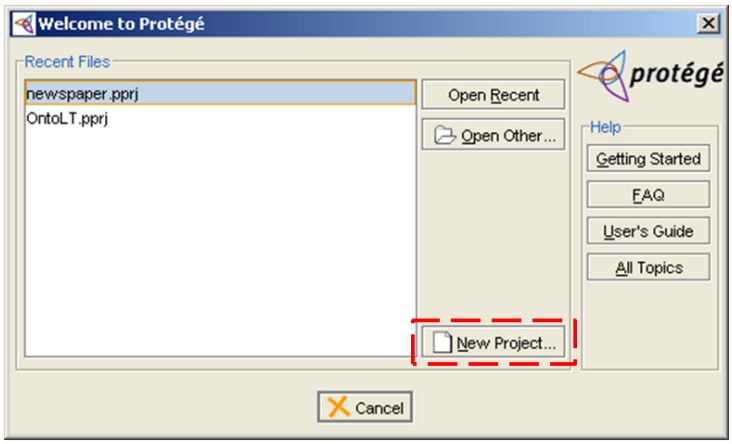


Figura 39: Interface de boas vindas do sistema Protégé.

3.O Protégé irá abrir o projeto escolhido. Feito isso, para carregar o OntoLP acesse na barra de menu as opções *Project*→*Configure...* (Figura 40).

4.A interface apresentada na Figura 41 aparecerá na tela, você deve selecionar "OntoLP" e pressionar o botão "ok".

5.Depois de executado o passo 4 aparecerá a aba do plug-in na interface do Protégé (Figura 42). Para iniciar a utilizá-lo clique sobre ela.

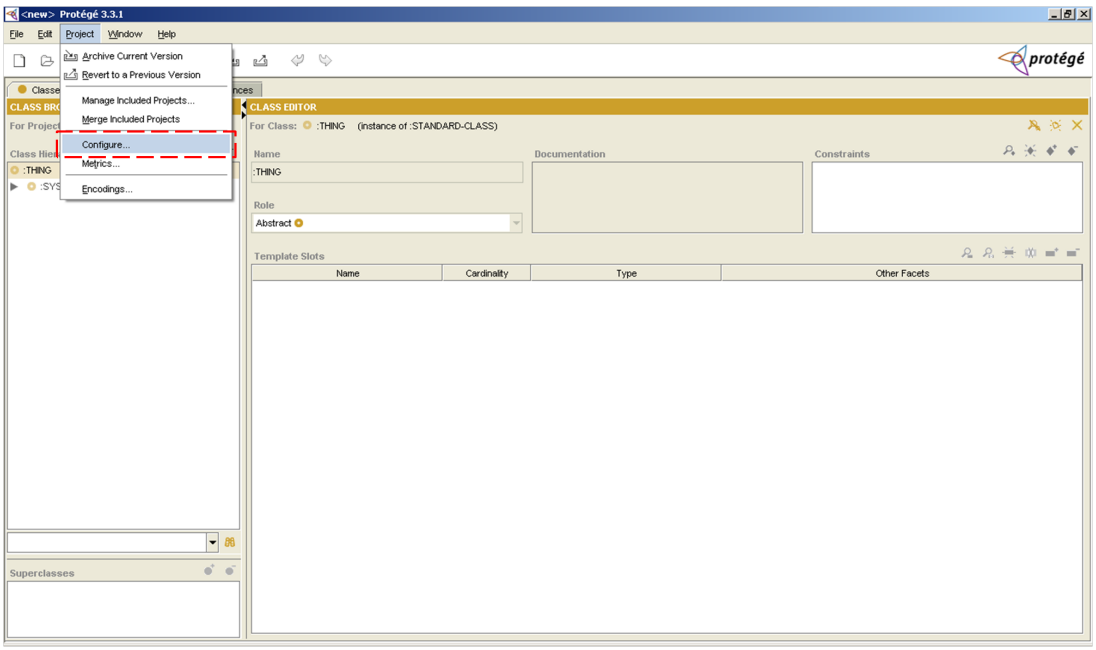


Figura 40: Menu de Configuração do Projeto.

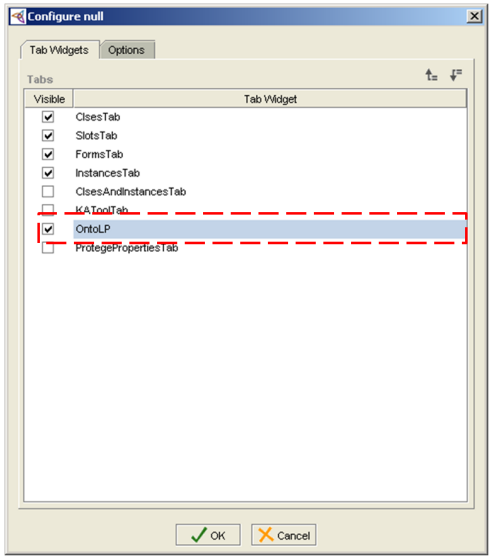


Figura 41: Interface de seleção de plug-in do Protégé.

4- Observação Importante

Todos os botões do OntoLP apresentam um texto indicando sua função sempre que o mouse é deixado sobre eles (Figura 43).

5- Aba de Importação do Corpus

Um guia rápido da Interface de Importação do Corpus é apresentado na Figura 44.

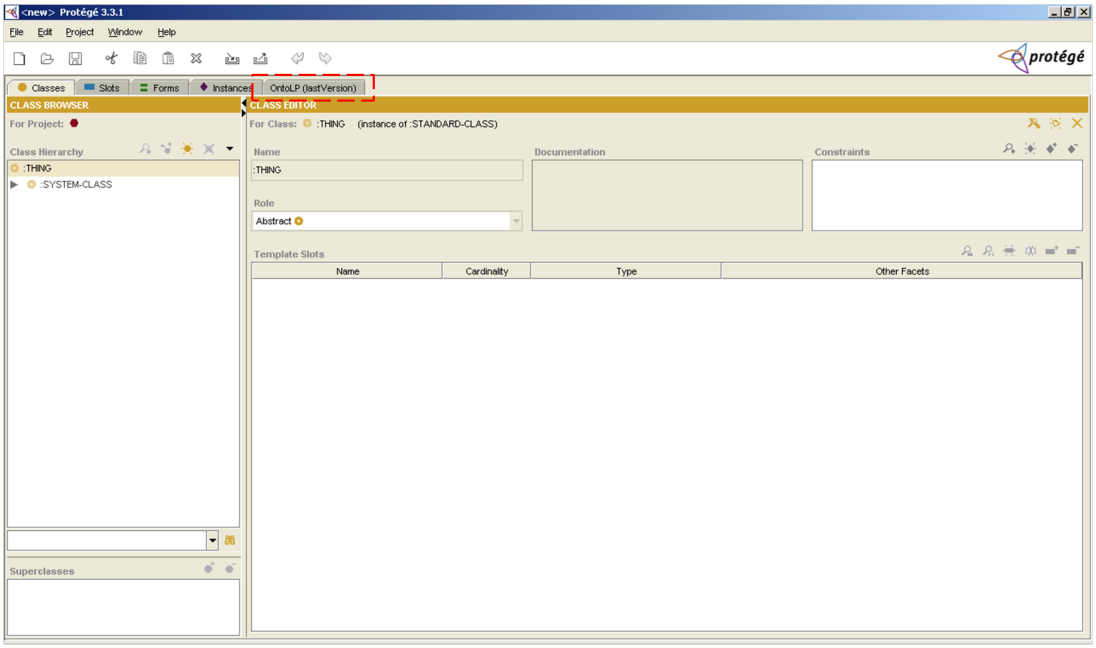


Figura 42: Interface do Protégé com a aba OntoLP.

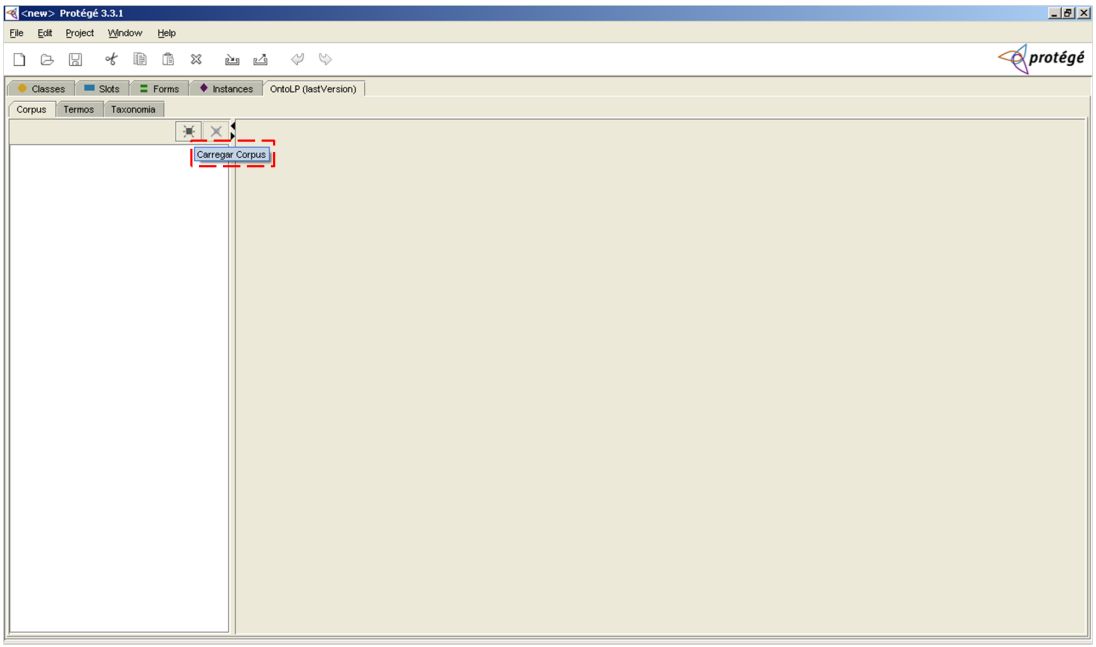


Figura 43: Exemplo de texto indicando a função de um botão.

5.2- Funcionalidades

Para carregar um corpus no OntoLP são necessárias as seguintes etapas:

1.Selecionar a aba “Corpus” (Figura 45).

2.Pressionar o botão “Carregar Corpus” (Figura 46).

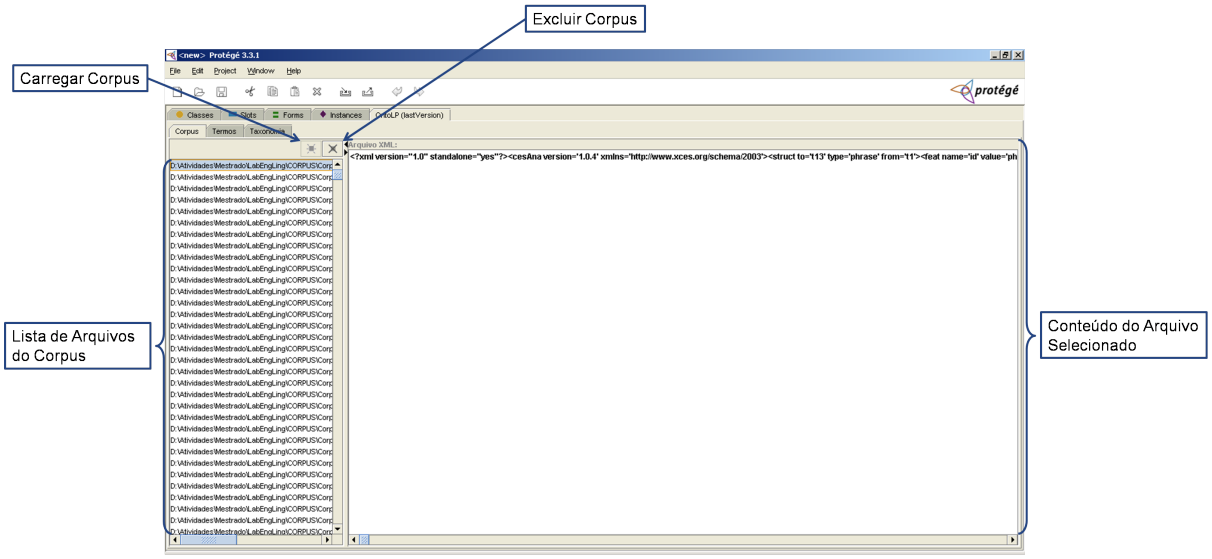


Figura 44: Interface de Importação do Corpus.

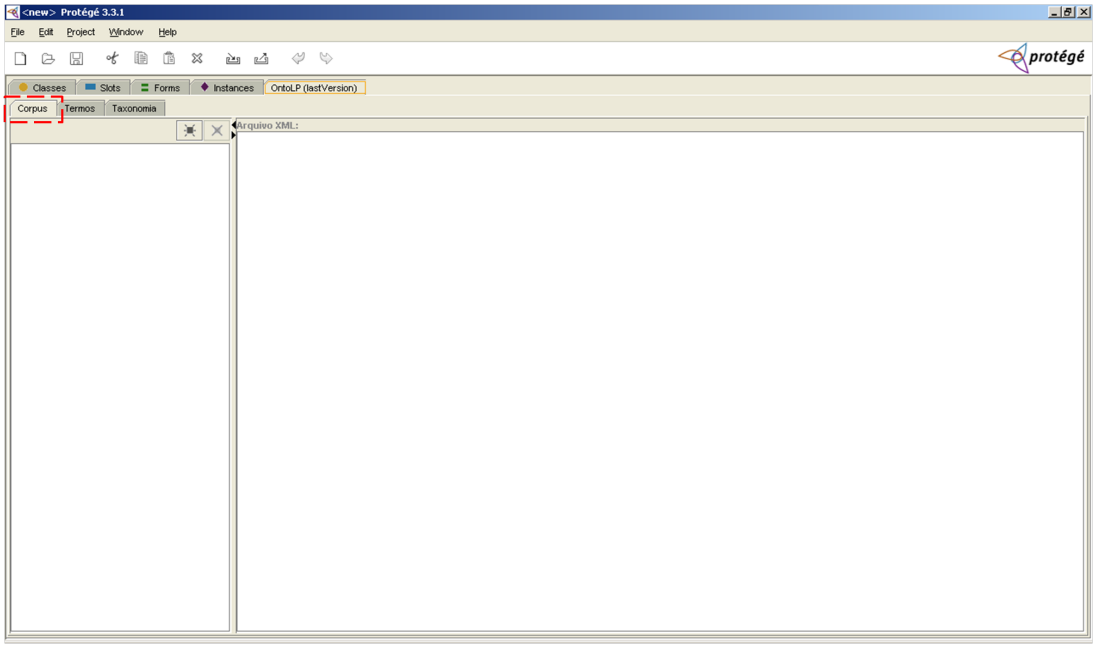


Figura 45: Interface de carga do corpus.

Figura 46: Botão para carregar o corpus de domínio.

3.Selecionar a lista de arquivos (ctrl+a) em formato XCES e pressionar “abrir” (Figura 47).

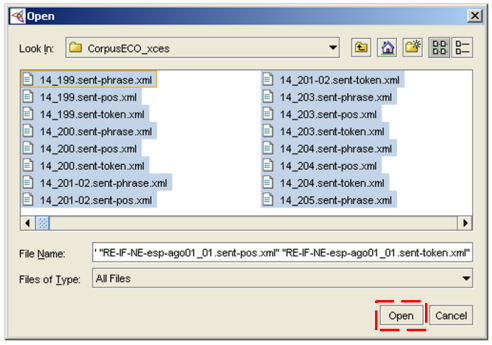


Figura 47: Interface de seleção do corpus.

4.Depois de carregado o corpus, é apresentada uma lista com os arquivos selecionados. Para visualizar o conteúdo de um arquivo basta selecioná-lo (Figura 48).

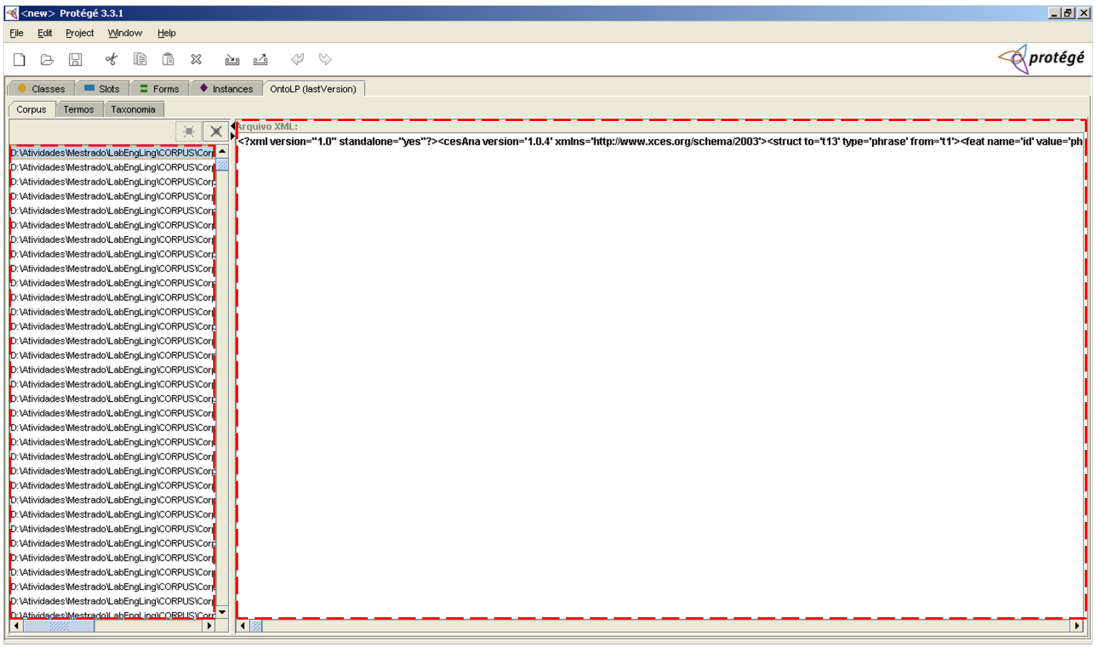


Figura 48: Corpus depois de carregado pelo usuário.

6- Aba de Extração de Termos

Nas Figuras 49 e 50 são apresentados guias rápidos para a Interface de Extração de Termos, que se subdivide em Filtro por Grupos Semânticos e Extração de Termos Simples e Complexos.

6.2- Funcionalidades

A extração de termos no OntoLP pode ser feita de duas maneiras: (Abordagem

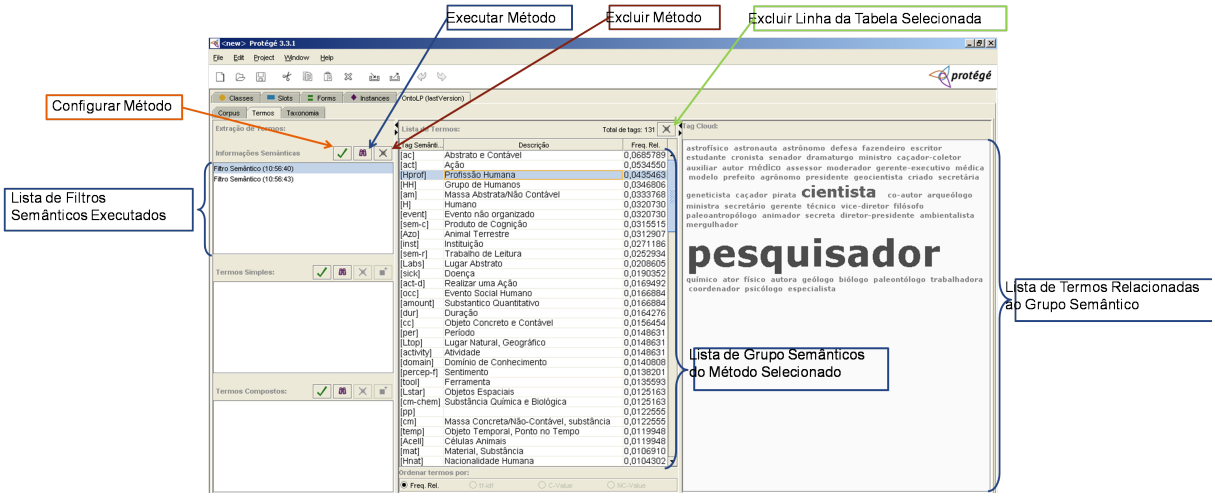


Figura 49: Interface de Extração de Termos (Filtro por Grupo Semântico).

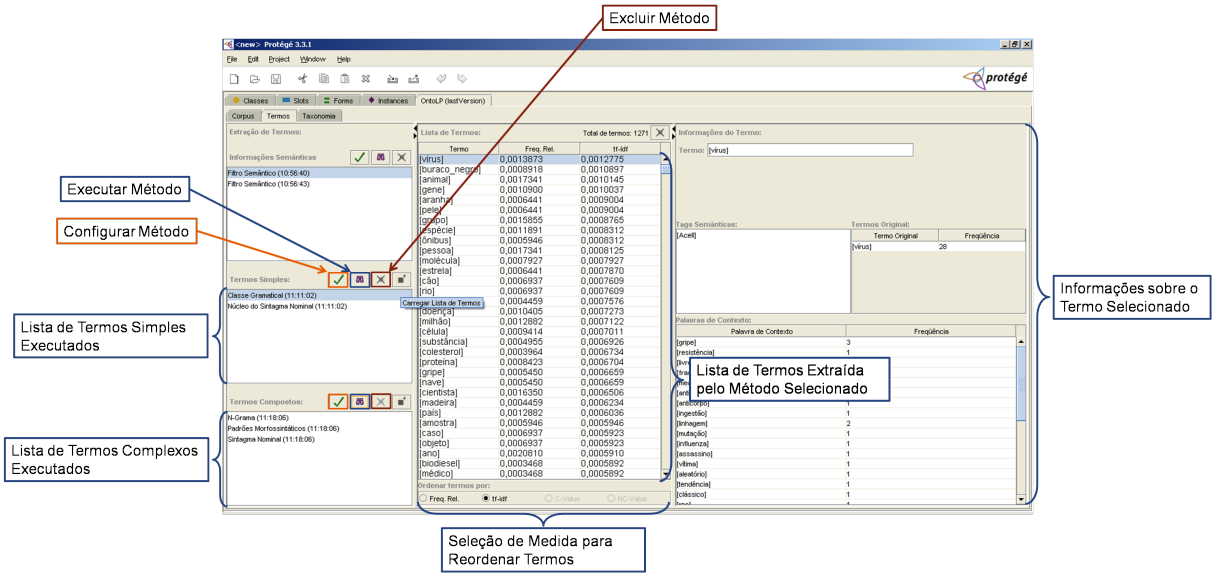


Figura 50: Interface de Extração de Termos (Termos Simples e Termos Complexos).

1) utilizando a saída de um método como entrada para outro, procurando melhorar os resultados do segundo; (Abordagem 2) os métodos são executados de forma independente. Cada uma delas explicada nas seções abaixo.

6.2.1- Abordagem 1:

A primeira etapa da abordagem 1 agrupa os termos extraídos do corpus em Grupos Semânticos. Por exemplo, os termos {braço, perna, tórax, cabeça} pertencem ao grupo "Anatomia". Já os termos {casa, prédio, apartamento, edifício} pertencem ao grupo "Construção". Para utilizarmos essas informações o método Filtro por Grupos Semânticos

deve estar habilitado, conforme as configurações abaixo (Figura 51):

1. Habilitar Filtro Semântico: habilita/desabilita o método.

2. Aplicar aos Cálculos Estatísticos dos Termos: quando habilitada, esta opção faz com que os valores de relevância do grupo semântico sejam utilizados no cálculo de relevância dos termos simples e complexos. Por padrão esta opção aparece desabilitada.

Figura 51: Painel de Configuração do Filtro por Grupos Semânticos.

Quando executado o método extrai os grupos semânticos encontrados no corpus e os apresenta ao usuário em ordem decrescente de relevância (Figura 52).

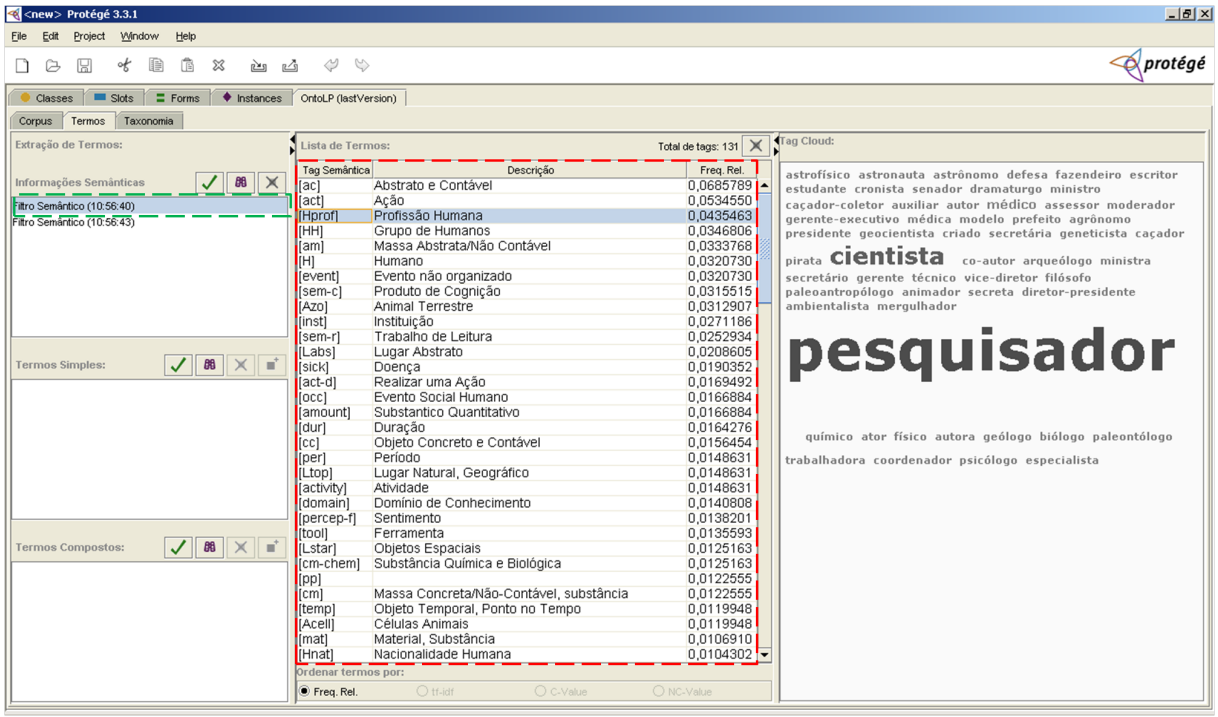


Figura 52: Para visualizar a lista de Grupos Semânticos selecione o método depois de executado (destacado em verde). A lista de Grupos Semânticos ordenada pela relevância é destacada em vermelho.

Depois de feita a extração dos Grupos Semânticos, o sistema possibilita ao usuário visualizar os termos presentes em cada grupo. A relevância do termo para seu grupo é dada pelo tamanho e cor de sua fonte, como destacado na Figura 53, onde "doença" é o termo mais relevante.

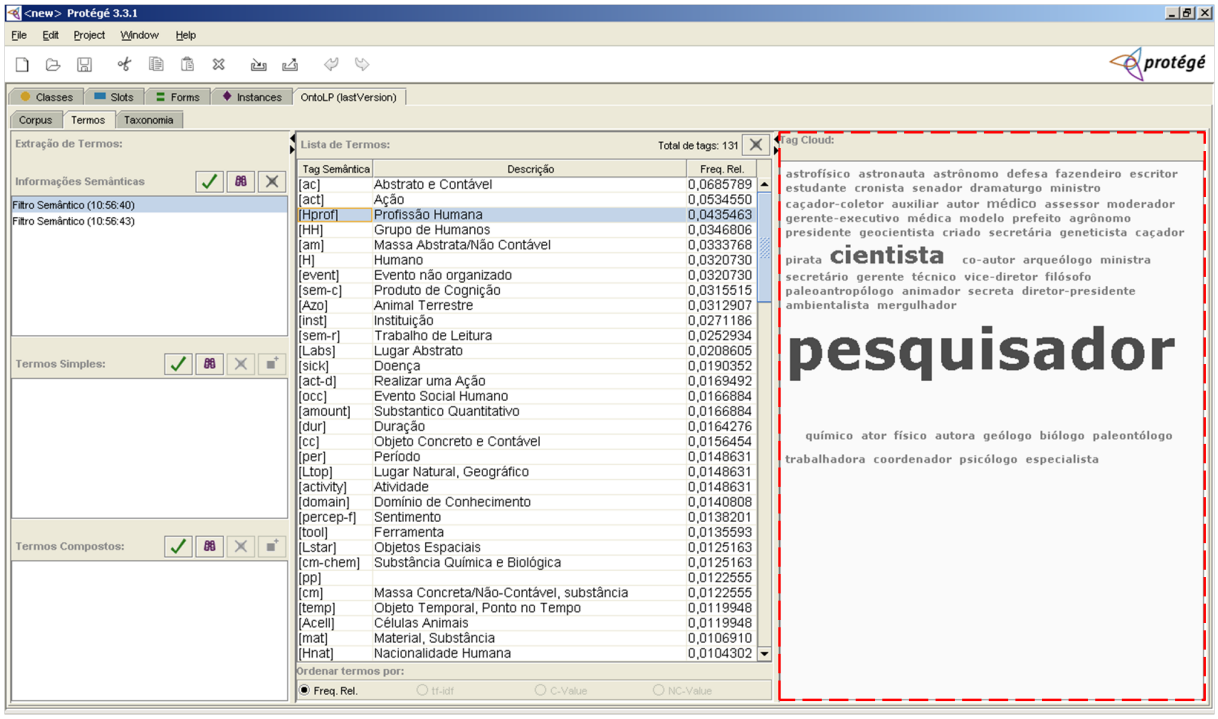


Figura 53: Termos extraídos para o Grupo Semântico [sick], sendo "doença" o termo mais relevante.

A ordem dos Grupos Semânticos e as diferentes fontes dos termos, ambos indicativos de relevância, são estratégias para auxiliar o engenheiro na exclusão dos grupos que não tem relação com o domínio em questão. Para realizar a exclusão de um grupo basta selecioná-lo e pressionar o botão destacado na Figura 54. Feito isso, os métodos de extração de termos simples e complexos (etapas posteriores) descartarão os termos pertencentes aos grupos excluídos.

Como o plug-in possibilita ao usuário executar a extração de grupos semânticos inúmeras vezes. Para indicar qual das listas deve estar em uso em determinado momento, basta mantê-la selecionada (Figura 55).

A etapa seguinte visa extrair termos simples (unigramas) através de dois diferentes

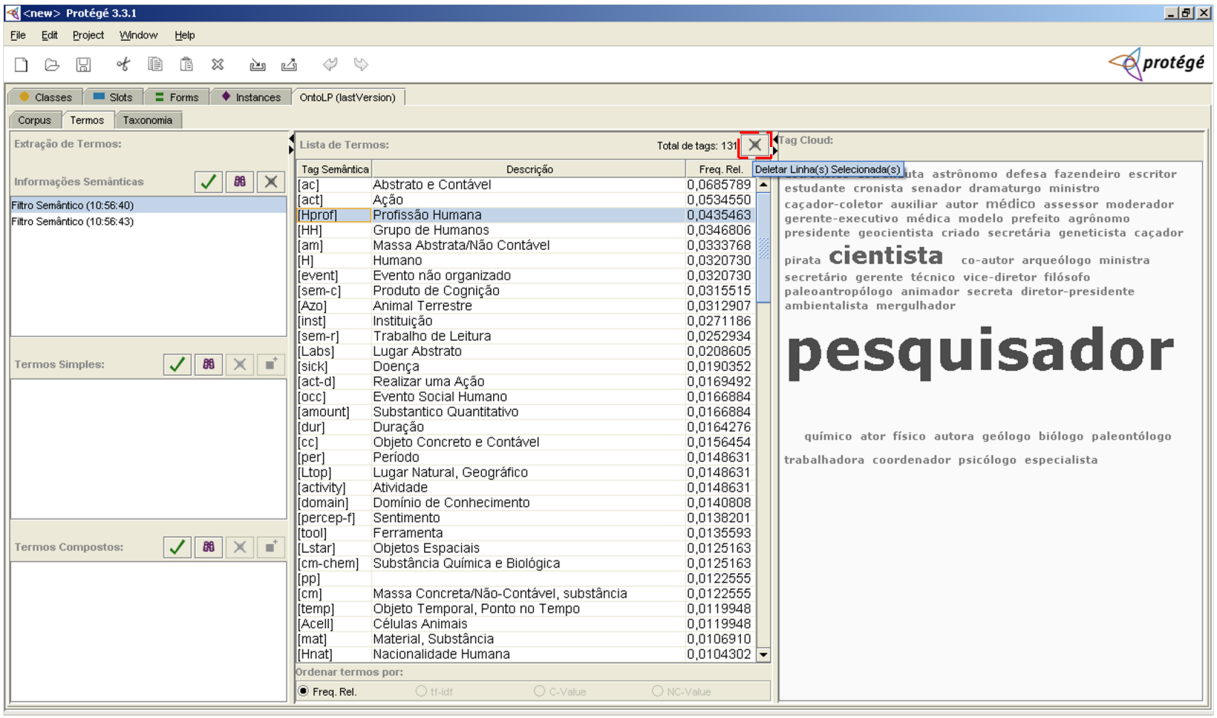


Figura 54: Botão de exclusão das linhas selecionadas na tabela.

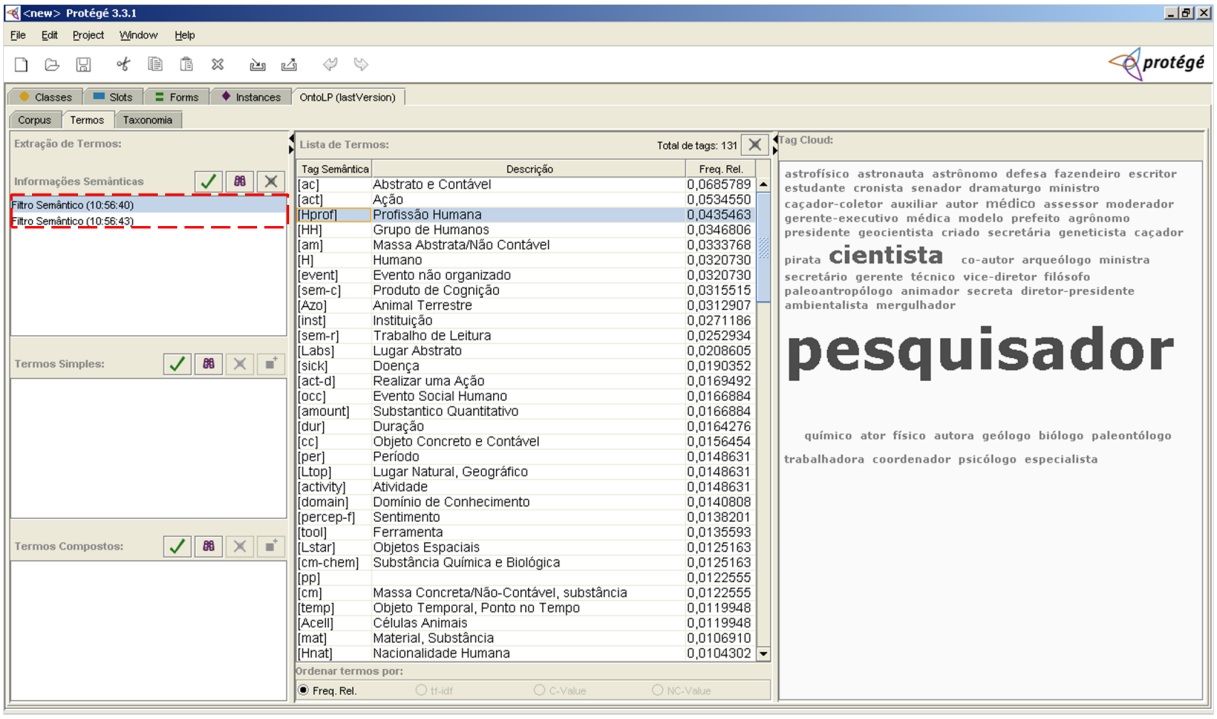


Figura 55: Execuções do método de Grupos Semânticos e seleção de uma lista de Grupos Semânticos.

métodos. Esses métodos possuem configurações distintas. Para carregar o painel de configuração deles pressione o botão destacado na Figura 56. É apresentada então uma janela com duas abas, uma para cada método (Figura 57), com as seguintes opções:

•Método Classe Gramatical

1.Habilitar método: habilita/desabilita o método.

2.Configurar medidas: possibilita selecionar as medidas de relevância dos termos e definir um limiar de corte para cada medida.

3.Classes gramaticais aceitas: possibilita selecionar quais classes gramaticais devem ser extraídas como possíveis conceitos de uma ontologia.

•Método Núcleo do Sintagma Nominal

1.Habilitar método: habilita/desabilita o método.

2.Configurar medidas: possibilita selecionar as medidas de relevância dos termos e definir um limiar de corte para cada medida.

3.Classes gramaticais aceitas: nesse caso, não é possível selecionar outras classes gramaticais, visto que o método seleciona somente núcleo de sintagmas nominais.

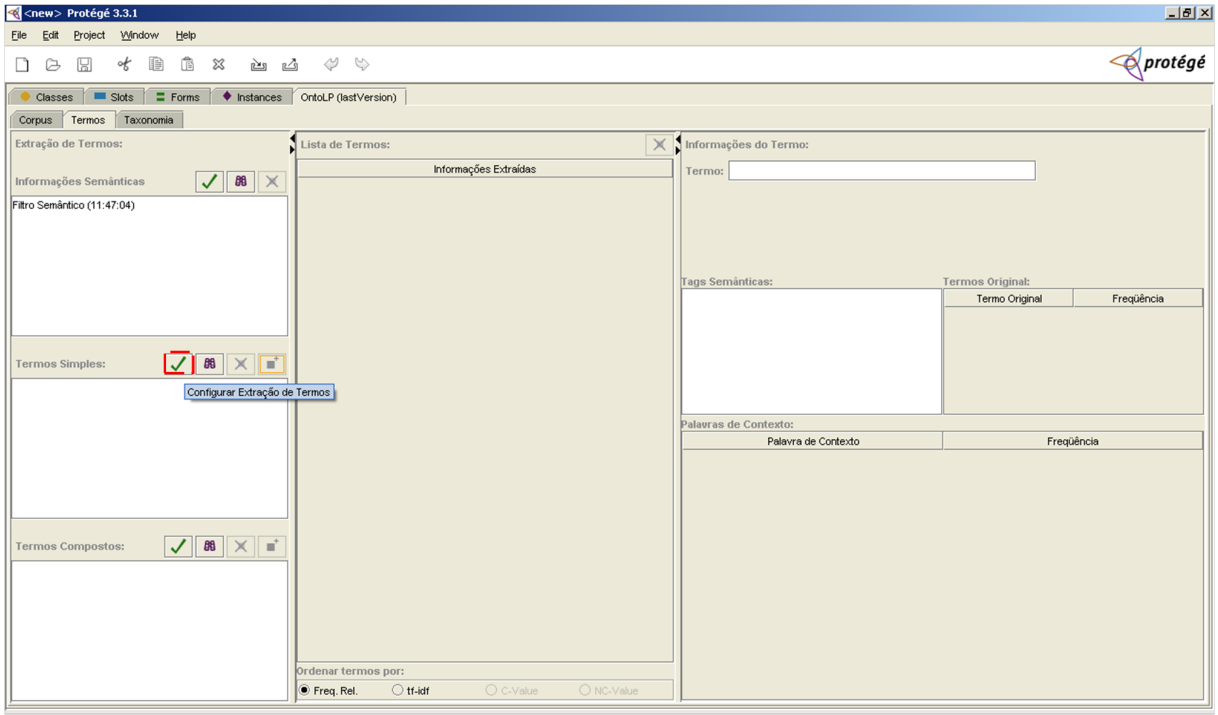


Figura 56: Botão de configuração dos métodos de extração de termos simples.

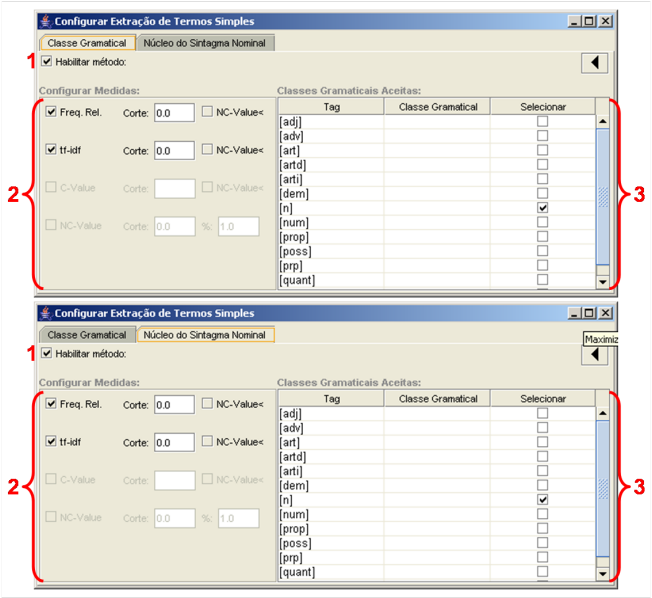


Figura 57: Interface de configuração dos métodos de extração de termos simples.

Depois de executados, os métodos de extração de termos possuem funcionalidades semelhantes as do Filtro por Grupos Semânticos. Assim como nos Grupos, a lista de termos simples pode ser visualizada (Figura 58) e editada pelo usuário, excluindo os termos irrelevantes (Figura 54).

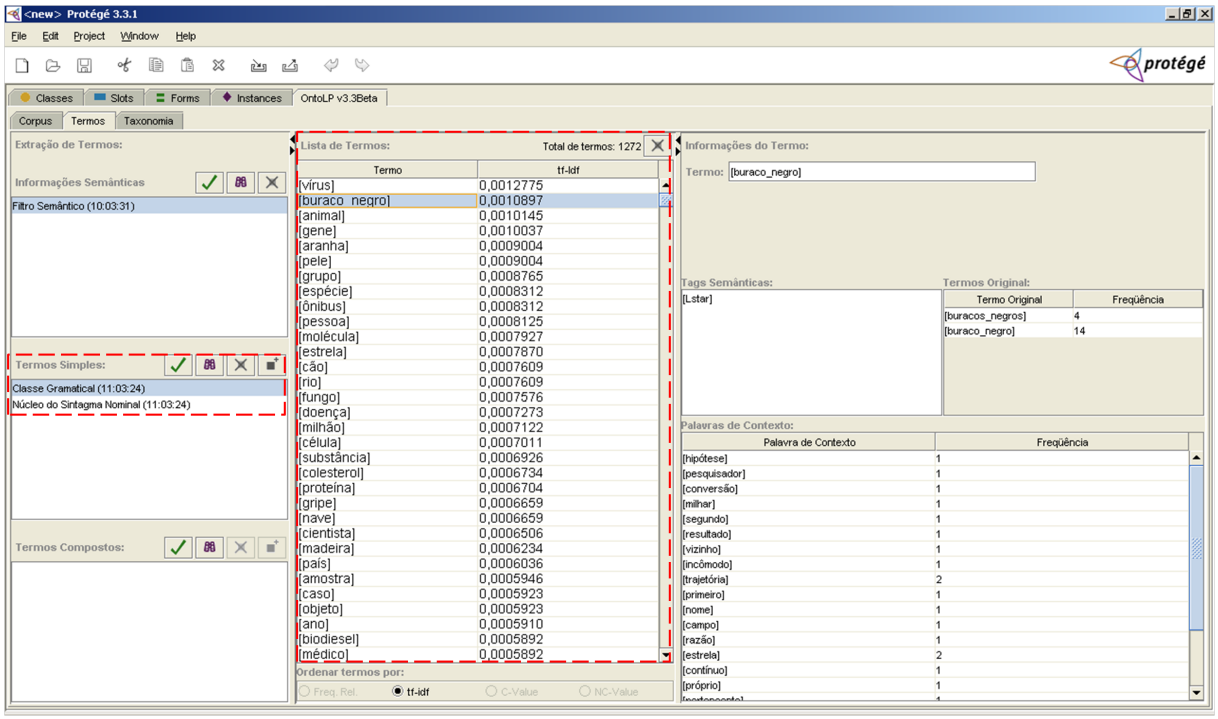


Figura 58: Seleção do método e lista de termos relacionada a ele.

Além disso, o plug-in disponibiliza informações sobre cada termo, como:

•A qual Grupo Semântico o termo está relacionado (Figura 59).

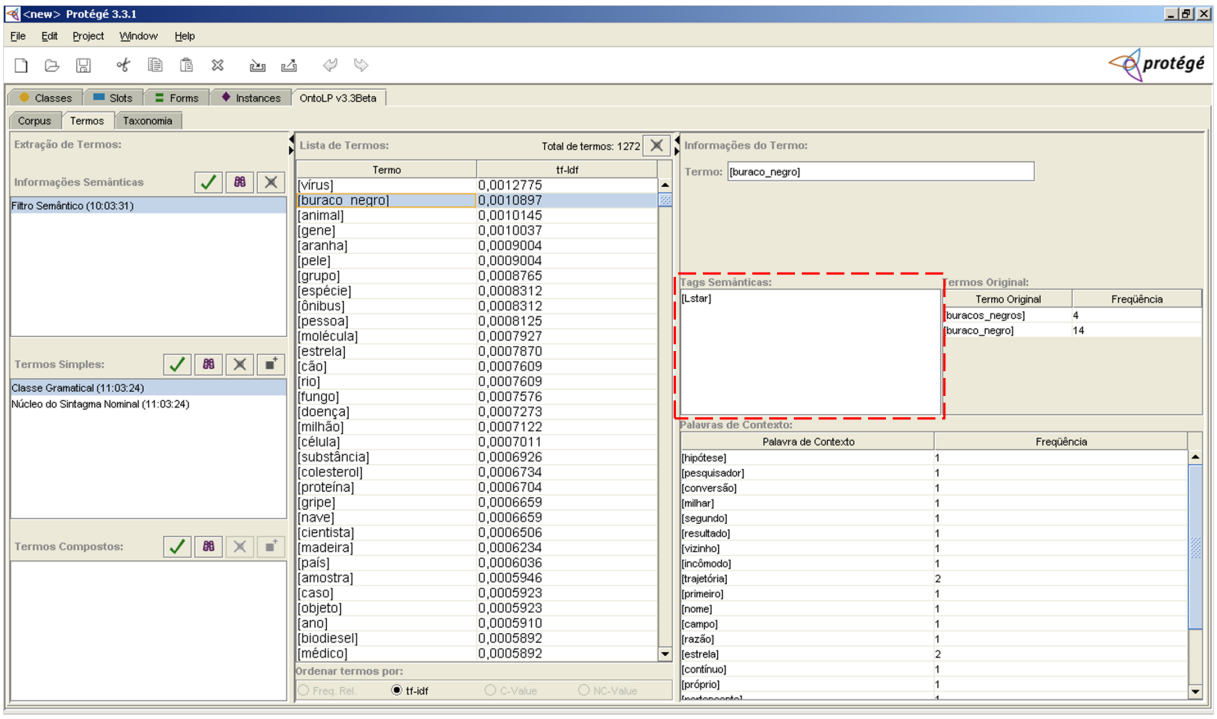


Figura 59: Grupo Semântico ao qual o termo está relacionado.

•Forma original dos termos no corpus e a freqüência de cada uma delas (Figura 60).

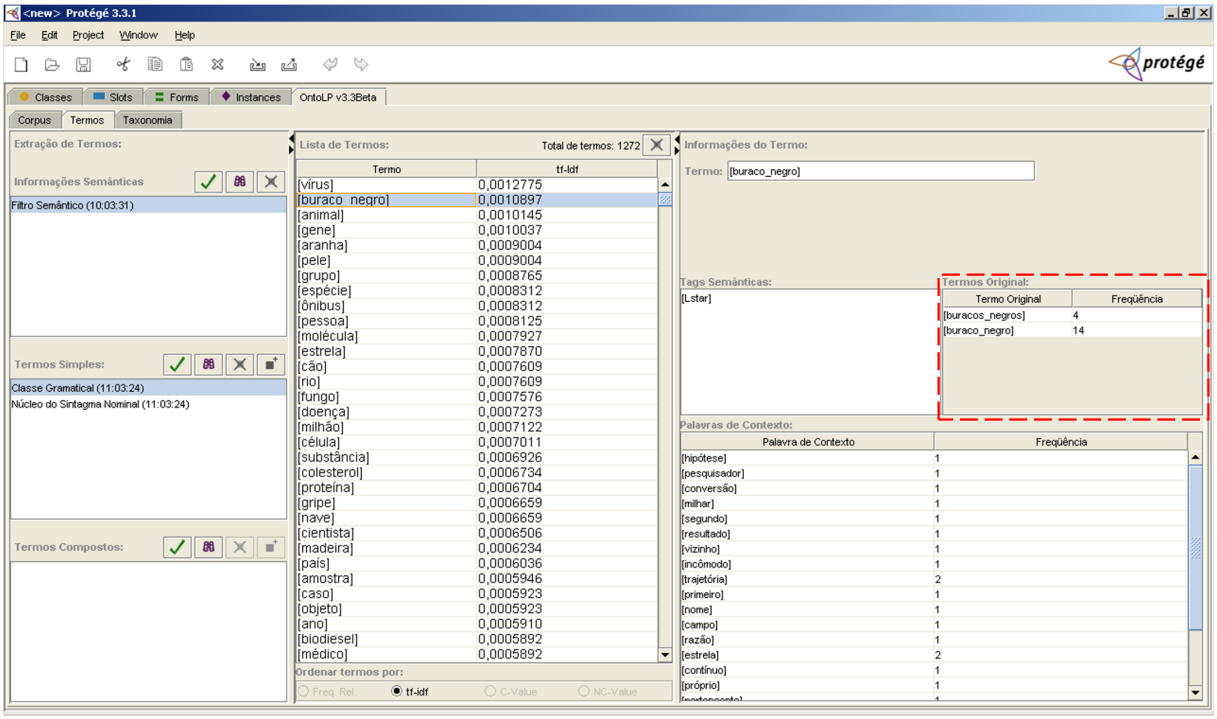


Figura 60: Forma original que o termo apareceu no corpus.

•Palavras que aparecem na mesma sentença do termo e a freqüência com que se repetem (Figura 61).

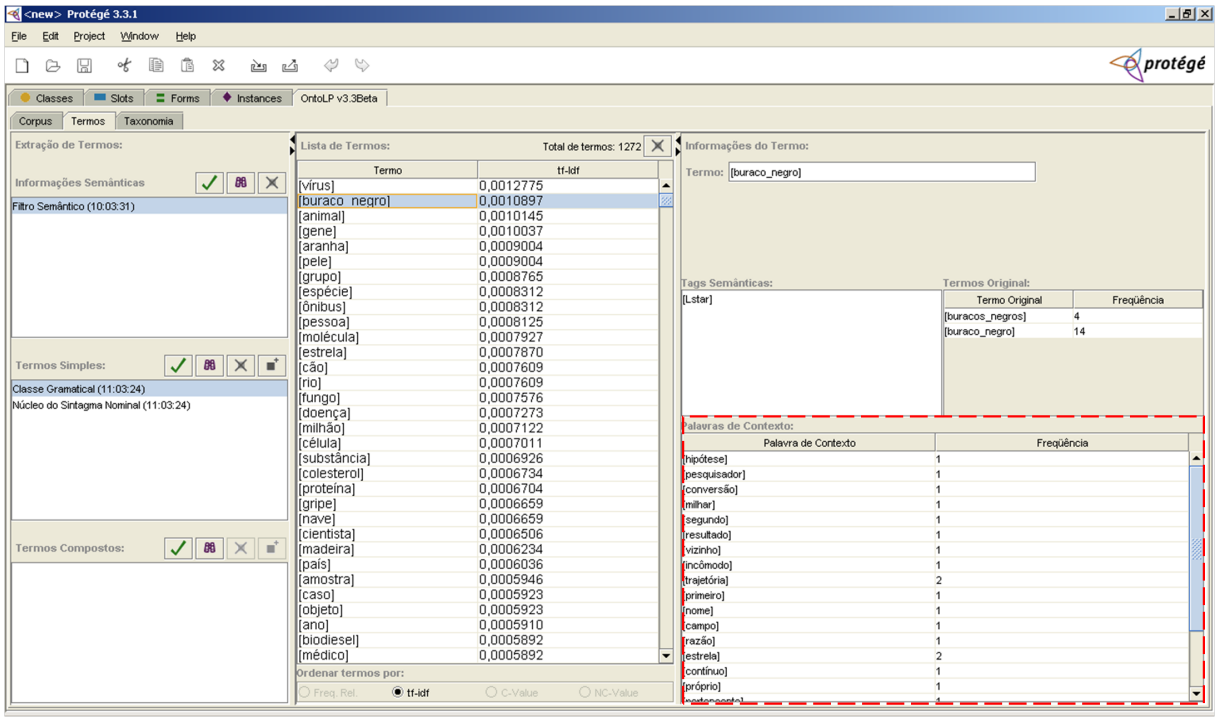


Figura 61: Lista de palavras que aparecem próximas ao termo.

A última etapa, extração de termos complexos, assim como a anterior, possui um conjunto de opções de configuração para cada método. Nas Figuras 62, 63 e 64 são apresentadas as interfaces de configuração, as opções são explicadas a seguir:

•N-Gram

1. Habilitar método: habilita/desabilita o método.
2. Configurar Medidas: possibilita selecionar as medidas de relevância dos termos e definir um limiar de corte para cada medida.
3. Configurações Gerais: possibilita definir o tamanho do termo bigrama em número de palavras e habilitar ou desabilitar a opção "Restrição por unigrama".
4. Classes Gramaticais Aceitas: possibilita selecionar quais classes gramaticais devem ser aceitas na construção de termos complexos.

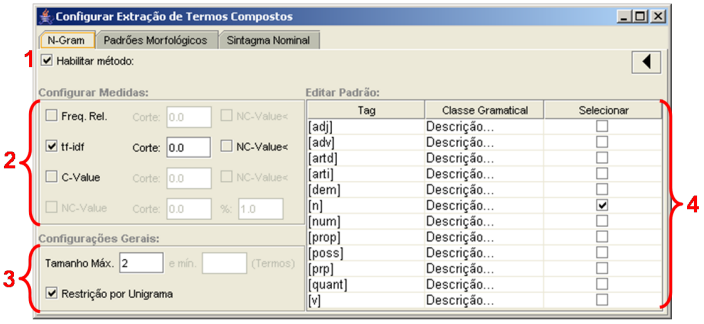


Figura 62: Interface de configuração do método N-Grama.

•Padrões Morfológicos

1.Habilitar método: habilita/desabilita o método.

2.Configurar Medidas: possibilita selecionar as medidas de relevância dos termos e definir um limiar de corte para cada medida.

3.Configurações Gerais: possibilita definir o tamanho máximo e mínimo dos termos em número de palavras e habilitar ou desabilitar a opção "Restrição por unigrama".

4.Editar padrões: essa opção ainda está desabilitada, a idéia é que o usuário possa selecionar os padrões que deseja extrair do corpus.

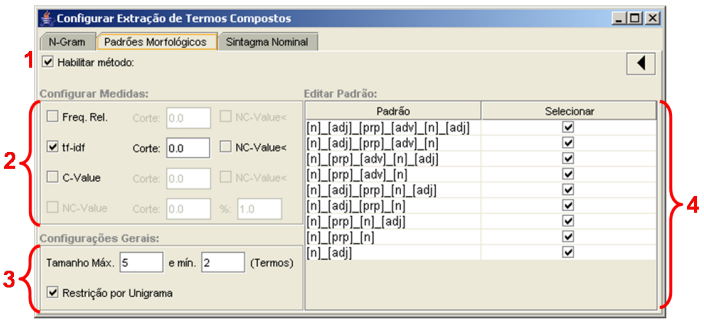


Figura 63: Interface de configuração do método Padrões Morfológicos.

•Sintagma Nominal

1.Habilitar método: habilita/desabilita o método.

2.Configurar Medidas: possibilita selecionar as medidas de relevância dos termos e definir um limiar de corte para cada medida.

3.Configurações Gerais: possibilita definir o tamanho máximo e mínimo dos termos em número de palavras e habilitar ou desabilitar a opção “Restrição por unigrama”.

4.Classes Gramaticais Aceitas: essa opção está desabilitada.

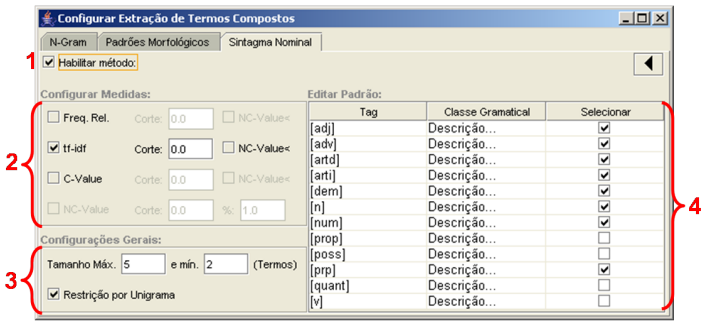


Figura 64: Interface de configuração do método Sintagma Nominal.

Para que a etapa de extração de termos complexos execute conforme a Abordagem 1, a opção “Restrição por unigrama” deve estar habilitada nos métodos utilizados. Quando habilitada o engenheiro deve selecionar uma das listas de termos simples para ser utilizada como entrada dos métodos de extração de termos complexos. Essa opção restringe os termos complexos, como demonstrado no exemplo da Figura 65. Os métodos de extração de termos complexos recebem uma lista de termos simples ({animal, pele}) e extraem somente termos constituídos por palavras presentes nessa lista ({animal_doméstico, pele_artificial}), os demais são descartados ({célula_hepática}).

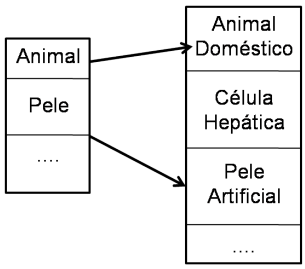


Figura 65: Exemplo de restrição por unigrama.

6.2.2- Abordagem 2:

A Abordagem 2 disponibiliza funcionalidades semelhantes as da Abordagem 1, entretanto, a saída de um método não é utilizada como entrada para outro (Filtro por

Grupos Semânticos→Termos Simples→Termos Complexos). Para utilizar esta abordagem desabilite o uso de Filtro por Grupos Semânticos e a opção "Restrição por unigrama" em cada método de extração de termos complexos. Dessa forma, esses métodos irão extrair todos os termos aptos, sem restringi-los com base na lista de termo simples.

6.3- Observação:

Para copiar os termos ou grupos semânticos basta selecionar a lista, copiar (ctrl+c) e colar (ctrl+v) em outro programa (Figura 66).

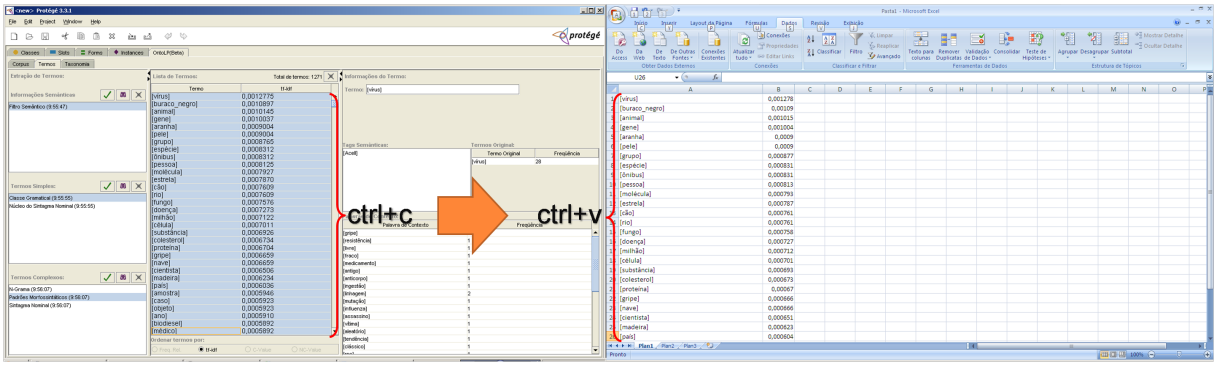


Figura 66: Exemplo de cópia de uma lista de termos do plug-in para o Excel.

7- Aba de Organização Hierárquica dos Termos (Taxonomia):

Um guia rápido para a Interface de Organização Hierárquica dos Termos é apresentado na Figura 67.

7.2- Organização Hierárquica dos Termos:

A aba de organização hierárquica de termos disponibiliza três métodos para o usuário: (1) Termos Complexos; (2) Padrões de Hearst e (3) Padrões de Morin. No primeiro caso, o usuário deve selecionar uma lista de termos simples e uma lista de termos complexos como entrada para o método. O método percorre as listas em busca de termos compostos (sistema_nervoso) que sejam constituídos de no mínimo uma palavra presente na lista de termos simples (sistema). Quando o teste é satisfeito o termo complexo é adicionado a taxonomia como hipônimo do termo simples. Na Figura 68 são apresentados alguns exemplos de organização hierárquica com base em termos compostos.

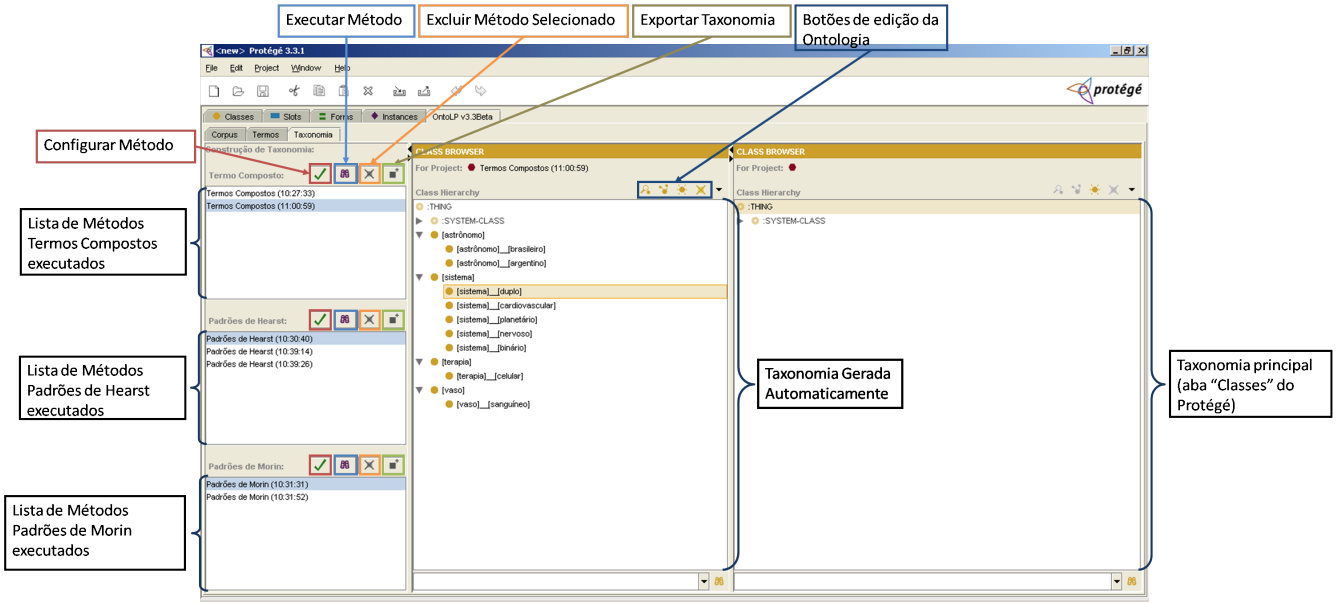


Figura 67: Aba de Organização Hierárquica dos Termos.

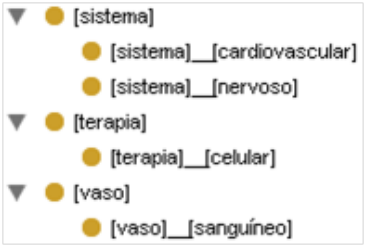


Figura 68: Exemplo de relação hierárquica extraída pelo método Termos Complexos.

Na Figura 69 é apresentada a interface de configuração do método e suas opções são descritas abaixo:

1. Habilitar método: habilita/desabilita o método.

2. Inicio do termo ("casa", "casa_bela"): essa opção compara somente a primeira palavra dos termos complexos com os presentes na lista de termos simples. Nesse caso, selecionaria relações como no exemplo "casa" e "casa_bela".

3. Fim do termo ("casa", "bela_casa"): essa opção compara somente a última palavra dos termos complexos com os presentes na lista de termos simples. Nesse caso, selecionaria relações como no exemplo "casa" e "bela_casa".

Caso o usuário deixe as opções 2 e 3 desabilitadas, o método selecionará ambos os

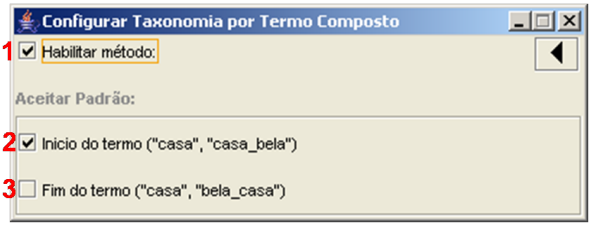


Figura 69: Painel de configuração do método Termos Complexos.

padrões, "casa"→"casa_bela" e "casa"→"bela_casa".

Nos dois últimos métodos, padrões de Hearst e de Morin, o sistema percorre o corpus anotado procurando por ocorrências de determinadas estruturas lingüísticas que indicam relações de hiperonímia/hiponímia entre conceitos. Como ambos os métodos trabalham de forma semelhante, eles possuem as mesmas opções de configuração, conforme apresentado na Figura 71 e explicadas abaixo:

1. Habilitar método: habilitar/desabilitar o método.

2. Mínimo uma tag semântica igual: aceita apenas relações onde exista no mínimo uma tag semântica igual entre os conceitos selecionados. Na Figura 70 é apresentado um exemplo, onde o conceito "neurônio" está no grupo "<an>" e o hipônimo "célula_nervosa" está nos grupos "<Acell>" e "<an>", esse último validando a restrição imposta.

Figura 70: Exemplo de restrição "Mínimo uma tag semântica igual".

3. Somente termos selecionados: essa opção restringe a extração de relações apenas aos termos presentes nas listas de termos simples e complexos selecionados na aba de Termos.

4. Selecionar Padrões: possibilita que o usuário selecione quais as estruturas lingüísticas quer considerar durante a extração de relações hierárquica.

8- Conclusão:

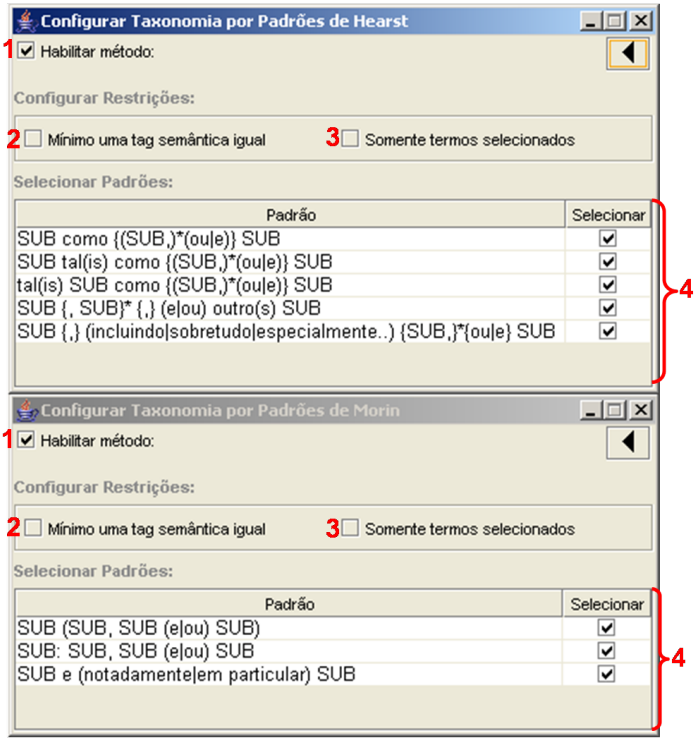


Figura 71: Janelas de configuração dos padrões de Hearst e Morin.

O plug-in OntoLP é uma ferramenta em construção, portanto, algumas inconsistências ainda podem ocorrer. Caso você perceba algum comportamento inadequado pedimos para reportar o erro (xxxxxxxx@gmail.com), indicando o procedimento causador. Outra limitação atual é a utilização de corpora muito grandes, o que está sendo corrigido.

Anexo C - Taxonomias do processo de avaliação da organização hierárquica

1 Thing
  1.1 CARNIVOROS
  1.2 HABITAT

Figura 72: Taxonomia extraída pelo método Padrões de Hearst.

1 Thing
  1.1 SOLO
  1.2 LAGO
  1.3 MANGUE

Figura 73: Taxonomia extraída pelo método Padrões de Morin.

1 Thing
- 1.1 BIOTA
- 1.2 CONSUMIDORES
- 1.3 CARNIVOROS
- 1.4 DETRITIVOS
- 1.5 HEMATOFAGOS
- 1.6 HERBIVOROS
- 1.7 INSETIVOROS
- 1.8 ONIVOROS
- 1.9 DECOMPOSITORES
- 1.10 BACTERIAS
- 1.11 FUNGOS
- 1.12 PRODUTORES_PRIMARIOS
- 1.13 BIOTOPO
- 1.14 CLIMA
- 1.15 ARTICO
- 1.16 EQUATORIAL
- 1.16 SUBARTICO
- 1.18 TEMPERADO
- 1.19 TROPICAL
- 1.20 HABITAT
  - 1.20.1 HABITAT_AQUATICO
    - 1.20.1.1 HABITAT_AQUATICO_DOCE
      - 1.20.1.1.1 HABITAT_AQUATICO_DOCE_LENTICO
      - 1.20.1.1.2 HABITAT_AQUATICO_DOCE_LOTICO
    - 1.20.1.2 HABITAT_AQUATICO_SALGADO
    - 1.20.2.3 HABITAT_AQUATICO_SALOBRO
  - 1.20.2 HABITAT_TERRESTRE
- 1.21 LAGO
- 1.22 LAGOA
- 1.23 PANTANO
- 1.24 REPRESA
- 1.25 CORREGO
- 1.26 RIO
- 1.28 LAGOAS_SALINAS
- 1.29 OCEANO
- 1.30 MAR
- 1.31 MANGUE
- 1.32 CAMPO
- 1.33 CAPOEIRA
- 1.34 DESERTO
- 1.35 DUNA
- 1.36 FLORESTA
- 1.37 SAVANA
- 1.38 SOLO
- 1.39 LATOSSOLO
- 1.40 PODZOL
- 1.41 SERPENTINO
- 1.42 AMENSALISMO
- 1.43 COMPETICAO
- 1.44 ESCLAVAGISMO
- 1.45 PARASITISMO
- 1.46 PREDATISMO
- 1.47 COMENSALISMO
- 1.48 FORESIA
- 1.49 INQUILINISMO
- 1.50 MUTUALISMO
- 1.51 INTERACOES_INTERESPECIFICAS
  - 1.51.1 INTERACOES_INTERESPECIFICAS_DESARMONICAS
  - 1.51.2 INTERACOES_INTERESPECIFICAS_HARMONICAS

Figura 74: Taxonomia extraída pelo método baseado em Termos Complexos.

1 Thing
- 1.1 BIOTA
  - 1.1.1 CONSUMIDORES
    - 1.1.1.1 CARNIVOROS
    - 1.1.1.2 DETRITIVOS
    - 1.1.1.3 HEMATOFAGOS
    - 1.1.1.4 HERBIVOROS
    - 1.1.1.5 INSETIVOROS
    - 1.1.1.6 ONIVOROS
  - 1.1.2 DECOMPOSITORES
    - 1.1.2.1 BACTERIAS
    - 1.1.2.2 FUNGOS
  - 1.1.3 PRODUTORES_PRIMARIOS
- 1.2 BIOTOPO
  - 1.2.1 CLIMA
    - 1.2.1.1 ARTICO
    - 1.2.1.2 EQUATORIAL
    - 1.2.1.3 SUBARTICO
    - 1.2.1.4 TEMPERADO
    - 1.2.1.5 TROPICAL
  - 1.2.2 HABITAT
    - 1.2.2.1 HABITAT_AQUATICO
      - 1.2.2.1.1 HABITAT_AQUATICO_DOCE
        - 1.2.2.1.1.1 HABITAT_AQUATICO_DOCE_LENTICO
          - 1.2.2.1.1.1.1 LAGO
          - 1.2.2.1.1.1.2 LAGOA
          - 1.2.2.1.1.1.3 PANTANO
          - 1.2.2.1.1.1.4 REPRESA
        - 1.2.2.1.1.2 HABITAT_AQUATICO_DOCE_LOTICO
          - 1.2.2.1.1.2.1 CORREGO
          - 1.2.2.1.1.2.2 RIO
      - 1.2.2.1.2 HABITAT_AQUATICO_SALGADO
        - 1.2.2.1.2.1 LAGOAS_SALINAS
        - 1.2.2.1.2.2 OCEANO
          - 1.2.2.1.2.2.1 MAR
      - 1.2.2.1.3 HABITAT_AQUATICO_SALOBRO
        - 1.2.2.1.3.1 MANGUE
    - 1.2.2.2 HABITAT_TERRESTRE
      - 1.2.2.2.1 CAMPO
      - 1.2.2.2.2 CAPOEIRA
      - 1.2.2.2.3 DESERTO
      - 1.2.2.2.4 DUNA
      - 1.2.2.2.5 FLORESTA
      - 1.2.2.2.6 SAVANA
  - 1.2.3 SOLO
    - 1.2.3.1 LATOSSOLO
    - 1.2.3.2 PODZOL
    - 1.2.3.3 SERPENTINO
- 1.3 INTERACOES_INTERESPECIFICAS
  - 1.3.1 INTERACOES_INTERESPECIFICAS_DESARMONICAS
    - 1.3.1.1 AMENSALISMO
    - 1.3.1.2 COMPETICAO
    - 1.3.1.3 ESCLAVAGISMO
    - 1.3.1.4 PARASITISMO
    - 1.3.1.5 PREDATISMO
  - 1.3.2 INTERACOES_INTERESPECIFICAS_HARMONICAS
    - 1.3.2.1 COMENSALISMO
    - 1.3.2.2 FORESIA
    - 1.3.2.3 INQUILINISMO
    - 1.3.2.4 MUTUALISMO

Figura 75: Ontologia do domínio de Ecologia de Comunidades, utilizada como referência.

Anexo D - Resultados obtidos por (CIMIANO; HOTHO; STAAB, 2005) para a Organização Hierárquica de Conceitos

Tabela 34: Resultados obtidos pela abordagem proposta por (CIMIANO; HOTHO; STAAB, 2005) durante a tarefa de organização hierárquica de conceitos, nos domínios de Turismo e Financeiro.

Domínio	**Precisão**	**Abrangência**	**F-measure**	**F-measure'**
Turismo	29,33%	65,49%	40,52%	44,69%
Financeiro	29,93%	37,05%	33,11%	38,85%

Anexo E - Questionário de Avaliação do plug-in OntoLP

Este questionário é composto de três etapas e faz parte de uma pesquisa de mestrado relacionada à Construção Semi-Automática de Ontologias. Como a tarefa de construção de ontologias é extremamente custosa o projeto visa disponibilizar uma ferramenta de auxílio às etapas de extração e organização hierárquica de termos provenientes de corpora em língua portuguesa. Na Parte I do questionário estão perguntas que procuram definir o nível de conhecimento do usuário sobre o processo de construção de ontologias. Nas Partes II e III são realizadas tarefas que visam, respectivamente, à extração de termos candidatos a conceitos e a organização desses em uma taxonomia. Na última parte do questionário estão descritos os matérias que devem ser retornados. Bom trabalho!

Parte I:

Avaliação da Experiência do Usuário:

1.Quais etapas de construção de uma Ontologia você já realizou?

a. Extração de termos candidatos a conceitos ( )

b. Organização hierárquica dos termos (taxonomia) ( )

c. Identificação de relações não hierárquicas ( )

d. Povoamento de ontologias (instâncias) ( )

e. Criação de axiomas ( )

2.Caso já tenha realizado a extração de termos candidatos a conceitos, qual critério foi utilizado?

a. Semântico (extração dos termos através da leitura do corpus) ( )

b. Semi-automático (com auxílio de um sistema de extração) ( )

3.Caso tenha selecionado a alternativa 'b' na questão anterior, quais sistemas já utilizou?

4.Caso já tenha realizado a organização hierárquica dos termos, qual abordagem utilizou?

a. Manual ( )

b. Semi automática ( )

5.Caso tenha selecionado a alternativa 'b' na questão anterior, quais sistemas já utilizou?

6.Você se considera: ( ) experiente ( ) razoavelmente experiente ( ) inexperiente na tarefa de construção de ontologias.

Parte II:

Extração de Termos - Abordagem 1

1.Executar o método de Filtro Semântico:

a. Mantenha a configuração padrão.

b. Extraia os Grupos Semânticos.

c. Exclua os Grupos Semânticos que considera não ter relação com o domínio.

d. Salve a lista de Grupos Semânticos final na planilha Excel.

e. Envie a lista final para xxxxxxxx@gmail.com assim que estiver pronta.

2.Executar os métodos de extração de termos simples (unigramas):

a. Mantenha a configuração padrão.

b. Extraia os termos.

c. Exclua os termos que considera não relevantes para o domínio, considerando cada um dos métodos (Classe Gramatical e Núcleo do Sintagma Nominal).

3.Execução dos métodos de extração de termos complexos (n-gramas, onde n>1):

a. Mantenha a configuração padrão.

b. Extraia os termos.

c. Exclua os termos que considera não relevantes para o domínio, considerando cada um dos métodos (N-Gram, Padrões Morfológicos e Sintagma Nominal).

Extração de Termos - Abordagem 2

1.Método Filtro Semântico:

a. Desmarcar "Habilitar Filtro Semântico"

2.Executar os métodos de extração de termos simples (unigramas):

a. Mantenha a configuração padrão.

b. Extraia os termos.

c. Exclua os termos que considera não serem relevantes para o domínio, considerando cada um dos métodos (Classe Gramatical e Núcleo do Sintagma Nominal).

3.Execução dos métodos de extração de termos complexos (n gramas, onde n>1):

a. Desmarque a opção "Restrição por Uni-grama" de todos os métodos.

b. Extraia os termos.

c. Exclua os termos que considera não serem relevantes para o domínio, considerando cada um dos métodos (N-Gram, Padrões Morfológicos e Sintagma Nominal).

Avaliação da Etapa de Extração (Pontue as funcionalidades utilizando o critério: 0-Não auxiliou e aumentou o trabalho; 1-Não auxiliou; 2-Auxiliou e 3-Auxiliou muito)

1.Quanto ao auxílio na seleção dos Grupos Semânticos:

a. Organização por relevância ( )

b. Tag clouds ( )

c. A definição do Grupo ( )

2.Quanto ao auxílio na Extração de Termos Simples:

a. Organização por relevância ( )

b. Tags Semânticas ( )

c. Termo Original ( )

d. Palavras de Contexto ( )

e. Filtro por Grupos Semânticos (etapa anterior) ( )

3.Quanto ao auxílio na Extração de Termos Complexos:

a. Organização por relevância ( )

b. Tags Semânticas ( )

c. Termo Original ( )

d. Palavras de Contexto ( )

e. Filtro por Grupos Semânticos ( )

f. Restrição por Uni gramas ( )

4.Concluindo, qual das abordagens você achou melhor para a extração de termos?

a. Abordagem 1 ( )

b. Abordagem 2 ( )

5.Esse espaço é destinado a sugestões e/ou críticas sobre a etapa de extração de termos:

_____-

Parte III:

Organização Hierárquica dos Termos

1.Método baseado em Termos Complexos:

a. Mantenha a configuração padrão.

b. Execute o método selecionando como entrada um par de listas de termos (simples e complexos) por vez.

c. Analise e altere as taxonomias geradas.

d. Exporte para a taxonomia final.

2.Método baseado nos Padrões de Hearst e Morin:

a. Mantenha a configuração padrão.

b. Execute o método.

c. Analise e altere as taxonomias geradas.

d. Exporte para a taxonomia final.

3.Após todos os passos anteriores, salve o projeto final (processo de salvar projetos no Protégé).

Avaliação da Etapa de Organização Hierárquica (Pontue as funcionalidades utilizando o critério: 0-Não auxiliou e aumentou o trabalho; 1-Não auxiliou; 2-Auxiliou e 3-Auxiliou muito)

1.Método baseado em Termos Complexos ( )

2.Método baseado nos Padrões de Hearst ( )

3.Método baseado nos Padrões de Morin ( )

4.Visualização da Taxonomia Inferida ao lado da Taxonomia em Construção ( )

5.Esse espaço é destinado a sugestões e/ou críticas sobre a etapa de organização hierárquica: