O desenho do novo Folheador

Relatório Técnico
31 de Dezembro de 2011

Hernani Costa
Linguateca, FCCN
hpcosta@dei.uc.pt

Resumo

Neste relatório é apresentado o novo Folheador, um serviço na rede que permite navegar e explorar relações semânticas da língua portuguesa. Além de facilitar a pesquisa em diferentes bases de conhecimento lexical para o português, uma característica importante do sistema é a sua ligação com uma interface de corpos que permite a exploração de palavras relacionadas em contextos reais. Uma outra característica do sistema é a apresentação de diferentes valores de confiança, calculados previamente, para cada relação semântica. Por fim, o sistema engloba também uma interface gráfica que permite explorar visualmente todas as relações semânticas de uma palavra pré-selecionada no Folheador.

1 Introdução

Neste relatório é apresentado o novo Folheador[1] (figura 1), um serviço que surge da necessidade da criação de uma interface na rede capaz de manusear e apresentar relações semânticas para os recursos públicos disponibilizados pela Linguateca[2].

De referir que a aparência do Folheador foi inspirada no antigo Folheador, criado por Hugo Gonçalo Oliveira para o PAPEL [3] (Gonçalo Oliveira, Santos e Gomes, 2010). De modo similar ao antigo Folheador, esta nova versão foi, primeiramente, pensada para visualizar as relações semânticas presentes no recurso referido anteriormente. Contudo, ao longo do seu desenvolvimento, os requisitos mudaram e foi necessário repensar toda a estrutura inicial de modo a possibilitar a inclusão de novos recursos. Foi através desta necessidade que o *Núcleo*[4] foi alvo de várias reestruturações durante o seu desenvolvimento, contribuindo por sua vez para um sistema mais otimizado e robusto.

Relativamente à interface, esta foi desenhada para facilitar a pesquisa nas várias bases de conhecimento lexical para o português, assim como a ligação a uma interface de corpos, permitindo deste modo a exploração das palavras relacionadas em contextos reais. Por fim, surgiu também a necessidade de apresentar a informação disponível de um modo mais visual. Deste modo, foi criada uma interface gráfica, que permite a visualização de todas as relações semânticas de uma palavra, selecionada no Folheador, num grafo.

Um outro requisito inicial passava pela atribuição de um valor de confiança a cada relação semântica. Foram criados dois tipos diferentes de valores de confiança: um valor simplesmente calculado tendo em conta a co-ocorrência dos termos da relação nos corpos disponibilizados pela Linguateca; e outro considerando o número de recursos em que a relação semântica é contemplada, ou seja o número de contextos diferentes de que a relação foi extraída.

Depois de apresentadas as principais características do sistema, o restante relatório contém mais cinco secções: a secção 2 que, de uma forma geral, apresenta a arquitetura do sistema; em seguida as secções 3 e 4 descrevem em pormenor como foi abordada a implementação da *Interface*[5] e do *Núcleo*, respetivamente; e por fim, antes de serem apresentadas as novas perpetivas de trabalho futuro na secção 6, são ainda apresentadas algumas observações sobre todo o trabalho desenvolvido, no âmbito do projeto, na secção 5.

[1] `http://linguateca.pt/Folheador`

[2] `http://www.linguateca.pt`

[3] `http://linguateca.pt/PAPEL`

[4] *Núcleo (Back Office)* é, tradicionalmente, o nome a que se chama o resto do sistema, que inclui o conteúdo, a que o Folheador dá acesso.

[5] Entenda-se *Interface*, como sendo a parte do sistema relacionada com a interação com o utilizador, tradicionalmente chamado em inglês de *Front Office*.

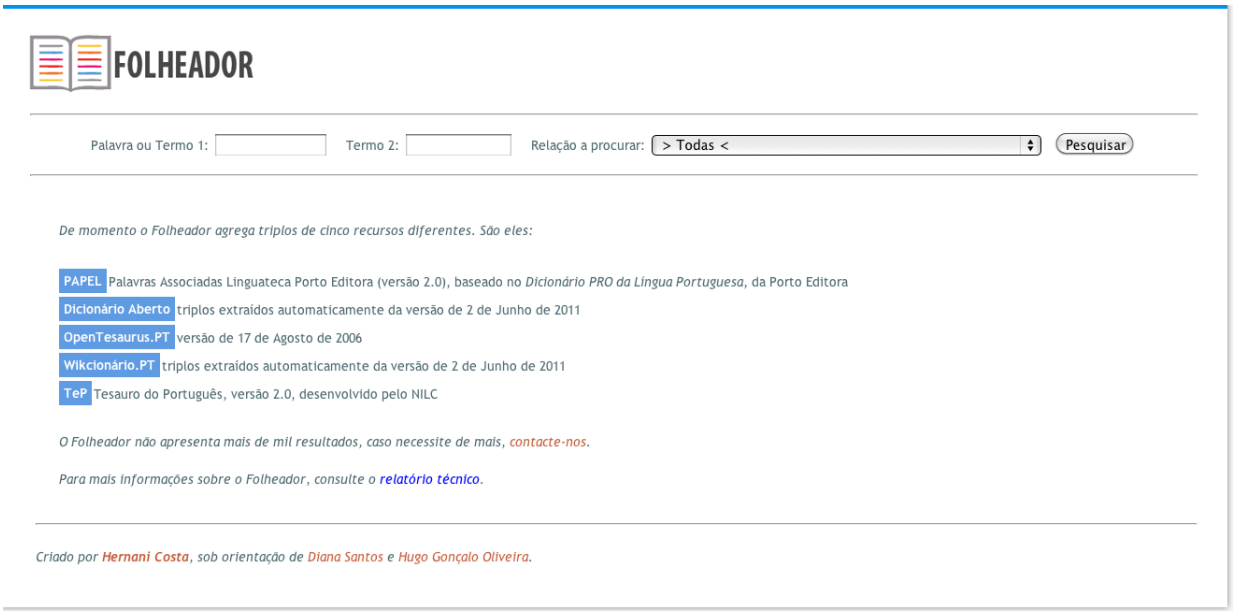

Figura 1: Página inicial do Folheador.

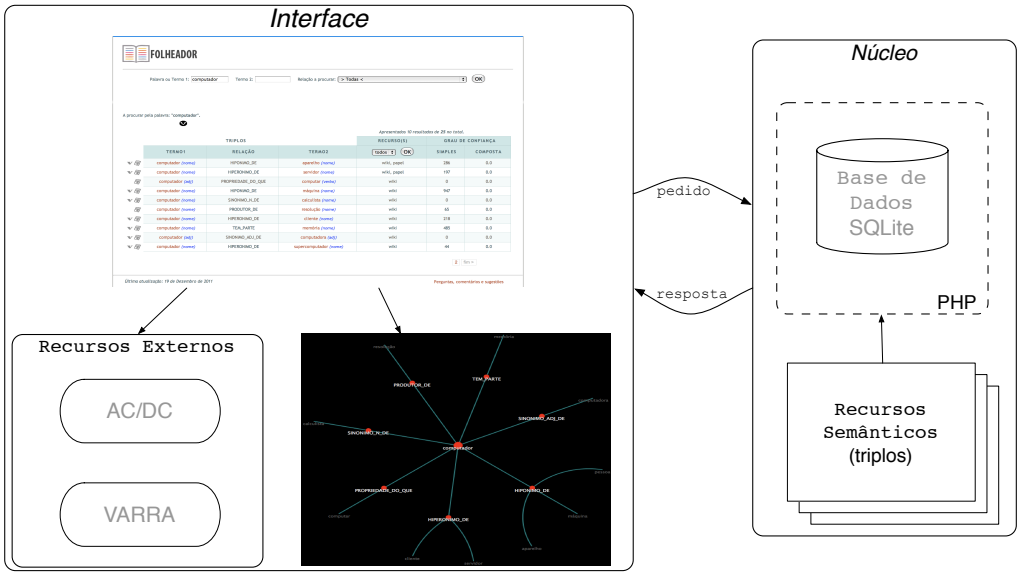

Figura 2: Arquitetura do sistema.

2 *Arquitetura do sistema*

A arquitetura inerente ao Folheador é apresentada na figura 2. Como se pode verificar, esta divide-se em duas componentes, a *Interface* e o *Núcleo*. A *Interface* engloba todos os componentes necessários para interagir com o utilizador. Mais especificamente, contém a própria interface do Folheador que, por sua vez, possibilita a invocação de `recursos externos` (especificamente, o `AC/DC` [6] e o `VARRA` [7]) e do visualizador gráfico. Por sua vez, o *Núcleo* disponibiliza, à *Interface*, a `base de dados` que contém toda a informação, ou seja, relações semânticas, valores de confiança, etc. É também no *Núcleo* que estão todas as linguagens dedicadas que foram usadas na criação da `base de dados`.

O único requisito do sistema é ter um servidor Apache[8] instalado para que este interprete a linguagem PHP. A escolha do PHP, como linguagem de programação principal deste projeto, deve-se à sua portabilidade[9], livre disponibilização e por ser uma linguagem que é interpretada do lado do servidor, permitindo

[6] `http://www.linguateca.pt/ACDC`
[7] `http://www.linguateca.pt/VARRA`
[8] `http://httpd.apache.org`
[9] Não está dependente da plataforma onde é executado.

deste modo não sobrecarregar o cliente.

Em seguida apresenta-se detalhadamente a estrutura da *Interface* (secção 3) e como pode ser efetuada a manutenção do *Núcleo* (secção 4).

3 A Interface do Folheador

Como se sabe, a interface de um sistema informático é toda a vertente de software, neste caso páginas na rede, que está direcionada ao utilizador. Contudo, por detrás da interface, visível ao utilizador, existe um conjunto de programas que estão a correr para que esta funcione, veja-se a figura 3.

De modo a ser mais percetível, são apresentados em detalhe os três módulos que a *Interface* engloba (Módulo: folheador, visualizador e recursos externos), assim como uma pequena descrição dos mesmos:

a) folheador

   i) `index.php` → *página principal*

   ii) `funcoes_indice.php` → *funções auxiliares*

   iii) `style.css` → *folha de estilos*

b) visualizador

   i) `visualizador.php` → *página principal*

   ii) `jit.js` → *biblioteca JIT*[10]

   iii) `visualizador.js` → *modelo JavaScript*[11]

   iv) `base.css` → *folha de estilos*

c) recursos externos

   i) `reencaminhaACDC.php` → *cria query*

   ii) `chamaACDC.php` → *envia pedido com a query, criada em i), ao AC/DC*

   iii) `reencaminhaVARRA.php` → *filtra as relações que não estão presentes no VARRA*

   iv) `chamaVARRA.php` → *envia pedido ao VARRA*

Módulo folheador

O módulo `folheador` abrange todos os programas responsáveis por apresentar a informação ao utilizador.

De modo a respeitar as boas práticas de programação, o formato (ficheiro `style.css`) e o conteúdo (ficheiro `index.php` e `funcoes_indice.php`) do módulo `folheador` foram separados. Mais especificamente, é na página principal, ficheiro `index.php`, onde é apresentada toda a informação visível ao utilizador. Contudo, o `index.php` só contém a estrutura lógica do Folheador, ou seja todas as funções necessárias para que as interações do utilizador sejam realizadas com sucesso estão descritas no ficheiro `funcoes_indice.php` (contém todas as funções auxiliares usadas repetidamente pelo `index.php`). Em adição, e como referido anteriormente, foi criada uma folha de estilos (ficheiro `style.css`) para formatar a informação apresentada na página principal.

Como podemos ver na figura 3, existe uma ligação à `base de dados` e outra ao ficheiro `log`. Embora se considere que a `base de dados` faz parte do *Núcleo*, esta é apresentada na figura de modo a que se compreenda que o módulo `folheador` vai buscar toda a informação à mesma. Por fim, temos o ficheiro `log` que, como o próprio nome indica, é onde todo o histórico de pesquisas e interações com a *Interface* é guardado. Além de permitir perceber quais as pesquisas mais frequentes, futuramente a informação guardada neste ficheiro possibilitará, a análise do desempenho do sistema e o estudo de futuras melhorias.

Módulo visualizador

O módulo `visualizador` engloba quatro ficheiros (ver figura 3).

O `visualizador.php`, página principal da interface gráfica, recebe um objeto em JavaScript Object Notation (JSON[12]), da página principal do Folheador (`index.php`). Em seguida, através de HTTP Request (do tipo GET), o modelo em JavaScript `visualizador.js` encarrega-se de transformar este objeto num grafo. Este processo é realizado usando os métodos disponibilizados pela biblioteca JIT[13] (`jit.js`).

Por fim, e à semelhança do módulo `folheador`, foi criada uma folha de estilos para formatar a disposição da interface (ficheiro `base.css`).

Na figura 9 é apresentado o aspeto gráfico que este módulo apresenta.

[10] `http://thejit.org/downloads/Jit-2.0.0b.zip`

[11] `http://thejit.org/static/v20/Jit/Examples/Hypertree/example1.html`

[12] `http://www.json.org`

[13] Acrónimo para JavaScript InfoVis Toolkit.

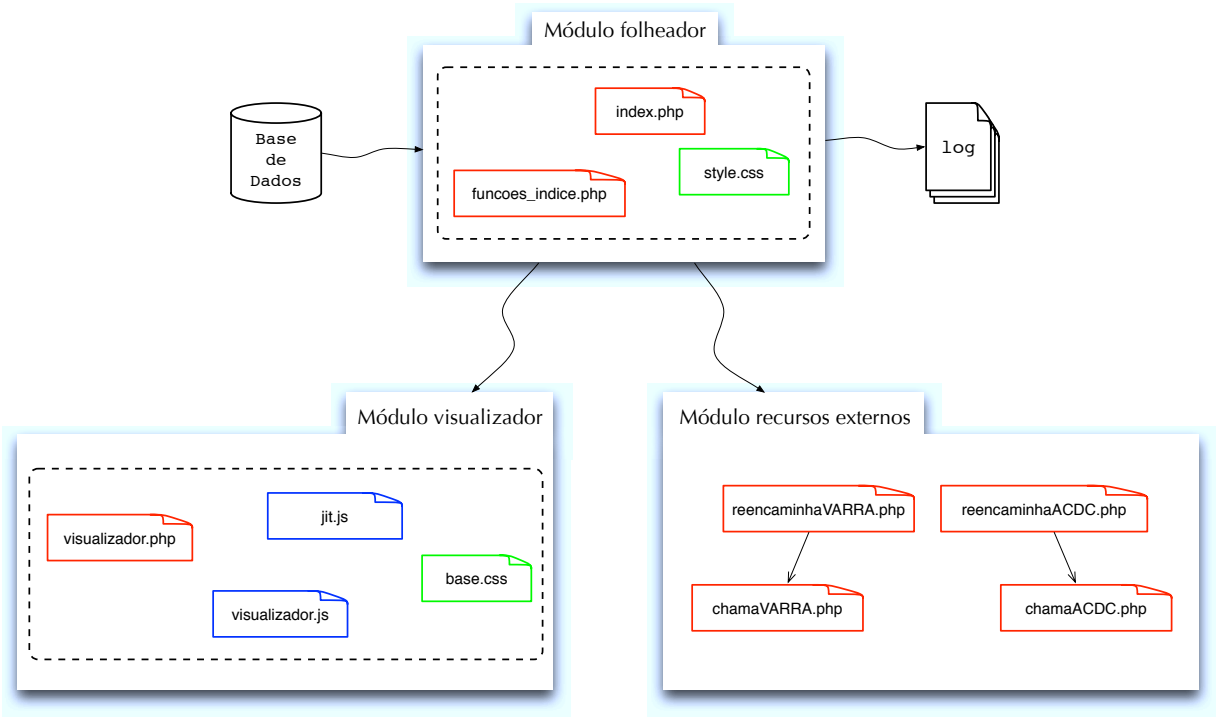

Figura 3: Arquitetura da *Interface* do sistema.

Módulo recursos externos

O terceiro módulo agrega os chamados `recursos externos`, mais especificamente o AD/DC (Santos e Bick, 2000; Santos, 2011) e o VARRA (Freitas et al., 2010).

Em seguida são apresentadas as soluções encontradas para invocar, tanto o AC/DC como o VARRA:

* no caso do AC/DC o ficheiro `reencaminhaACDC.php` serve para criar a expressão de pesquisa: `'MU (meet [lema="' .$termo1. '"] [lema="' .$termo2 .'"] s)'`, onde `$termo1` e `$termo2` correspondem aos dois termos de uma determinada relação semântica. Em seguida essa expressão é enviada para o ficheiro `chamaACDC.php`. Este submete de uma forma automática um formulário com os dados necessários para invocar o programa em Perl "`/cgi-bin/acesso.pl`" que se encontra no servidor da Linguateca;

* no caso do VARRA, o ficheiro `reencaminhaVARRA.php` filtra somente as relações que estão contempladas no VARRA e o ficheiro `chamaVARRA.php` submete, de uma forma automática, um formulário com os dados necessários para invocar o programa em Perl "`/cgi-bin/varra.pl`", que também se encontra no servidor da Linguateca.

3.1 Pesquisa e Navegação

O Folheador permite pesquisar de quatro modos diferentes:

a) **por um termo**[14]: `Palavra ou Termo 1`, ver figura 4. Ao mesmo tempo é possível selecionar a `Relação a procurar` (ver figura 5);

b) **por dois termos:** `Palavra ou Termo 1` e `Termo 2` (figura 4). Ao mesmo tempo é possível selecionar a `Relação a procurar` (figura 5);

c) **selecionando um determinado termo nos resultados apresentados**: todos os termos resultantes da pesquisa têm um endereço associado. Ao clicar num determinado termo, o Folheador tem o mesmo comportamento que em a);

d) **apresentar todas as relações**: para apresentar todos os resultados, as caixas de pesquisas têm de estar em branco, só deste modo é apresentada toda a informação, de uma forma aleatória, contida na base de dados do Folheador.

[14] Um termo pode conter mais que uma palavra, por exemplo "*correio electrónico*".

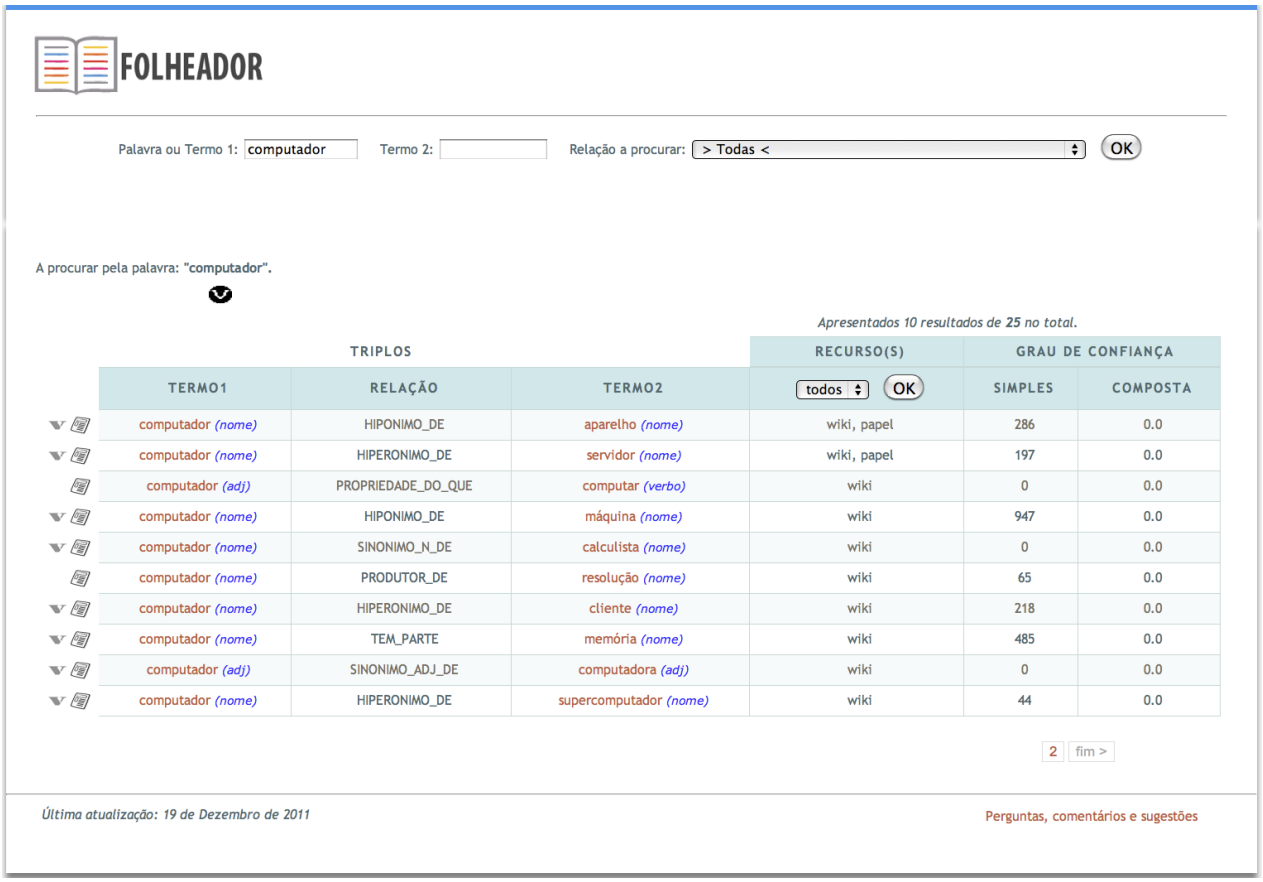

Figura 4: Interface de pesquisa do Folheador.

Além da pesquisa, o Folheador permite filtrar os resultados. Para isso estão disponíveis as seguintes opções:

1. **filtrar por relação**: esta opção (ver figura 5 e tabela 1) permite filtrar os resultados por:
   1.1 relação semântica (por exemplo, `PARTE` e `LOCAL`);
   1.2 sub-relação semântica (por exemplo, `PARTE_DE`, `LOCAL_ORIGEM_DE`);
   1.3 apresentar todos as relações (ou seja, opção "> `Todas` <", figura 5 e tabela 1).
2. **filtrar por recurso**: esta opção só é apresentada caso o Folheador devolva informação para uma determinada pesquisa. Neste menu, somente são apresentados os recursos ou recurso em que existe o termo ou termos que são pesquisadas (veja-se a figura 6).

De modo a não sobrecarregar o sistema, os resultados das pesquisas são apresentados em várias páginas, em que cada uma mostra 20 resultados (na figura 4 somente são apresentados 10 resultados para exemplificar). Para alterar este valor, basta alterar a variável `$linhas_por_pagina`, no ficheiro `index.php`. A navegação nas páginas é conseguida através do menu, apresentado no canto inferior direito da tabela de resultados, que contém a numeração das páginas.

De referir também que os resultados acessíveis ao utilizador estão limitados. Para alterar o valor, que por defeito é 1000, é necessário alterar a variável `$MAX_RESULTADOS` no ficheiro `index.php`.

3.2 Grau de Confiança dos Dados

Com o objetivo de associar um grau de confiança aos triplos extraídos dos vários recursos foram pensados dois tipos de valores de confiança: `simples` e `composto`. O valor `simples` resulta do número de vezes que os termos de cada triplo co-ocorrem nos corpos disponibilizados pelo AC/DC. Para a criação do mesmo, foi criado um ficheiro em PHP para invocar o CQP (Evert, 2005), instalado no servidor da Linguateca. O programa percorre todos os triplos, que à priori estão na base de dados, e procura no AC/DC o número de co-ocorrências dos seus termos em todas as frases dos corpos. Em seguida esses valores são guardados na base de dados. Para

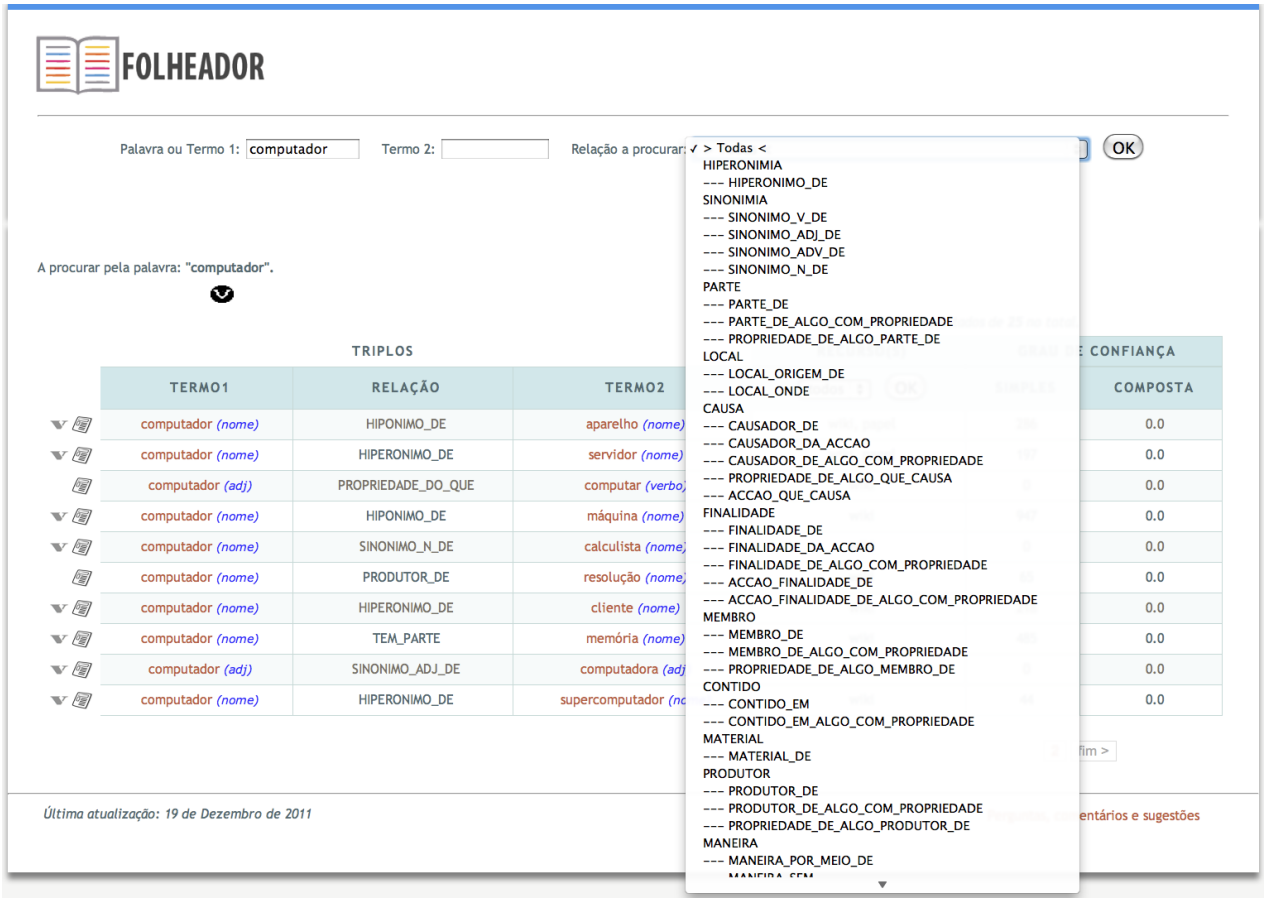

Figura 5: Relações subjacentes ao Folheador.

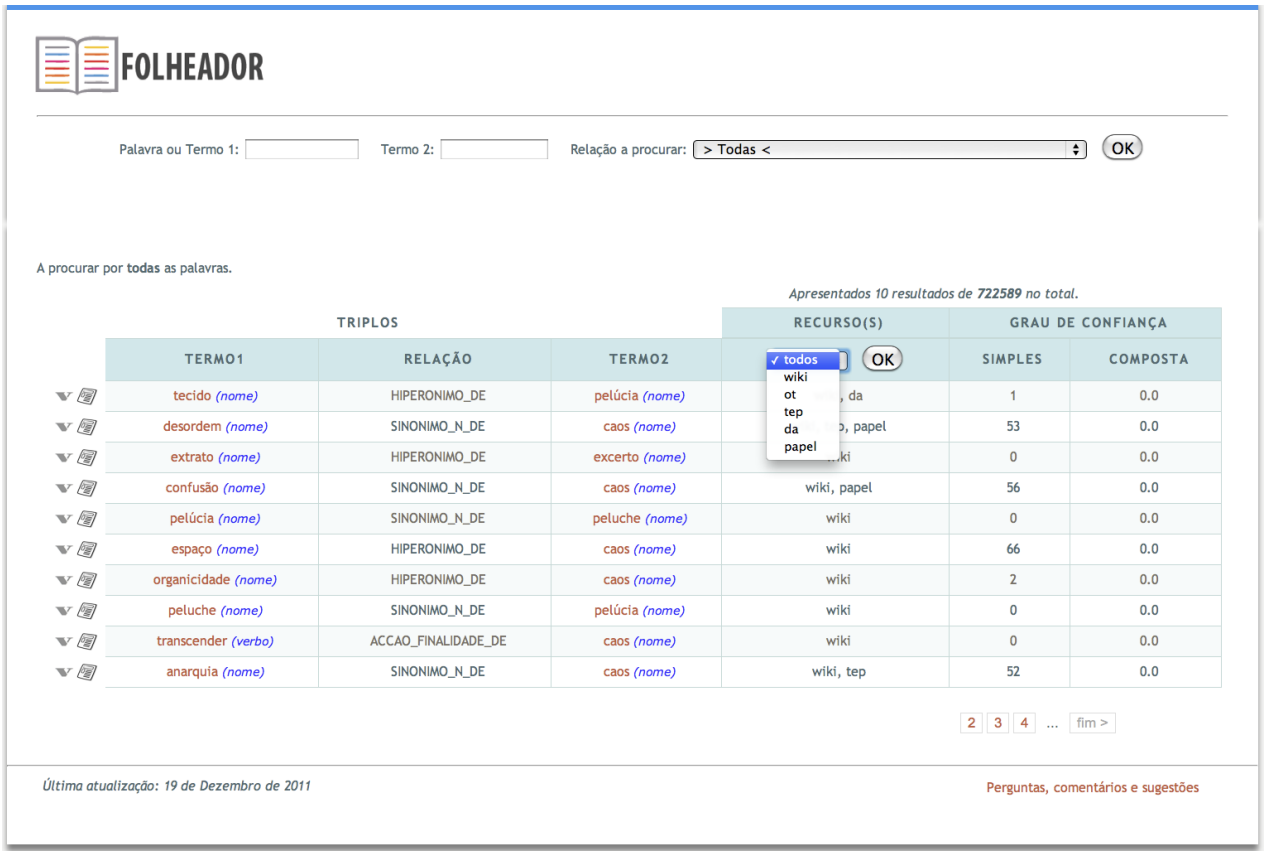

Figura 6: Recursos subjacentes ao Folheador.

mais detalhes sobre o modo que o valor `simples` foi implementado veja-se a secção 4.

Foi também pensado um valor `composto`. Este resultaria do número de co-ocorrências dos termos, conectados através de padrões léxico-sintáticos que denotassem a relação semântica dos triplos (pensámos usar os mesmos padrões [15] usados no projeto VARRA). No entanto, a sua implementação não foi possível devido ao curto espaço de tempo do projeto.

Contudo, se considerarmos que um determinado triplo foi extraído em mais que um recurso, poderemos inferir que este tem uma maior probabilidade de estar correto do que outro triplo que só foi extraído de um recurso (ver coluna 'RECURSO(S)' da figura 4). Deste modo, pode-se interpretar este resultado como um outro valor de confiança associado ao triplo.

3.3 Ligações a outros Recursos

Neste momento, a interface disponibiliza ligação a dois recursos da Linguateca, nomeadamente ao AC/DC e ao VARRA (ver figura 2).

Sendo o AC/DC um serviço na rede que fornece acesso a um grande conjunto de corpos da língua portuguesa, a sua inclusão foi imperativa. No interface do Folheador, em apenas um clique, é possível a consulta de todas as frases nos corpos do AC/DC que contêm os dois argumentos de um triplo (ver figura 4, ícone ). A figura 7 apresenta alguns resultados para as palavras "`computador`" e "`aparelho`" no contexto do AC/DC.

Um outro serviço disponibilizado pela Linguateca, que também é possível invocar através do interface, é o VARRA. Este serviço usa um conjunto de padrões léxico-sintácticos [15] que normalmente transmitem relações semânticas, que podem ser encontradas em corpos. São elas:

a) `SINONIMO_DE`
b) `HIPERONIMO_DE`
c) `PARTE_DE`
d) `CAUSADOR_DE`
e) `FINALIDADE_DE`
f) `INSTRUMENTO_PARA`

Depois de clicar no endereço do VARRA (ver figura 4, ícone ), o Folheador passa os parâmetros necessários ao serviço. Este, depois de processar os dados, apresenta somente as frases em que os termos co-ocorrem com o padrão semântico correspondente à relação semântica. Na figura 8 são apresentadas duas frases retornadas pelo serviço para a relação "`computador HIPONIMO_DE máquina`".

3.4 O visualizador gráfico

De modo a facilitar a visualização da informação disponibilizada pela interface do Foheador, foi criado um visualizador gráfico (ver figura 9, ícone ). Este, além de mostrar visualmente informação de um determinado termo, permite navegar graficamente na informação apresentada. Como podemos ver na imagem 9, o visualizador gráfico apresenta os triplos a que um determinado termo (neste caso '`computador`', ver figura 5) está associado, as suas relações semânticas e os termos a que está ligado (por exemplo, "`HIPONIMO_DE máquina`").

Como referido na secção 3.1, o Folheador disponibiliza várias opções de pesquisa, no entanto somente duas dão acesso ao visualizador gráfico: pesquisar *a) por um termo* e *c) selecionando um determinado termo nos resultados apresentados* (para mais detalhe sobre estas duas opções ver secção 3.1). Somente através destas duas opções é possível visualizar o termo que procuramos no visualizador gráfico. Uma vez no visializador, é possível clicar em qualquer nó do grafo, permitindo deste modo que possamos com maior facilidade analisar os nós relacionados com o termo que procuramos.

No seu desenvolvimento, optamos por usar o JavaScript InfoViss Toolkit[16] devido a inúmeras vantagens, tais como: a sua facilidade de integração com o Folheador, a sua linguagem de programação (JavaScript), o seu desempenho na realização na tarefa pretendida e a sua disponibilidade (acesso livre).

4 O Núcleo do Folheador

O *Núcleo* do Folheador é composto por um conjunto de programas, ficheiros e uma base de dados, e pode ser dividido em duas partes: a primeira que tem como objetivos criar toda a estrutura da base de dados e inserir os valores na mesma (ver figura 10); a segunda parte abrange todos os programas necessários ao cálculo e inserção dos valores de confiança (ver figura 12). Em seguida é apresentada uma pequena descrição dos ficheiros que aparecem nas figuras referidas anteriormente (10 e 12):

a) Descrição da figura 10:
   i) `criaBD.php` → *contém toda a estrutura da base de dados; permite criar a base de*

[15] `http://linguateca.pt/acesso/padroes_relacoes_semanticas.php`

[16] `http://thejit.org`

par=2530: Ela trazia irregularmente do Paraguai computadores, **aparelhos** eletrônicos e uísque .
par=2548: Em outubro, Wanderlei usou cerca de r $ 15 mil que Márcia havia juntado com os seus contrabandos para comprar duas televisões, dois videocassetes, um **aparelho** de som, uma filmadora e um computador .
: · Para você que quer ter uma coleção de músicas para seu computador ou mesmo para ouvir no seu novo **aparelho** de MP3, não perca essa oportunidade .
: Usando o CyberTracker, um software com o qual os ecologistas podem registrar suas observações em campo usando computadores portáteis conectados a **aparelhos** de posicionamento global (GPS) , os rastreadores puderam reunir dados que comprovam a degradação da população local da espécie .
: · Para você que quer ter uma coleção de músicas para seu computador ou mesmo para ouvir no seu novo **aparelho** de MP3, não perca essa oportunidade .
: Segundo a pesquisa, 16,6 % dos domicílios brasileiros têm computadores de mesa, contra 95,7 % que têm **aparelhos** de TV .
par=49126: O **aparelho** está equipado com modernos instrumentos de telecomunicações, primeiros-socorros, páraquedas e computador .
par=saude16727: Os avanços da ecografia, enquanto tecnologia, resultam da evolução da Infor mática, afinal, estes **aparelhos** são computadores que analisam o som e a imagem .

Figura 7: Algumas frases retornadas pelo AC/DC para as palavras `aparelho` e `computador`.

Relação	Procura	Exemplo
máquina HIPERONIMO_DE computador	padrões usados	*par=Mais--94a-2*: E também de ensinar **máquinas** como **computadores** a identificarem 'ses objetos . (NSC)
máquina HIPERONIMO_DE computador	padrões usados	*par=ext328388-soc-95a-2*: **Máquinas** como os **computadores**, os faxes e os videofones devem poder comunicar entre si sem falhas, o que supõe um trabalho de programação importante . (CP)

Figura 8: Frases que exemplificam a relação `computador HIPONIMO_DE máquina`.

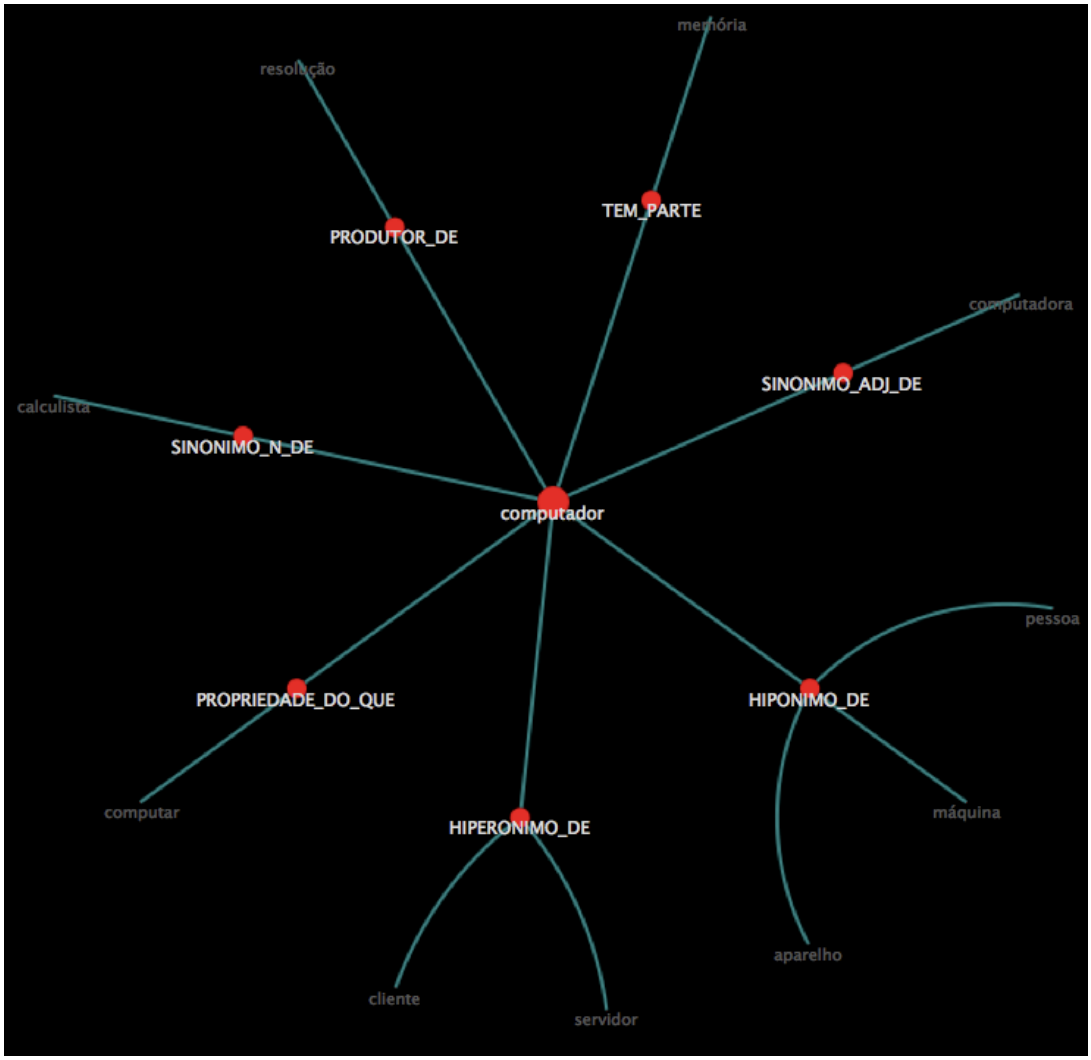

Figura 9: Interface gráfica do Folheador.

dados

ii) `insereDadosNaBD.php` → *permite inserir triplos na base de dados*

iii) `funcoes_insere.php` → *funções auxiliares do ficheiro* ***insereDadosNaBD.php***

b) Descrição da figura 12:

calcula

i) `calcula_valConfSimples.php` → *calcula valor de confiança simples*

ii) `funcoes_cqp.php` → *funções auxiliares do ficheiro* ***calcula_valConfSimples.php;*** *contém as funções que permitem invocar o CQP (Evert, 2005)*

iii) `indice_ficheiro-recurso.conf` → *associa o nome do recurso com o nome do ficheiro*

insere

i) `insereSimples.php` → *insere os valores do grau de confiança simples na base de dados*

c) Ficheiros comuns às duas imagens:

i) `triplos` → *ficheiros com os triplos dos vários recursos*

ii) `Base de Dados` → *ficheiro SQLite* [17] que representa a base de dados

De modo a termos uma melhor perceção da estrutura da base de dados, é apresentado o diagrama da base de dados na figura 11. A figura apresenta todas as tabelas, ligações e valores dos campos que constituem a base de dados.

Relativamente às relações semânticas (diretas) que estão presentes nesta versão do sistema, estas são apresentadas na tabela 1.

De seguida descreve-se como foi criada a base de dados e como inserir dados na mesma (secção 4.1), como é possível adicionar novos recursos (secção 4.2) e como podem ser calculados os valores de confiança dos dados previamente inseridos (secção 4.3).

[17] `http://www.sqlite.org/index.html`

4.1 Como criar e inserir dados na base de dados?

Relativamente ao sistema de gestão da base de dados, optou-se por usar o SQLite [17] , pois já tem embutido o motor de base de dados SQL. Diferentemente de outros motores de bases de dados SQL, o SQLite não tem um processo de servidor separado. O SQLite lê e escreve diretamente para um ficheiro no disco rígido facilitando a sua portabilidade. Para além da sua licença ser de domínio público, o que permite a sua livre utilização para qualquer tipo de finalidade, uma outra vantagem da sua utilização passa pela ausência de configurações adicionais.

Para criar a base de dados basta executar o ficheiro em PHP `criaBD.php` (ver figura 10). Este descreve a estrutura inicial da base de dados. O ficheiro também insere as categorias gramaticais, relações e sub-relações (inversas e diretas) que o Folheador abrange, nas tabelas: `relacoes`, `catGram`, `subRelacaoDireta` e `subRelacaoInversa`, respetivamente (ver figura 11). Para executar o programa basta usar o seguinte comando: `$php -q criaBD.php`.

Depois de criada a estrutura da base de dados e inseridos os valores iniciais nas tabelas referidas anteriormente, é necessário introduzir, de uma forma manual, na tabela `recursos`, o(s) nome(s) do(s) recurso(s) que se pretende introduzir na base de dados. Este passo é necessário, pois aquando da inserção dos dados, estes são identificados pelo identificador do recurso a que pertencem. Por outras palavras, o campo `recursos` da tabela `triplos` terá associado o(s) identificador(es) do(s) recurso(s) a que pertence (campo `id` da tabela `recursos`).

Em seguida, como podemos observar na figura 10, os ficheiros `funcoes_insere.php` e `insereDadosNaBD.php` têm como finalidade guardar os `triplos` na `base de dados`. Para isso, é necessário executar no terminal o seguinte comando: `$php -q insereDadosNaBD.php`. De um modo simples, o comando anterior guarda na base de dados todos os triplos presentes no(s) ficheiro(s) que pretendemos processar (identificados como `triplos` na figura 10). Mais especificamente: os termos dos triplos são guardados na tabela `termos`; e usando os seus identificadores (`ids`) é criada uma nova entrada na tabela `triplos` (campos `termo1ID` e `termo2ID`). De um modo homólogo, o identificador da relação (tabela `relacoes`, campo `id`), sub-relação direta (tabela `subRelacaoDireta`, campo `id`) e inversa (tabela `subRelacaoInversa`, campo `id`) são usados para

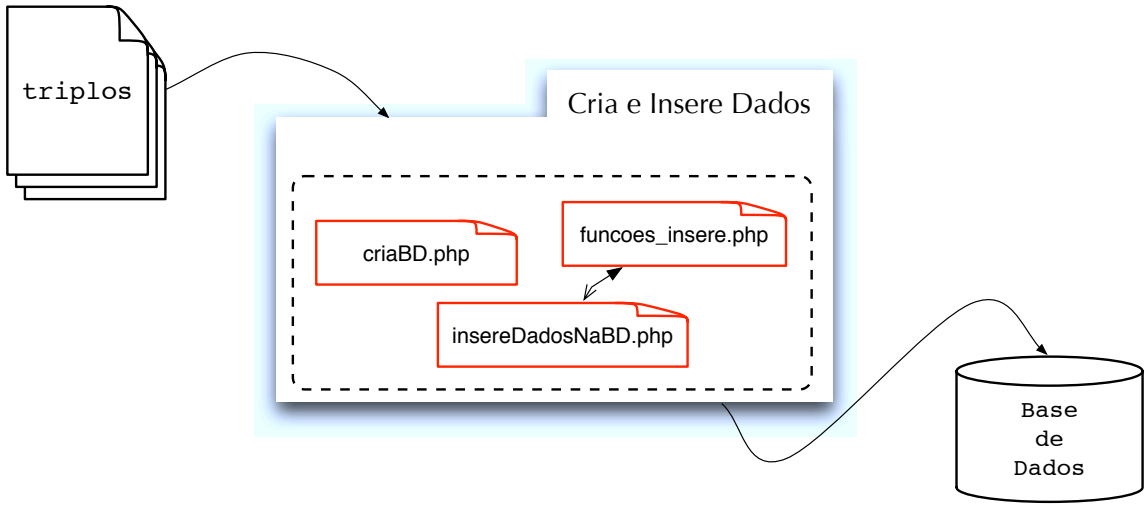

Figura 10: Criação e inserção de dados na base de dados.

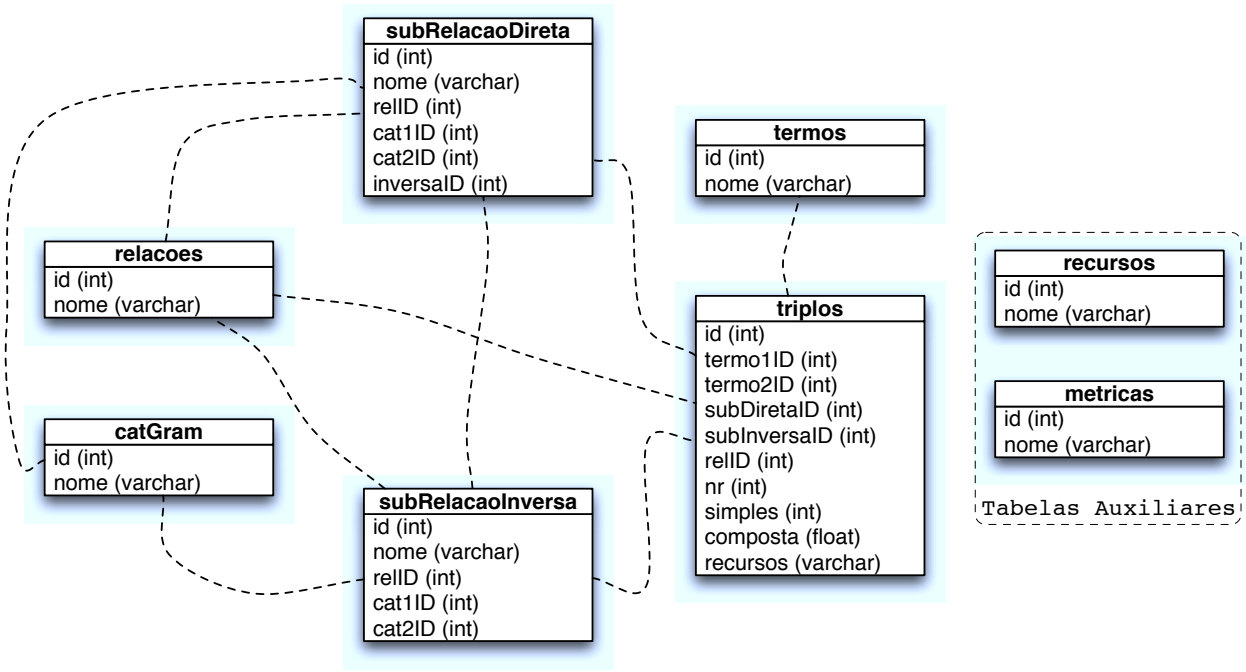

Figura 11: Diagrama da base de dados.

preencher os campos `relID`, `subDiretaID` e `subInversaID` da tabela triplos, respetivamente.

Um modo simples de verificar se os valores foram inseridos corretamente na base de dados passa por usar o SQLite Manager[18], uma extensão gratuita do browser Mozilla Firefox[19].

4.2 Como adicionar novos recursos?

Nesta secção são apresentados os recursos já existentes na base de dados, assim como os passos necessários para adicionar novos recursos.

De momento o Folheador agrega triplos de cinco recursos diferentes: PAPEL 2.0, TeP 2.0 (Dias Da Silva, Oliveira e De Moraes, 2002; Dias Da Silva e De Moraes, 2003; Maziero et al., 2008), OpenThesaurus.PT[20], Wikcionário[21] e Dicionário Aberto[22] (Simões e Farinha, 2010).

- ◇ O PAPEL resulta da extração automática de triplos do Dicionário PRO da língua portuguesa da Porto Editora.
- ◇ O TeP e o OpenTesaurus.PT são dois

[18] `https://addons.mozilla.org/en-US/firefox/addon/sqlite-manager`
[19] `http://www.mozilla.org`
[20] `http://openthesaurus.caixamagica.pt`
[21] `http://pt.wiktionary.org`
[22] `http://www.dicionario-aberto.net`

Cat1	Sub-Relação	Cat2
SINONIMIA		
verbo	SINONIMO_V_DE	verbo
adj.	SINONIMO_ADJ_DE	adj.
adv.	SINONIMO_ADV_DE	adv.
nome	SINONIMO_N_DE	nome
		HIPERONIMIA
nome	HIPERONIMO_DE	nome
PARTE		
nome	PARTE_DE	nome
nome	PARTE_DE_ALGO_COM_PROPRIEDADE	adj.
adj.	PROPRIEDADE_DE_ALGO_PARTE_DE	nome
		MEMBRO
nome	MEMBRO_DE	nome
nome	MEMBRO_DE_ALGO_COM_PROPRIEDADE	adj.
adj.	PROPRIEDADE_DE_ALGO_MEMBRO_DE	nome
CONTIDO		
nome	CONTIDO_EM	nome
nome	CONTIDO_EM_ALGO_COM_PROPRIEDADE	adj.
		MATERIAL
nome	MATERIAL_DE	nome
FINALIDADE		
nome	FINALIDADE_DE	nome
nome	FINALIDADE_DA_ACCAO	verbo
nome	FINALIDADE_DE_ALGO_COM_PROPRIEDADE	adj.
verbo	ACCAO_FINALIDADE_DE	nome
verbo	ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE	adj.
		CAUSA
nome	CAUSADOR_DE	nome
nome	CAUSADOR_DA_ACCAO	verbo
nome	CAUSADOR_DE_ALGO_COM_PROPRIEDADE	adj.
adj.	PROPRIEDADE_DE_ALGO_QUE_CAUSA	nome
verbo	ACCAO_QUE_CAUSA	nome
PRODUTOR		
nome	PRODUTOR_DE	nome
nome	PRODUTOR_DE_ALGO_COM_PROPRIEDADE	adj.
adj.	PROPRIEDADE_DE_ALGO_PRODUTOR_DE	nome
		MANEIRA
adv.	MANEIRA_POR_MEIO_DE	nome
adv.	MANEIRA_SEM	nome
adv.	MANEIRA_COM_PROPRIEDADE	adj.
adv.	MANEIRA_SEM_ACCAO	verbo
verbo	ACCAO_PARA_MANEIRA	adv.
REFERENTE		
adj.	PROPRIEDADE_DE_ALGO_REFERENTE_A	nome
adj.	PROPRIEDADE_DO_QUE	verbo
		LOCAL
nome	LOCAL_ORIGEM_DE	nome
nome	LOCAL_ONDE	verbo

Tabela 1: Relações Semânticas presentes na base de dados.

tesauros da língua portuguesa[23] criados manualmente, contendo relações de sinonímia.

◇ Os restantes dois recursos, Wikcionário e Dicionário Aberto, são resultado da extração automática de relações

[23]Os thesaurus estão organizados em synsets que contêm, por isso, palavras relacionadas por sinonímia. Para os convertermos num formato de triplos, consideramos que existia um triplo de sinonímia entre cada par de termos dentro do mesmo synset. Por exemplo: x SINONIMO-DE y, onde x e y são itens lexicais que pertencem ao mesmo synset.

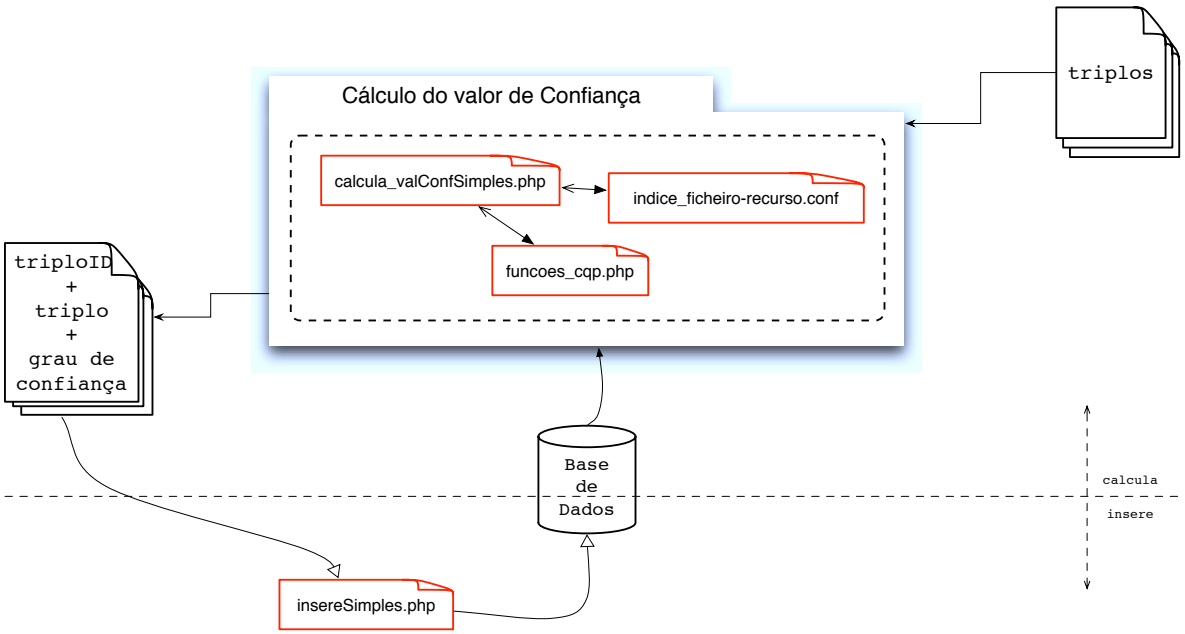

Figura 12: Cálculo e inserção do valor de confiança na base de dados.

semânticas no âmbito do projeto Onto.PT (Gonçalo Oliveira e Gomes, 2010; Gonçalo Oliveira et al., 2011). Estes incluem os triplos extraídos do Wikcionário da língua portuguesa e do Dicionário Aberto, respetivamente.

Apresentados os recursos, que neste momento já estão inseridos na base de dados, seguidamente descrevem-se os passos necessários para inserir um novo recurso:

1) primeiro é necessário inserir o nome do recurso na tabela `recursos`, campo `nome` (ver figura 11);
2) em seguida, é necessário alterar o ficheiro `indice_ficheiro-recurso.php`, de modo a fazer corresponder o nome do ficheiro que se pretende inserir com o nome do recurso;
3) por último, é preciso executar o programa `insereDadosNaBD.php` (ver secção 4.1).

Caso o novo recurso a inserir tenha outras relações ou categorias gramaticais, diferentes das apresentadas na tabela 1, é necessário introduzir manualmente na base de dados:

- o nome das categorias gramaticais (tabela `catGram`);
- o nome das novas relações, sub-relações diretas e as suas inversas (tabela `relacoes`, `subRelacaoDireta` e `subRelacaoInversa`, respetivamente). No caso das sub-relações, também é necessário fazer corresponder os identificadores das categorias gramaticais (campo `id`, tabela `catGram`) e da relação a que as sub-relações pertencem (campo `id`, tabela `relacoes`) aos campos `cat1ID`, `cat2ID` e `relID`, respetivamente.

Como podemos verificar, o processo de adição de um novo recurso tem de ser feito de um modo manual. Deste modo, é imperativo a criação de uma interface de administração capaz de facilitar este processo (ver secção 6), eliminando por completo qualquer tipo de falha que possa advir deste processo.

4.3 Como calcular e inserir os valores de confiança?

O processo de cálculo e inserção do valor do grau de confiança é relativamente simples. Os ficheiros envolvidos no processo foram descritos anteriormente na secção 4, tópico "b) Descrição da figura 12".

Como se verifica na figura 12, o processo de cálculo e inserção está dividido em duas partes. A primeira apresenta o processo de cálculo e a segunda a inserção dos valores na base de dados. De modo a que seja mais percetível, estes dois passos são detalhados em seguida.

Cálculo do valor do grau de confiança simples

Para calcular o valor do grau de confiança,

primeiramente é necessário associar o ficheiro de triplos, que se pretende processar, ao nome do recurso a que pertencem. Isto é possível através do ficheiro `indice_ficheiro-recurso.conf`. Em seguida executa-se a seguinte linha de código: `$php -q calcula_valConfSimples.php`. O programa anterior encarrega-se de criar um novo ficheiro. Este novo ficheiro, apresentado na figura 12 como `triploID + triplo + grau de confiança`, contém os triplos que leu do ficheiro original (`triplos`), o identificador do triplo (usando a informação contida na `Base de Dados`) e o valor de confiança atribuído. A opção de criar este ficheiro intermédio serve dois propósitos: futura disponibilização à comunidade e a facilidade de inserir os valores de confiança na base de dados (ver tópico seguinte). De modo a que seja possível ter acesso ao identificador do triplo, é necessário que anteriormente os triplos tenham sido guardados na base de dados, ou seja, para calcular e inserir os valores de confiança é obrigatório criar e inserir os dados/triplos previamente na base de dados (ver secção 4.1).

Relativamente ao valor do grau de confiança, este resulta do número de co-ocorrências dos termos na mesma frase. Este valor é calculado usando todos os corpos disponibilizados pelo serviço AC/DC. Mais especificamente, o ficheiro `calcula_valConfSimples.php` faz chamadas ao AC/DC com a seguinte questão: `MU(meet[lema='').$termo1.('']lema='').$termo2.('']s);`, em que `$termo1` e `$termo2` correspondem aos termos do triplo.

Inserir os valores na base de dados

Como referido anteriormente, a criação de um ficheiro com os identificadores dos triplos, os triplos em si e os seus graus de confiança, é uma mais valia no processo de inserção. Usando este ficheiro é possível atualizar o grau de confiança na tabela `triplos`, campo `simples`, de uma forma direta, ou seja como o ficheiro tem associado o identificador ao valor de confiança, o processo de inserção é muito mais rápido.

Para associar os valores de confiança (ver tópico anterior) aos triplos (previamente inseridos na base de dados, ver secção 4.2), é somente necessário executar o seguinte comando no terminal: `$php -q insereSimples.php`.

Embora o processo de cálculo e inserção não seja muito complicado de realizar, seria uma mais valia ser efetuado numa interface de administração (ver secção 6).

5 Observações Finais

Neste relatório técnico é apresentado o novo Folheador e todas as suas componentes. De um modo geral, o trabalho resulta num serviço na rede que permite navegar e explorar relações semânticas da língua portuguesa. Este, além de facilitar a pesquisa em diferentes bases de conhecimento lexical para o português, tem a vantagem de estar ligado a uma interface de corpos, permitindo deste modo a exploração de palavras relacionadas em contextos reais. Além de apresentar diferentes valores de confiança, para cada relação semântica, o serviço incorpora uma pequena interface gráfica que permite explorar visualmente todas as relações semânticas de uma palavra selecionada no Folheador. Depois de apresentadas as características, mais gerais do serviço, apresentamos em seguida alguns aspectos mais particulares.

Como seria de esperar, foram criadas as condições necessárias para uma inclusão de novos recursos na base de dados mais fácil. Também foram implementados os métodos necessários para poder, através da linguagem PHP, invocar o AC/DC e o VARRA, facilitando a interligação de *"linguagens para a web"* (HTML e PHP) com uma ferramenta de corpos, o CQP (Evert, 2005), permitindo, deste modo, o cálculo dos valores de confiança e a própria visualização de exemplos em contexto através da interface do Folheador.

Em termos de *design* foi criada uma página na rede que se pode considerar: simples, funcional, compreensível, de fácil utilização e com linhas modernas. De referir também que código do Folheador se encontra na rede, mais precisamente hospedado no google code[24]. O projeto é disponibilizado sob a licença GNU GPL v3, permitindo deste modo que a comunidade o descarregue e o altere de acordo com as suas necessidades.

Contudo, pode-se concluir que apesar de muitos alicerces terem sido acimentados, ainda ficaram muitos retoques por dar (ver secção 6). No entanto, podemos considerar que os objetivos propostos para esta versão foram atingidos com sucesso.

6 Trabalho Futuro

Como referido anteriormente, o desenvolvimento do Folheador sofreu várias reestruturações. Estas alterações não só permitiram solidificar vários pontos, como também contribuíram para o desenvolvimento de novas ideias. Em seguida são

[24] `http://code.google.com/p/folheador`

apresentadas algumas delas:

Embora não tenha havido tempo para calcular o valor de confiança composto, todas as condições estão criadas (ver secção 3.2) para que este seja integrado na próxima iteração do projeto.

Neste momento o Folheador permite invocar dois recursos externos (ver secção 3.3). Contudo, não é tido em conta a não co-ocorrência dos termos no(s) recurso(s). Deste modo, quando já sabemos de antemão que não existem co-cocorrências no(s) recurso(s) externo(s), a(s) ligações(ões) para o(s) mesmo(s) não deverá(ão) ser apresentada(s).

Uma caraterística que no futuro seria importante integrar no Folheador, é a inclusão de frases de exemplo que já foram validadas pelos varredores do VARRA. Através da análise das mesmas, acredita-se que é possível criar um novo valor de confiança. Ou seja, usando os dossiês já avaliados, pode-se tirar partido do número de vezes que os termos foram validados com sucesso nos vários contextos.

A criação de uma interface de administração faz todo o sentido neste projeto, pois existe a necessidade de:

- inserir novos dados de um modo amigável (por exemplo, a nova versão do PAPEL (Gonçalo Oliveira et al., 2011)). Ou seja inserir novos recursos na base de dados automaticamente e sem que o administrador tenha de executar os programas na linha de comandos;
- disponibilizar a opção de calcular os valores de confiança sem que seja necessário correr a linha de comandos;
- gerir a base de dados, por exemplo adicionar novas categorias gramaticais, nomes de recursos, etc.;
- visualizar os acessos que foram feitos ao Folheador, permitindo deste modo analisar quais as pesquisas mais frequentes.

O Folheador deverá oferecer a ligação aos recursos abrangidos pelo mesmo, ou seja, sugere-se que haja ligações à página do OpenThesaurus.PT, Wikcionário e Dicionário Aberto, assim como a novos recursos que venham a ser integrados no projeto.

Apesar de podermos considerar que a pesquisa e navegação no Folheador esteja bastante completa, ainda é possível tornar a mesma mais versátil, permitindo por exemplo fixar a categoria gramatical do(s) termo(s) na pesquisa.

Sugere-se também que o visualizador gráfico seja integrado na página de resultados do Folheador, de modo a oferecer uma alternativa visual às pesquisas feitas pelos utilizadores.

Por fim, também seria uma mais valia para a comunidade disponibilizar os vários recursos subjacentes ao Folheador, juntamente com os valores de confiança associados, e/ou disponibilizar a própria base de dados.

Para terminar, pensamos que, apesar do curto espaço de tempo desta iteração, o projeto contribui com um serviço que acreditamos que venha a ser muito útil à comunidade de processamento da língua portuguesa.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e presentemente pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

Agradeço a orientação do Hugo Gonçalo Oliveira e da Diana Santos. Também estou grato aos comentários do Fernando Ribeiro em relação a uma versão preliminar do Folheador.

Referências

[Dias Da Silva, Oliveira e De Moraes2002] Dias Da Silva, Bento C., Mirna F. de Oliveira, e Helio R. De Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 189–196, Berlin/Heidelberg. Springer-Verlag.

[Dias Da Silva e De Moraes2003] Dias Da Silva, Bento Carlos e Helio Roberto De Moraes. 2003. A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115.

[Evert2005] Evert, Stefan. 2005. The CQP Query Language Tutorial (CWB version 2.2.b90).

Relatório técnico, University of Stuttgart, Stuttgart, Germany.

[Freitas et al.2010] Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, e Violeta Quental. 2010. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. `http://www.linguateca.pt/Diana/download/resFreitasetalELC2010.pdf`.

[Gonçalo Oliveira et al.2011] Gonçalo Oliveira, Hugo, Leticia Anton Pérez, Hernani Costa, e Paulo Gomes. 2011. Uma rede léxico-semânticas de grandes dimensões para o português, extraída a partir de dicionários electrónicos. 3(2):23–38, December, 2011.

[Gonçalo Oliveira e Gomes2010] Gonçalo Oliveira, Hugo e Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. Em *Proc. 5th European Starting AI Researcher Symposium (STAIRS 2010)*, pp. 199–211. IOS Press.

[Gonçalo Oliveira, Santos e Gomes2010] Gonçalo Oliveira, Hugo, Diana Santos, e Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, Maio, 2010.

[Maziero et al.2008] Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, e Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390–392.

[Santos2011] Santos, Diana. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3(2):113–128, Junho, 2011.

[Santos e Bick2000] Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em *Proceedings of 2nd International Conf. on Language Resources and Evaluation*, LREC'2000, pp. 205–210.

[Simões e Farinha2010] Simões, Alberto e Rita Farinha. 2010. Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, (16), Setembro, 2010.