

Manual OntoLP

Sumário:

1-Introdução ao OntoLP.....	2
2-Instalação do OntoLP.....	2
3-Executando o OntoLP.....	2
4-Observação Importante.....	4
5-Aba de Carga do Corpus.....	5
6-Aba de Extração de Termos.....	7
7- Aba de Organização Hierárquica dos Termos (Taxonomia).....	17
8-Conclusão.....	19

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 um corpus de entrada no formato XCES (padrão de representação de informações lingüísticas adotado no projeto PLNBR) e o analisador sintático que disponibiliza tais informações deve ser o PALAVRAS.

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. Cada uma dessas abas é explicada melhor nas seções 5, 6 e 7.

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.
3. Na figura 1 é demonstrado como deve ficar a pasta “plugins”.

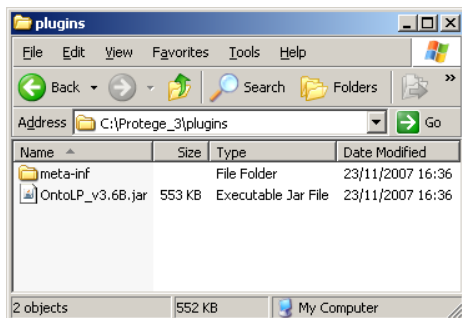


Figura 1. Arquivos de instalação do OntoLP.

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 2) selecione “new Project” ou o projeto em uso.

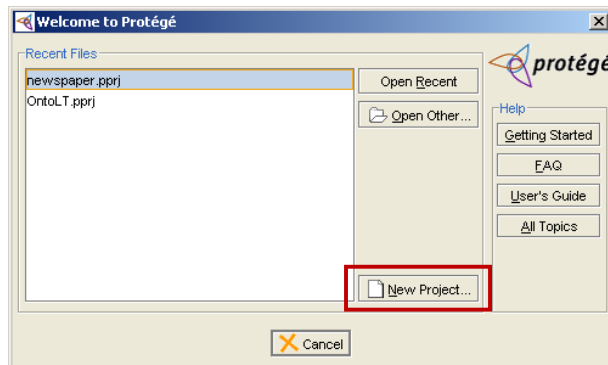


Figura 2. Welcome to Protégé.

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

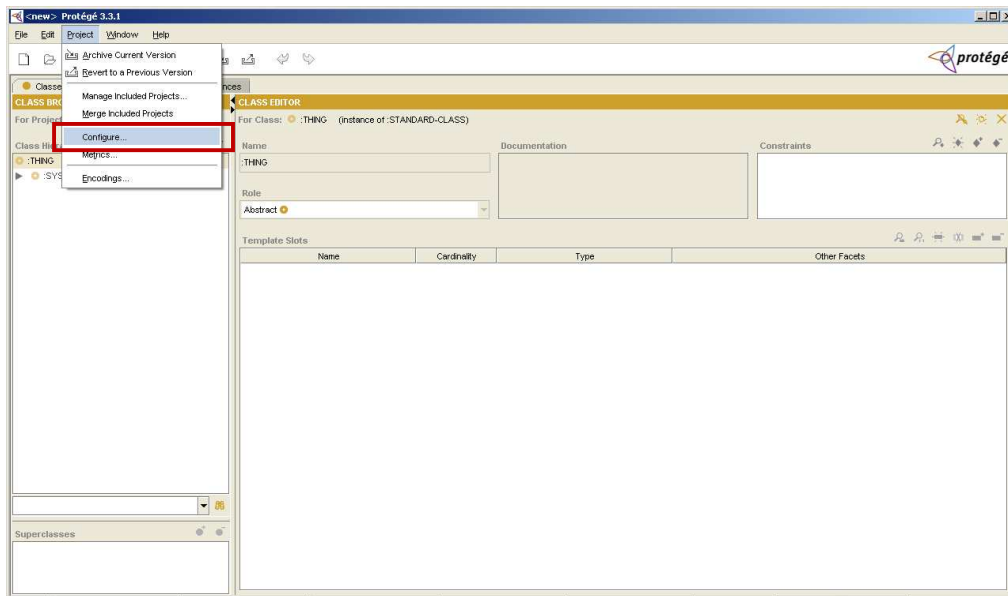


Figura 3. Menu de configuração do projeto.

4. A interface apresentada na figura 4 aparecerá na tela, você deve selecionar “OntoLP” e clicar no botão “ok”.

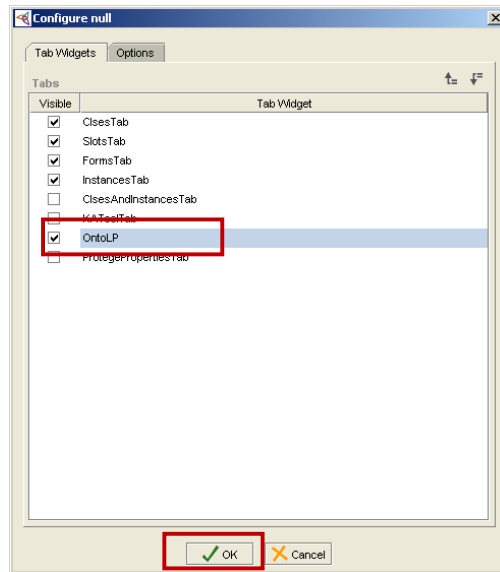


Figura 4. Interface de seleção de plug-ins do Protégé.

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

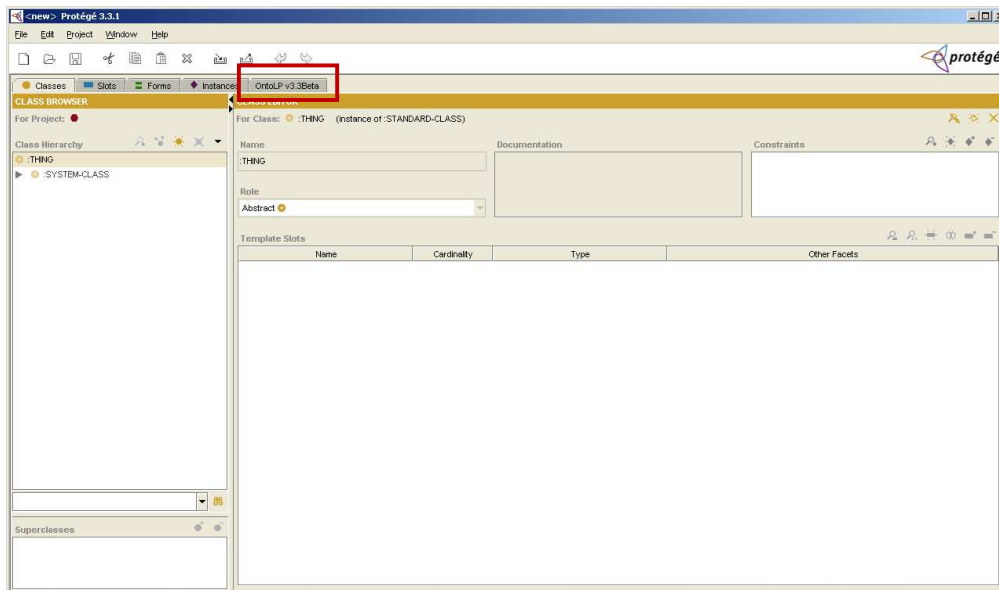


Figura 5. Interface do Protégé com a aba OntoLP.

4-Observação Importante:

Todos os botões do OntoLP apresentam um texto indicando sua função sempre que o mouse é deixado sobre eles (figura 6).

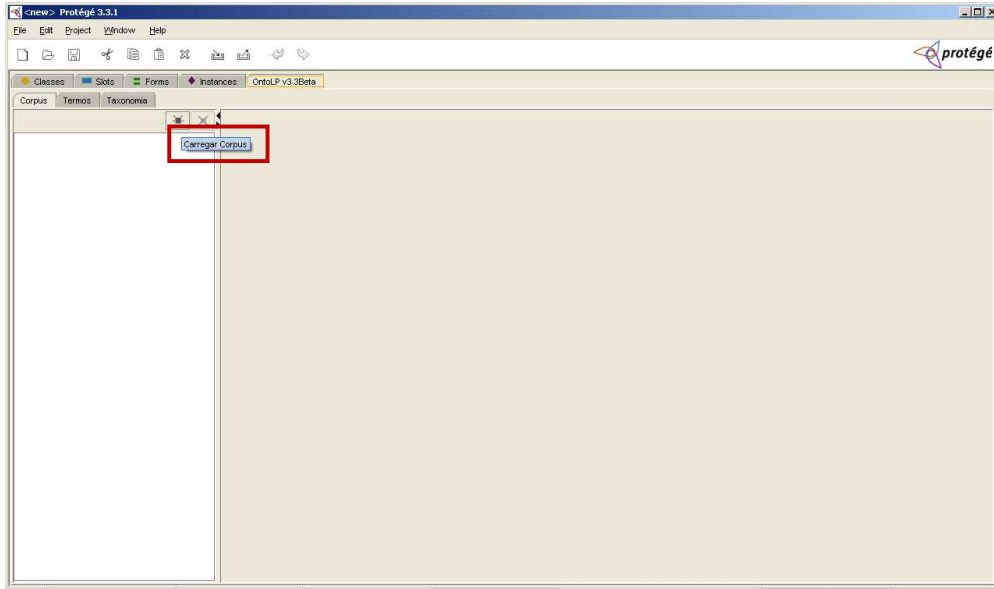


Figura 6. Exemplo de texto indicando a função de um botão.

5-Aba de Carga do Corpus:

5.1-Guia Rápido da Aba de Carga do Corpus (figura 7):

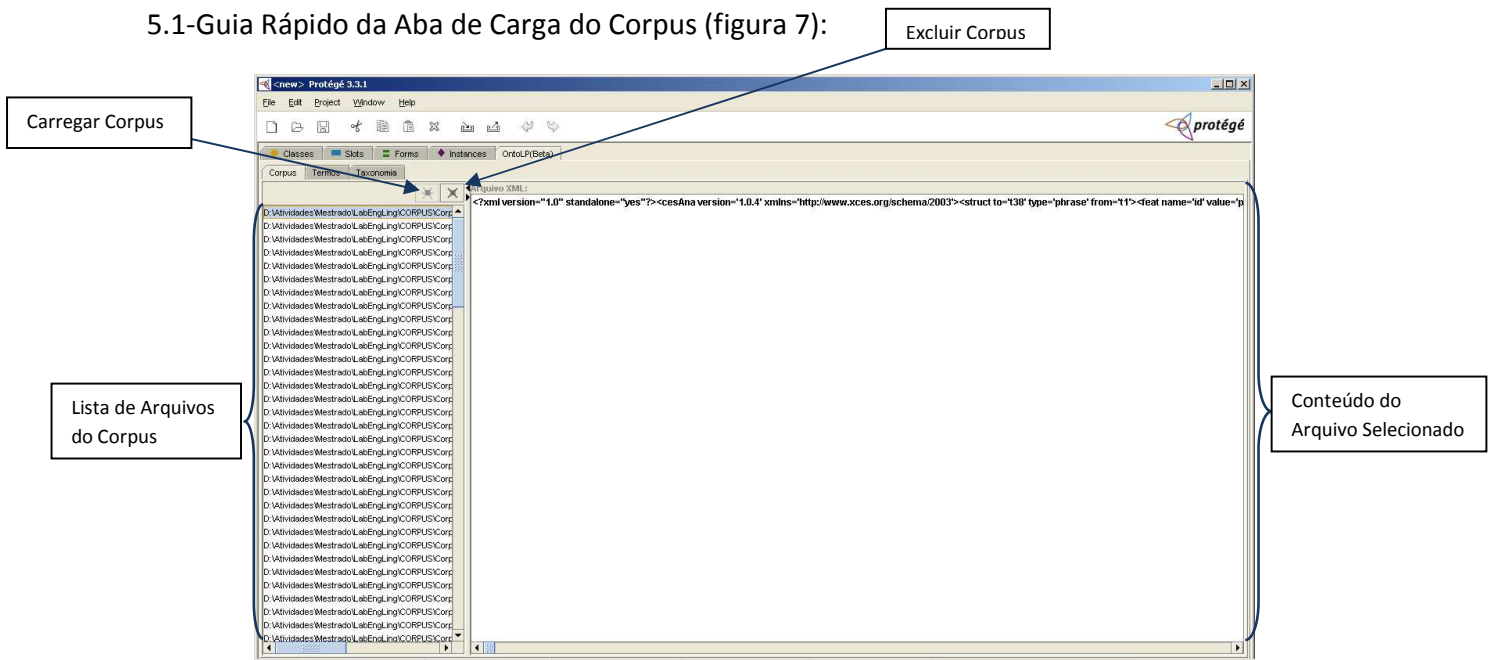


Figura 7. Interface de Carga do Corpus.

5.2-Funcionalidades:

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

1. Clicar sobre a aba “Corpus” (figura 8).

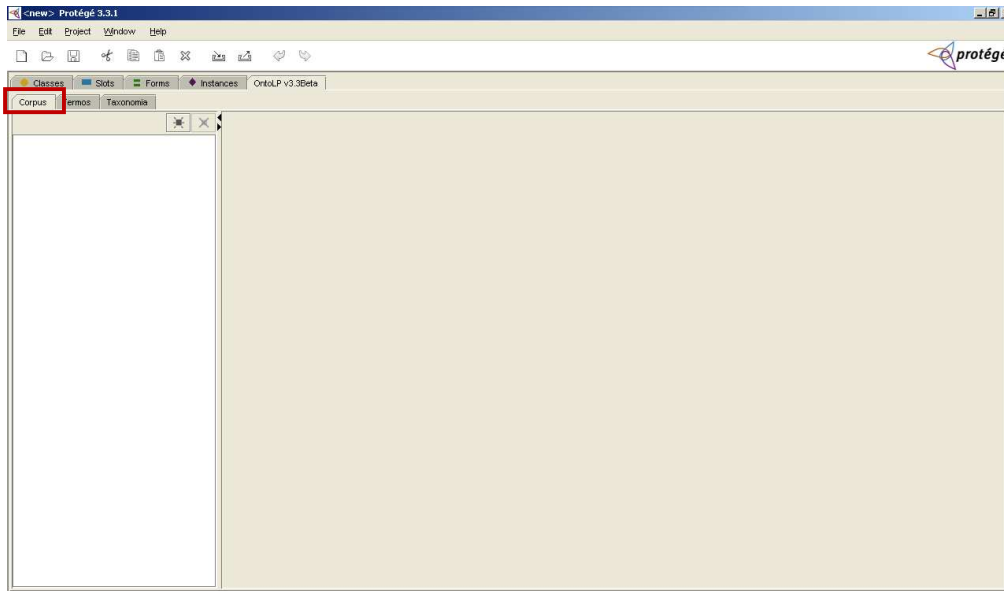


Figura 8: Interface de carga do corpus.

2. Clicar no botão “Carregar Corpus” (figura 9).



Figura 9. Botão para carregar um corpus de domínio.

3. Selecionar a lista de arquivos (ctrl+a) em formato XCES e pressionar “abrir” (figura 10).

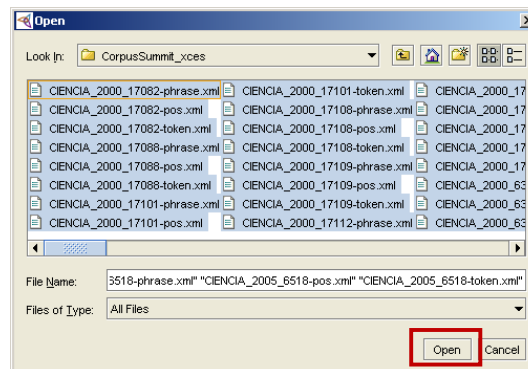


Figura 10. Interface de seleção do corpus.

4. Depois de carregado o corpus, é apresentada uma lista com os arquivos selecionados. Para visualizar o conteúdo de um arquivo basta clicar sobre ele (figura 11).

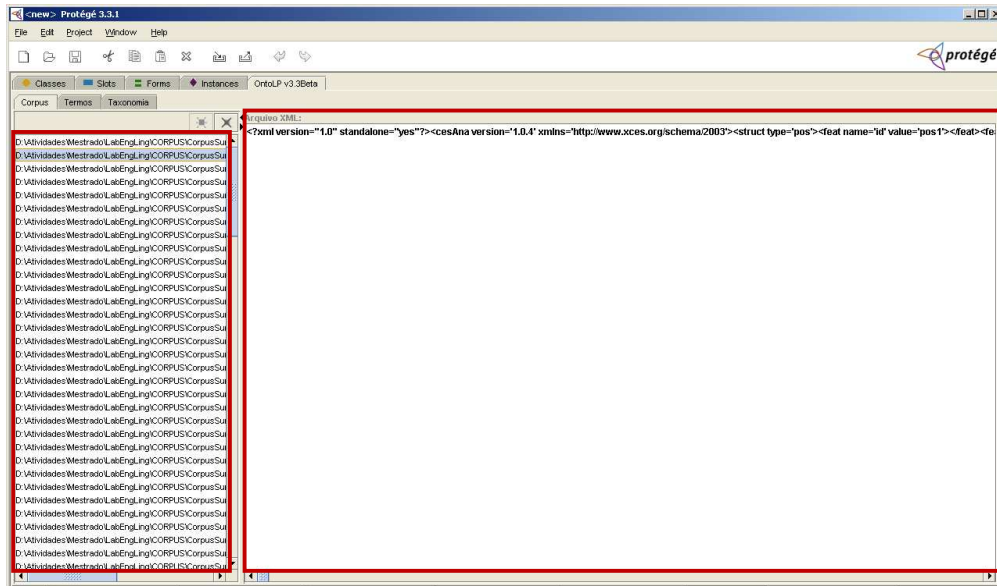


Figura 11. Corpus depois de carregado pelo usuário.

6-Aba de Extração de Termos:

6.1- Guia Rápido da Aba de Extração de Termos:

Executar Método

Excluir Método

Excluir linha da tabela selecionada

Configurar Método

Lista de Filtros Semânticos Executados

Lista de Termos relacionados ao Grupo Semântico

Lista de Grupos Semânt. do Método Selecionado

Semântica	Descrição	Freq. Rel.
[ac]	Abstrato e Contável (alternativa, chance, laz...	0,0685789
[act]	Ação	0,0534550
[Hprof]	Professional human (marinheiro, imples <Ha...	0,0435463
[H]	Group of humans (organisations, teams, co...	0,0346806
[ann]	Massa Abstratação Contável (habilidade, le...	0,0333168
[H]	Human, umbrella tag	0,0320730
[event]	(EVENTO) non-organised event (-CONTRO...	0,0320730
[sem-c]	cognition product (concept, plan, system, co...	0,0315515
[Azo]	Land-animal (raposa)	0,0312907
[inst]	(institution)	0,0271186
[sem-r]	read-work (biografia, dissertação, e-mail, fic...	0,0252934
[Labs]	Abstract place (anverso, auge)	0,0208605
[sick]	disease (sne, AIDS, sida, alcoolismo, cp <...	0,0190352
[act-o]	Realizar uma Ação (tentativa, teste, homena...	0,0189492
[occ]	occasion, human/social event (copa do mun...	0,0166884
[amount]	quantity noun (bocada, teor, sem-fim)	0,0166884
[dur]	duration noun (test: durar+, imples <unit>, e...	0,0164276
[cc]	Objeto Concreto e Contável (geralmente obj...	0,0156454
[per]	(period) and <temp>	0,0148631
[Ltop]	Geographical, natural place (promontório, p...	0,0148631
[activity]	Atividade (correria, manejo)	0,0148631
[domain]	domain (subject matter, profession, cf <gen...	0,0143008
[percep-r]	what you feel (senses or sentiment, pain, e...	0,0139201
[tool]	umbrella tag (abana-moscas, lapis, co...	0,0135593
[Lstar]	Star object (planets, comets: planeta, quasar)	0,0125163
[cm-chem]	chemical substance, also biological (acetien...	0,0125163
[op]		0,0125163
[cm]	concrete mass/non-countable, umbrella tag...	0,0122555
[temp]	temporal object, point in time (amanhecer, n...	0,0119348
[AceH]	Células Animais (bacteria, Células Sanguíne...	0,0119348
[mat]	material, substance (argila, bronze, granito, ...	0,0106910
[Hnat]	Nationality human (brasileiro, alemão), also...	0,0104502

Figura 12. Interface de Extração de Termos (Filtro por Grupo Semântico).

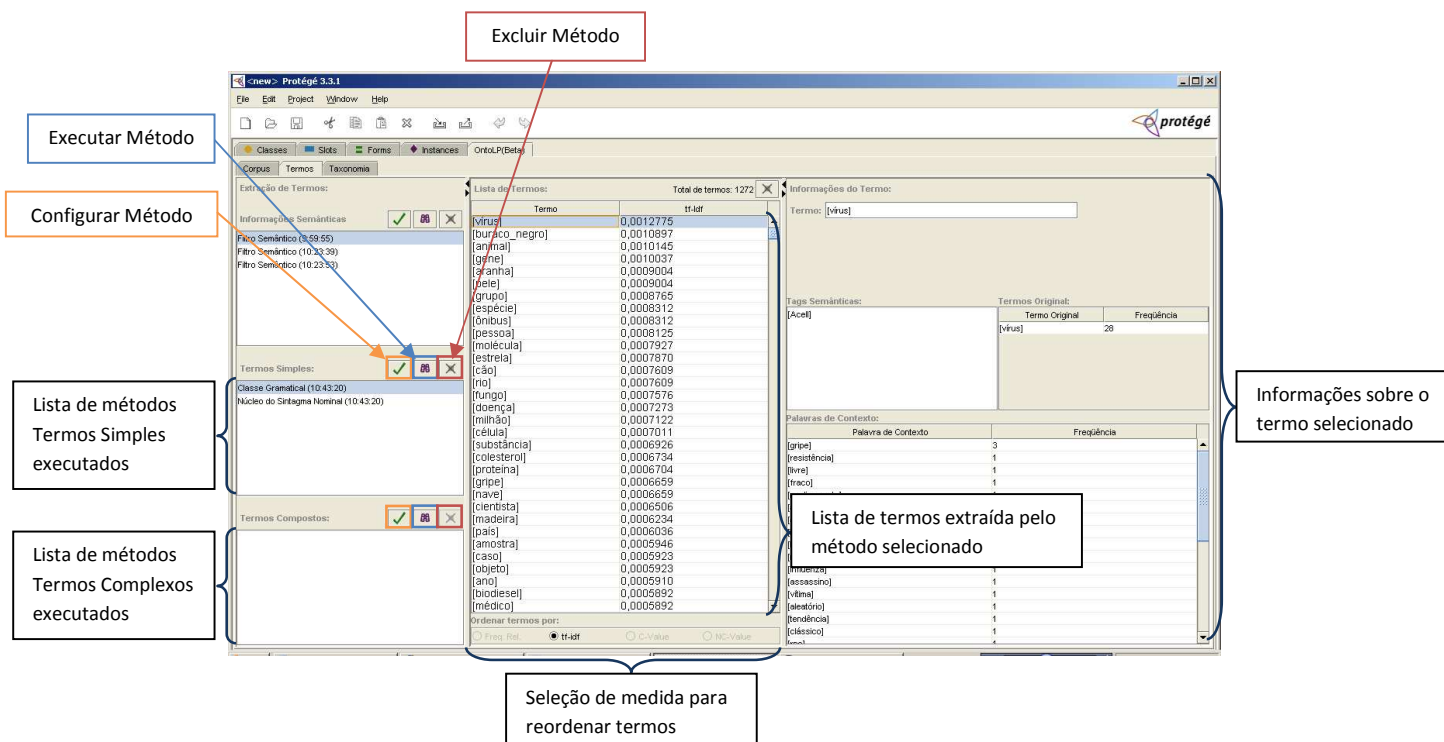


Figura 13. Interface de Extração de Termos (Termos Simples e Termos Complexos).

6.2- Funcionalidades:

A extração de termos no OntoLP pode ser feita de duas maneiras: (Abordagem 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 das abordagens será 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, enquanto, 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 14):

1. Habilitar Filtro Semântico: habilita ou 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.

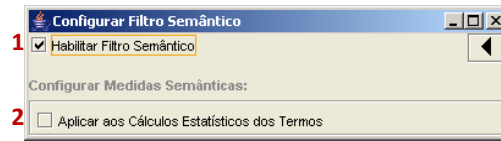


Figura 14. Painel de Configuração do Filtro por Grupos Semânticos.

Quando executado o método extrai os grupos semânticos encontrados no corpus e os apresenta ao usuário em ordem decrescente de relevância (figura 15).

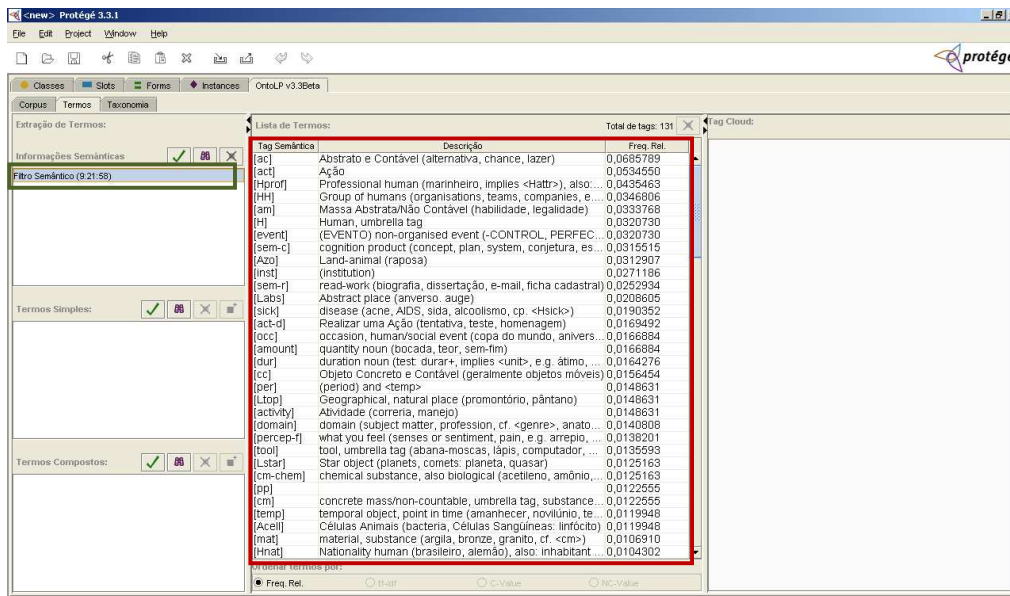


Figura 15. Para visualizar a lista de Grupos Semânticos clique no método depois de executado (destacado em verde). A lista de Grupos Semânticos ordenada pela relevância é destacada em vermelho.

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 16, onde “doença” é o termo mais relevante.

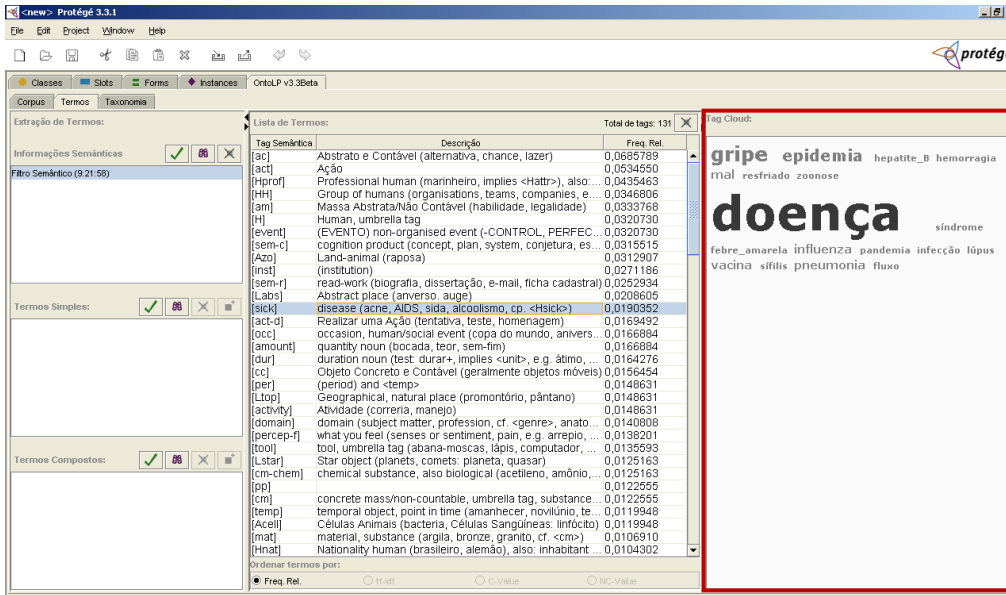


Figura 16. Termos extraídos para o Grupo Semântico [sick], sendo “doença” o termo mais relevante.

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 17. Feito isso, os métodos de extração de termos simples e complexos (etapas posteriores) descartarão os termos pertencentes aos grupos excluídos.

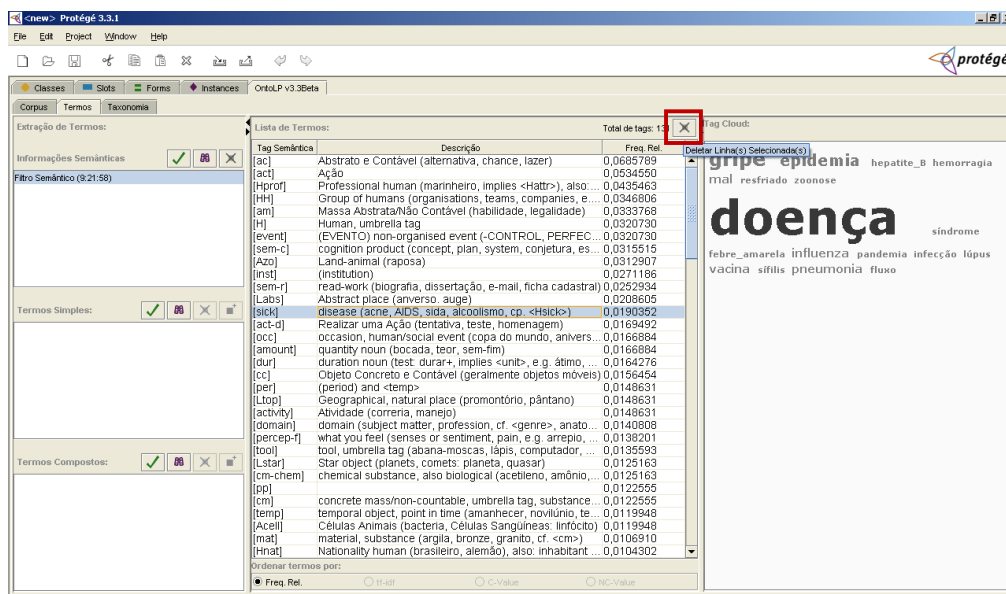


Figura 17. Botão de exclusão das linhas selecionadas na tabela.

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 18).

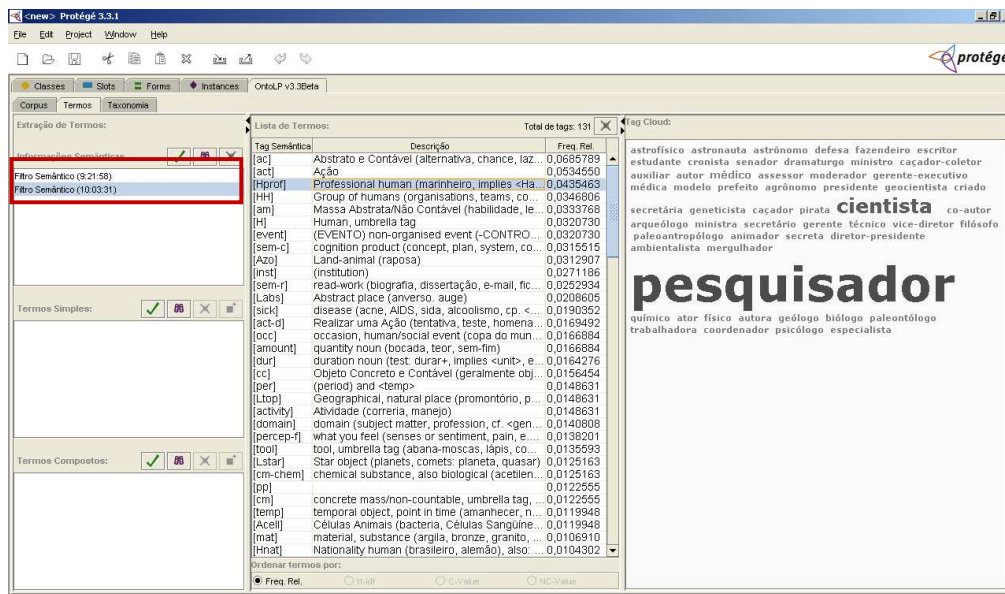


Figura 18. Execuções do método de Grupos Semânticos e seleção de uma lista de Grupos Semânticos.

A etapa seguinte visa extrair termos simples (uni gramas) através de dois diferentes métodos. Esses métodos possuem configurações distintas. Para carregar o painel de configuração dos métodos pressione o botão destacado na figura 19. É apresentada então uma janela com duas abas, uma para cada método (figura 20), com as seguintes opções:

- Método Classe Gramatical
 1. Habilitar método: habilita ou 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 ou 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.

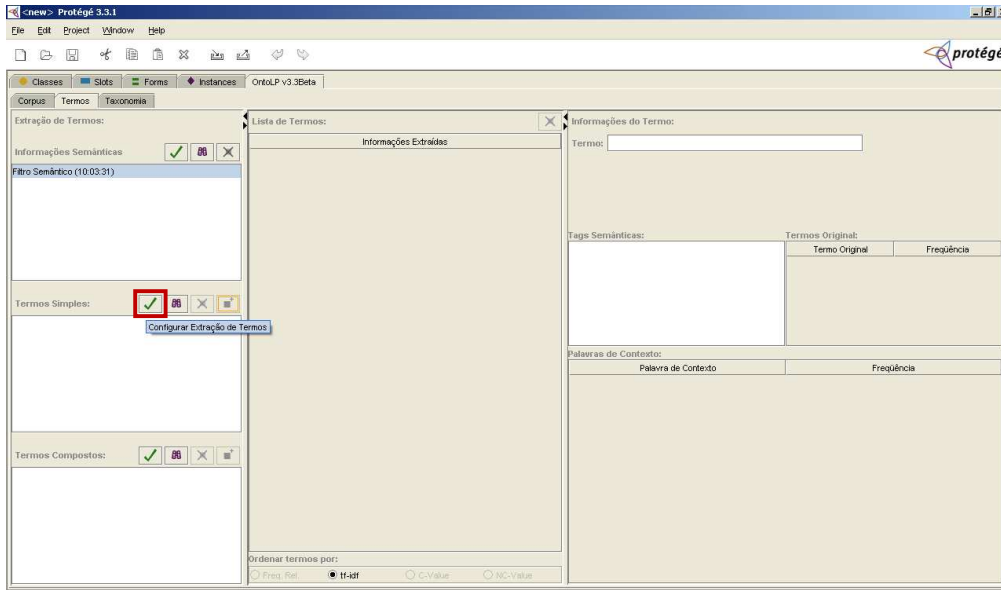


Figura 19. Botão de configuração dos métodos de extração de termos simples.

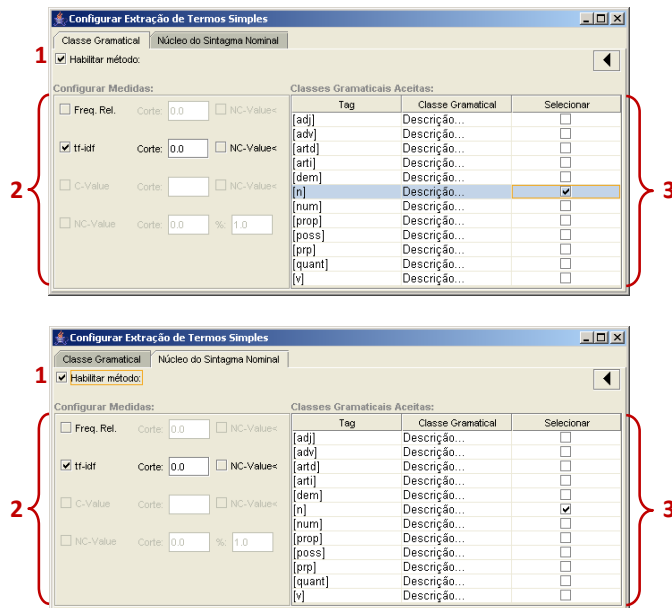


Figura 20. Interface de configuração dos métodos de extração de termos simples.

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 21) e editada pelo usuário, excluindo os termos irrelevantes (figura 17).

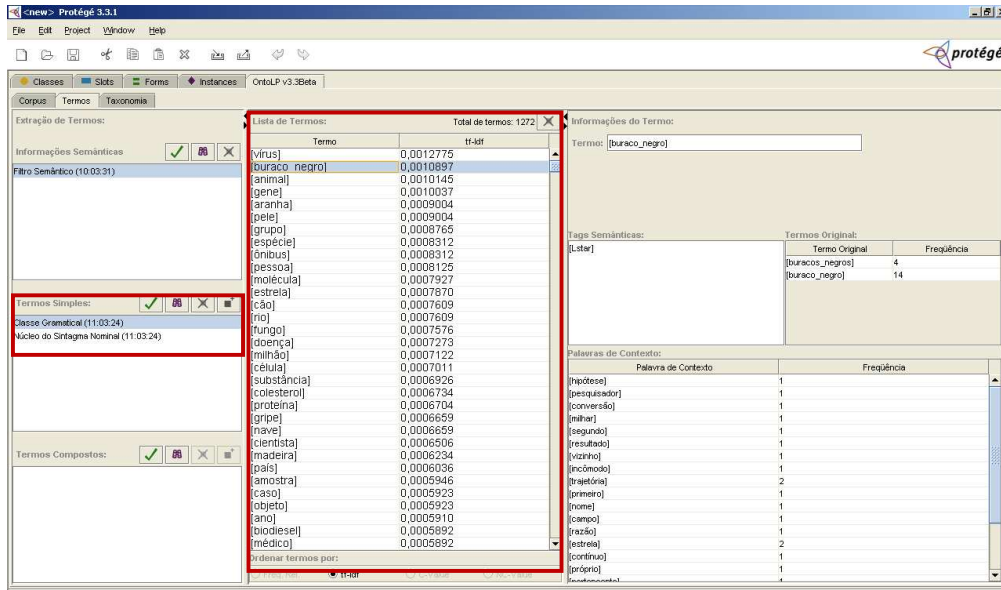


Figura 21. Seleção do método e lista de termos relacionada a ele.

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

3. A qual Grupo Semântico o termo está relacionado (figura 22).

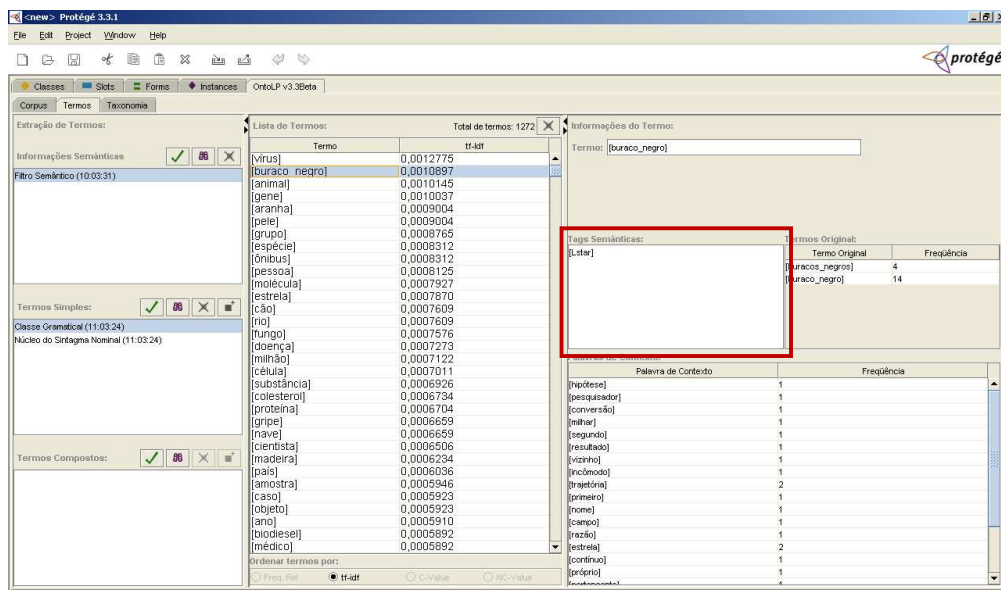


Figura 22. Grupo Semântico ao qual o termo está relacionado.

4. Forma original dos termos no corpus e a freqüência de cada uma delas (figura 23).

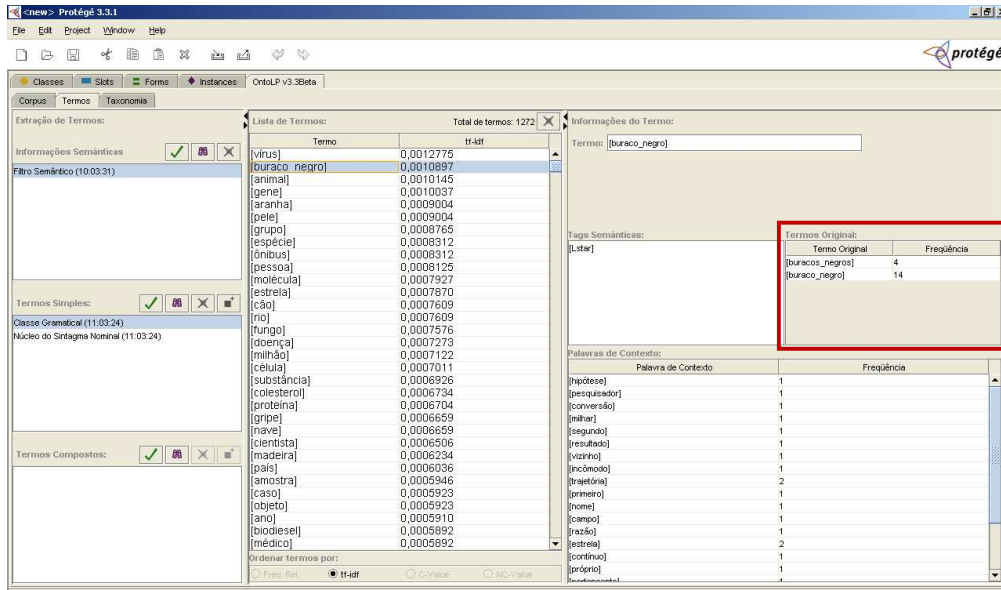


Figura 23. Forma original que o termo apareceu no corpus.

5. Palavras que aparecem na mesma sentença do termo e a frequência com que se repetem (figura 24).

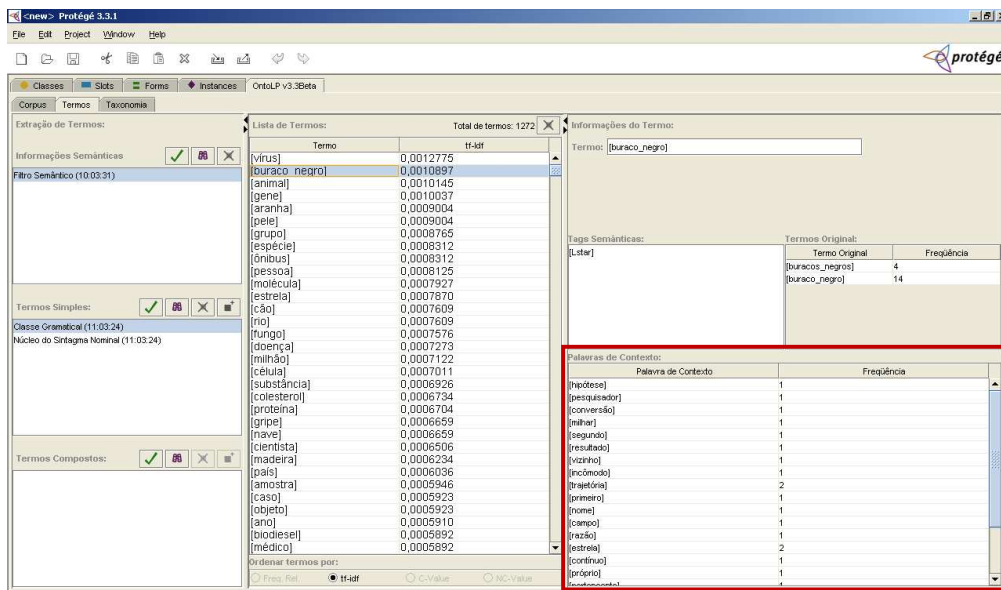


Figura 24. Lista de palavras que aparecem próximas ao termo.

A última etapa, extração de termos complexos, assim como a anterior, possui um conjunto de opções de configuração para cada método. Nas figuras 25, 26 e 27 são apresentadas as interfaces de configuração, as opções são explicadas a seguir:

- N-Gram

1. Habilitar método: habilita ou 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.

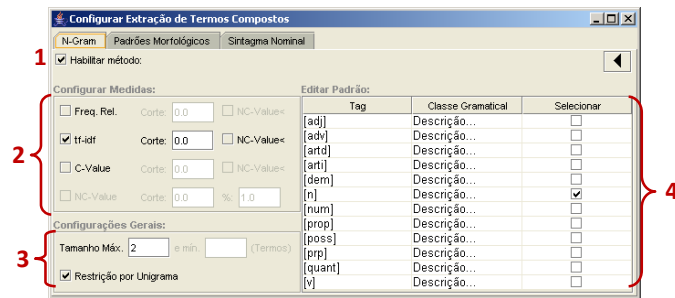


Figura 25. Interface de configuração do método N-Gram.

- Padrões Morfológicos

1. Habilitar método: habilita ou 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.

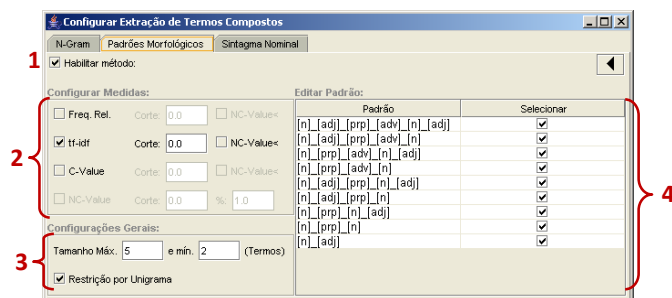


Figura 26. Interface de configuração do método Padrões Morfológicos.

- Sintagma Nominal

1. Habilitar método: habilita ou 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.

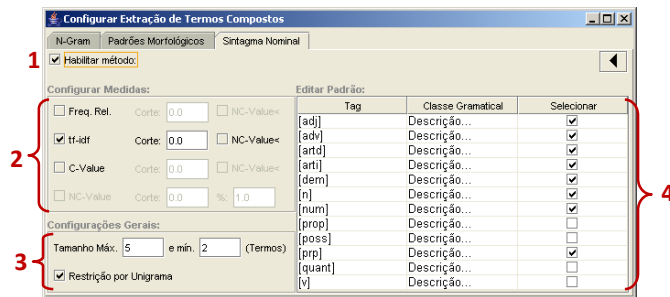


Figura 27. Interface de configuração do método Sintagma Nominal.

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 28. 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}).

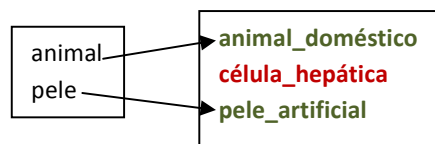


Figura 28. Exemplo de restrição por unigrama.

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.

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

7.1-Guia Rápido:

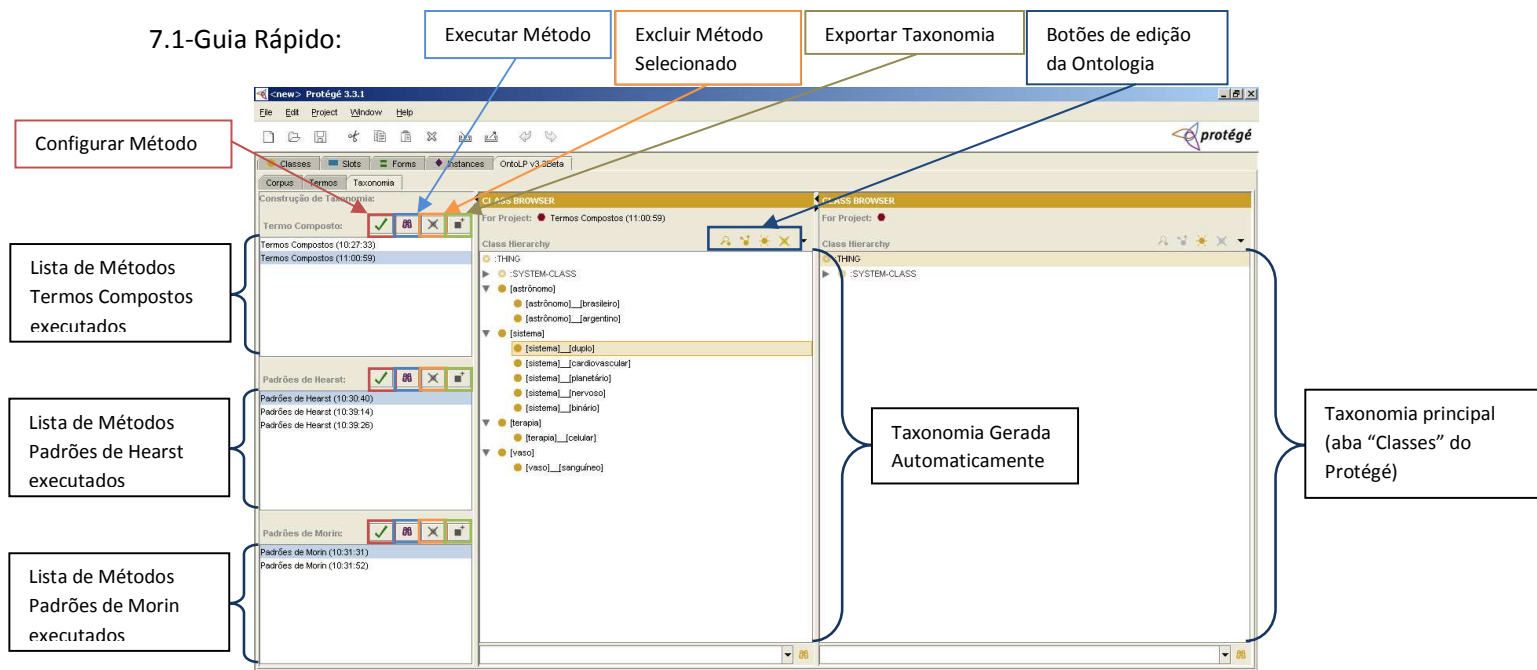


Figura 29. Aba de Organização Hierárquica dos Termos.

7.2-Organização Hierárquica dos Termos:

A aba de organização hierárquica de termos disponibiliza três métodos para o usuário: (1) Termos Compostos; (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 complexos ([**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 30 são apresentados alguns exemplos de organização hierárquica com base em termos compostos.

- ▼ ● [sistema]
 - [sistema]_[cardiovascular]
 - [sistema]_[nervoso]
- ▼ ● [terapia]
 - [terapia]_[celular]
- ▼ ● [vaso]
 - [vaso]_[sanguineo]

Figura 30. Exemplo de relação hierárquica extraída pelo método Termos Compostos.

Na figura 31 é apresentada a interface de configuração do método e suas opções são descritas abaixo:

1. Habilitar método: habilita ou desabilita o método.

2. Início 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**”.

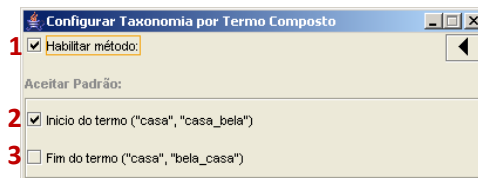


Figura 31. Painel de configuração do método Termos Complexos.

Caso o usuário deixe as opções 2 e 3 desabilitadas, o método selecionará ambos os 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 33 e explicadas abaixo:

1. Habilitar método: habilitar ou 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 32 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.

▼ ● [neurônio] <an>{anatomia}
 ● [célula_]nervoso] Acell(células animais), <an>{anatomia}

Figura 32. Painel de configuração do método Termos Complexos.

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.

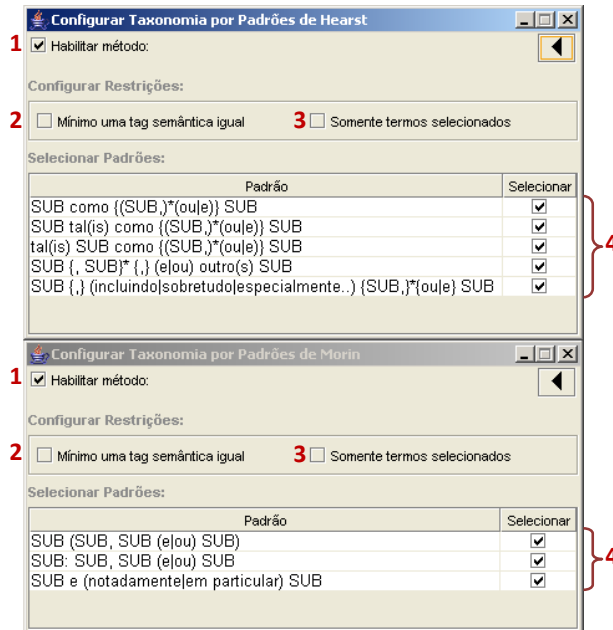


Figura 33. Janelas de configuração dos padrões de Hearst e Morin.

7-Conclusão:

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