

Introdução à Utilização do Corpógrafo

Um Pequeno Tutorial

1. Introdução

O Corpógrafo é uma plataforma destinada à análise e ao trabalho sobre corpora com o objectivo essencial de produzir diversos recursos linguísticos em suporte informático. O Corpógrafo pretende apoiar os investigadores da língua portuguesa num conjunto de tarefas que vão desde a compilação de corpora, à extracção e organização do conhecimento gerado a partir deles.

A criação de recursos linguísticos (dicionários, tesauros, listas de regras, etc.) em suporte informático é hoje em dia uma prioridade para todos, mesmo para aqueles que não trabalham directamente com a língua. De facto, numa sociedade cada vez mais baseada no conhecimento, todos necessitamos, directa ou indirectamente, de bons recursos linguísticos que apoiem as nossas actividades de produção e o acesso a esse conhecimento. Desde os populares correctores ortográficos até às mais complexas aplicações de recolha de informação ou tradução automática, muitos são os sistemas que baseiam a sua eficácia em recursos linguísticos completos. A penetração e a importância destas ferramentas na nossa sociedade serão cada vez maiores, justificando assim claramente a relevância da produção de bons e variados recursos linguísticos.

O Corpógrafo pretende ser uma pequena contribuição nesse esforço de construção, integrando num único ambiente de trabalho várias ferramentas que habitualmente se encontram dispersas e cuja integração nem sempre é fácil. O Corpógrafo está em desenvolvimento, mas neste momento já permite algumas tarefas interessantes para investigadores, estudantes e todos aqueles que, de alguma forma, trabalhem com a língua, como por exemplo, tradutores.

Utilizando o Corpógrafo, qualquer utilizador poderá compilar e organizar o seu *corpus pessoal* e estudá-lo ou pesquisá-lo de várias formas. O Corpógrafo encontra-se neste momento optimizado para pesquisa terminológica, possuindo várias técnicas de extracção terminológica e a possibilidade de organização dessa informação em bases de dados dedicadas. Mas possui também funcionalidades genéricas de pesquisa por *expressões regulares*, de estudo de *N-Gramas*, de organização de corpus e outras que vão sendo adicionadas a pedido dos utilizadores.

Neste pequeno tutorial de utilização do Corpógrafo pretendemos mostrar algumas das funcionalidades disponíveis e mostrar também como se podem resolver algumas situações mais problemáticas da sua utilização. O utilizador será acompanhado passo a passo ao longo de um processo de pesquisa terminológica, que ilustrará as actuais capacidades do Corpógrafo. O Corpógrafo é uma ferramenta em construção, possuindo, por isso, ainda algumas limitações que, no entanto, não impedem o trabalho sobre corpora. Por outro lado, o facto de a ferramenta não estar terminada e fechada permite a sua expansão à medida que os utilizadores vão sugerindo funcionalidades úteis, pelo que esperamos também a sua participação no desenvolvimento do Corpógrafo.

2. Preparação para este Tutorial

O Corpógrafo está dividido em quatro grandes áreas de trabalho:

1. *Gestor*, onde se encontram as ferramentas de edição e *pré-processamento* dos ficheiros;
2. *Pesquisa*, onde se realizam as pesquisas genéricas;
3. *Centro de Conhecimento*; onde se encontram as ferramentas que permitem gerir e organizar o conhecimento, bem como realizar pesquisas especificamente relacionadas com informação conceptual, como é o caso da pesquisa de termos e de relações semânticas;
4. Centro de Comunicação, onde se pode encontrar toda a documentação existente sobre o Corpógrafo, enviar comunicações ao Administrador do Corpógrafo e verificar as mensagens que recebeu.

Seguidamente, abordaremos em pormenor cada uma destas áreas de trabalho.

Para uma melhor compreensão do Corpógrafo e para que possa ir experimentando na prática as suas várias funcionalidades, iremos trabalhar com dois textos reais. Escolhemos dois textos de exemplo e não apenas um, uma vez que algumas das funcionalidades do Corpógrafo só se poderão demonstrar com mais de um ficheiro. Para obter os textos de exemplo, aceda ao sítio da Linateca em www.linguateca.pt, no menu clique em **Produção Escrita** e grave no seu computador o texto nº 27, "Um centro de recursos para o processamento computacional do português", e o texto nº 2, "Computational processing of Portuguese: working memo", da lista apresentada. Os ficheiros gravados serão do tipo HTML, como muitos

outros que poderá descarregar da rede através de pesquisas usando motores de pesquisa como, por exemplo, o *Google* (www.google.com) ou o *Tumba!* (www.tumba.pt).

Os textos que acabou de descarregar descrevem alguma da actividade produzida na Linguateca e contêm alguma da terminologia associada à área do Processamento da Linguagem Natural e da Linguística Computacional.

Vamos começar por trabalhar com o texto nº 27. Como poderá ver, fazendo duplo clique sobre o documento de texto, o texto contém cabeçalho, algumas tabelas e figuras, resumo em português e em inglês, agradecimentos, referências bibliográficas e ainda alguma informação sobre a autora.

Este texto servirá para ilustrar o que pode ser feito com o Corpógrafo, assim como o caminho a percorrer para fazê-lo.

Nota:

O Corpógrafo permite trabalhar não só com ficheiros HTML, mas também com outros tipos de ficheiros como, por exemplo, PDF, PS, DOC e RTF.

3. Gestor

O Corpógrafo dá-lhe acesso a uma área pessoal onde pode armazenar ficheiros, que constituirão os seus corpora. É através do *Gestor* que fará a gestão dos corpora que criar.

O *Gestor* é a base do Corpógrafo. Tudo o que se faz aqui condiciona as posteriores pesquisas, não só ao nível da eficácia das mesmas, mas também ao nível da sua qualidade.

Aqui encontrará as ferramentas necessárias para preparar o seu corpus de forma a torná-lo pesquisável. Para poder vir a efectuar uma pesquisa rápida e eficaz, terá de preparar aqui todos os ficheiros que for introduzindo no Corpógrafo, isto é, terá de “limpá-los”, verificar se a divisão em frases (*fraseamento*) foi feita correctamente, classificá-los e categorizá-los. Todas as operações relacionadas com o carregamento, pré-processamento e organização dos ficheiros que introduziu no Corpógrafo serão realizadas no *Gestor*.

As pesquisas que efectuar terão, certamente, um objectivo. Mediante o seu objectivo, os textos que inserir no Corpógrafo terão de ser *pré-processados*. Por exemplo, as *pesquisas terminológicas* requerem uma operação de *fraseamento* mais rigorosa para se poder verificar mais precisamente o contexto de determinadas ocorrências, enquanto que outras pesquisas, como é o caso da *pesquisa de colocações*, exigem apenas uma limpeza de eventuais caracteres estranhos, resultantes de uma extracção imperfeita do texto. Importa, portanto, definir quais os seus objectivos antes de iniciar a “limpeza” dos textos.

Questões a pensar nesta fase:

- Qual é o nosso objectivo de pesquisa com este texto?
- Que operações de *pré-processamento* são necessárias para a concretização desse objectivo?

Dada a riqueza terminológica do texto de exemplo nº 27, que retirámos do sítio da Linguateca, poderemos aproveitá-lo para fazer um levantamento terminológico em português europeu, pelo que iremos definir este como o nosso objectivo.

Comecemos então por carregar o texto para a área pessoal do Corpógrafo.

A partir da página principal, clique em **Gestor** e em **Ficheiros**. No menu **Adicionar Ficheiros**, no lado esquerdo do ecrã, clique em **Do meu Computador**. Selecciona o texto a carregar e clique em **Carregar**. Será imediatamente direccionado para a ficha de **Dados Gerais** do ficheiro. Aqui poderá preencher os campos que achar relevantes, sendo que o único que é obrigatório é o do **Idioma**, uma vez que se este não estiver definido, não poderão ser efectuadas pesquisas. Depois de preencher os campos que desejar, não se esqueça de clicar em **Guardar**.

Carregar Ficheiros para o Corpógrafo**Dica:**

Poderá salvar uma cópia do ficheiro original no disco. Assim, terá sempre acesso ao original e poderá descarregar esse ficheiro directamente do Corpógrafo. Esta opção já está pré-definida aquando do carregamento do ficheiro para o Corpógrafo.

Editar Ficheiros

The screenshot shows the 'Corpógrafo - Linguateca' interface in Microsoft Internet Explorer. The left sidebar contains a tree view with 'Gestor' and 'Ficheiros' expanded. The main area displays the 'Dados Gerais' form for a file named 'Art_02.htm'. The form fields are as follows:

- Nome do Ficheiro: Art_02.htm
- Título do documento: (empty)
- Idioma do Texto: Português (PT)
- Acesso público: ☐
- Data de introdução: 2005-09-13 11:14:25
- Data de alteração: 2005-09-13 11:15:53

Below the form, the 'Ficheiro Original' section states: 'O documento original está disponível para ser descarregado aqui' with a download icon. The 'Descrição' field contains the text 'comentario de'. A 'Guardar' button is located at the bottom of the form.

Fig. 1 - Dados Gerais do Ficheiro

Para além dos Dados Gerais do ficheiro, poderá ainda associar-lhe informação relativa à sua proveniência (autor, fonte, editora, etc.) e também classificá-lo semanticamente. No menu **Ficheiro** clique em **Fontes** e será direccionado para a área onde poderá associar autores, fontes, editoras, etc. aos seus ficheiros.

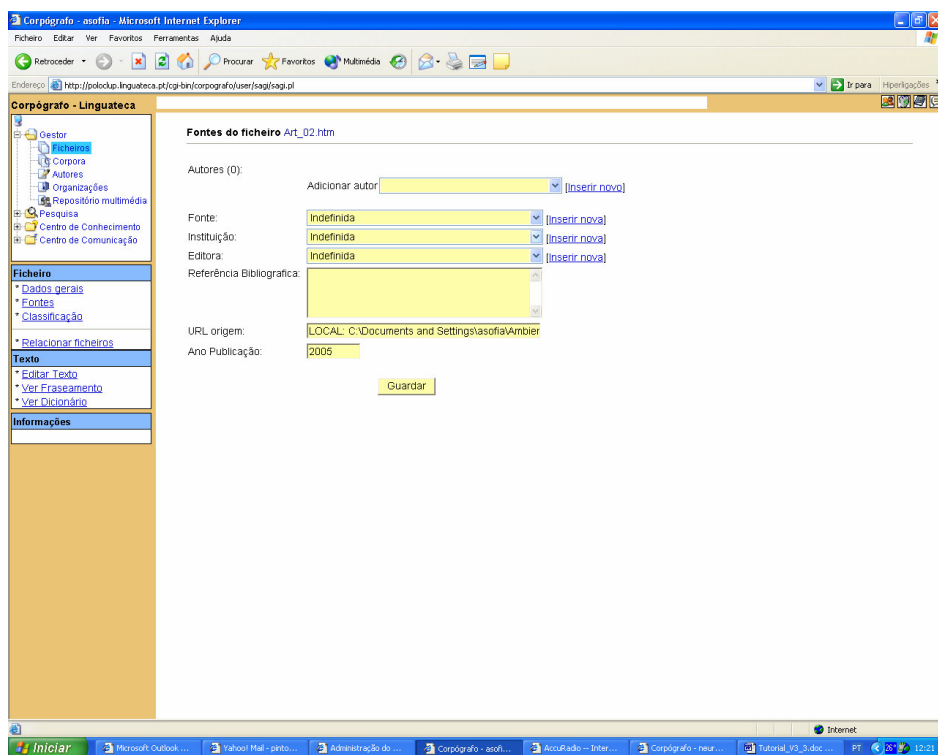


Fig. 2 - Fontes do Ficheiro

Tendo em conta que vários textos podem ter um mesmo autor e estar associados a uma mesma organização, o Corpógrafo permite-lhe agora fazer a gestão dos autores e organizações para que não precise de inserir um mesmo autor e/ou organização sempre que insere um novo ficheiro. Isto é, se tem vários ficheiros de um autor e/ou provenientes de uma mesma editora, basta que insira esse autor e/ou editora uma única vez.

Clique no botão **Inserir novo**, ao lado de **Adicionar Autor**. Na janela que vai aparecer, digite o nome do autor que pretende inserir, se desejar, preencha os restantes campos e clique em **Inserir Autor**.

Para inserir a **Fonte** do ficheiro clique em **Inserir nova**, digite o nome da fonte a inserir, preencha os campos que forem do seu interesse e clique em **Inserir Organização**. O processo para inserir uma **Instituição** e/ou uma **Editora** é o mesmo. Note que Fontes, Instituições e Editoras são

Nota:

Para associar uma organização (fonte, editora, etc.) a um autor terá de inserir primeiro a organização e só depois o autor

Organizações e, no caso de, por exemplo, a Fonte e a Editora serem uma mesma Organização, só terá de a inserir uma vez (ou em Fonte, ou em Editora) para que esta fique imediatamente disponível para ser seleccionada.

Autores e Organizações

Tanto os Autores como as Organizações podem ser inseridos a partir do menu lateral principal, uma vez que cada um tem uma área própria. Clicando em **Autores** terá acesso à lista de todos os autores que tiver inserido e poderá editá-los ou adicionar novos autores.

O mesmo acontece com as Organizações. Clicando em **Organizações** terá acesso à lista de todas as organizações que tiver inserido e poderá editá-las, assim como inserir novas organizações.

Todos os autores e organizações aqui inseridos ficarão disponíveis para serem associados aos ficheiros.

Para classificar semanticamente o ficheiro, também no menu **Ficheiro**, clique em **Classificação**. Aí poderá definir o Médium, Domínio, Subdomínio do ficheiro, etc. Se não encontrar uma determinada classe no Corpógrafo para classificar o seu texto, poderá sugerir a adição da referida classe através do *Centro de Comunicação* (ver **Centro de Comunicação**).

Em seguida, terá de “limpar” o ficheiro de forma a torná-lo pesquisável. No menu **Texto** clique em **Editar Texto** para ter acesso à área onde poderá realizar todas as tarefas de edição, processamento e tratamento de texto.

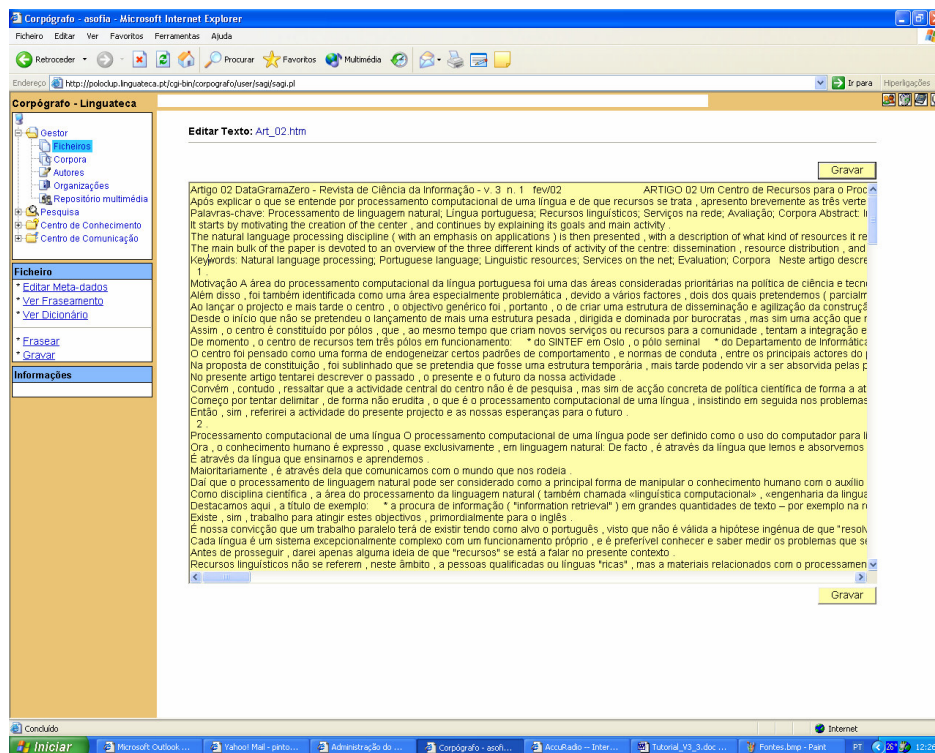


Fig. 3 – Tratamento de Texto

Para “limpar” o texto, isto é, remover tudo o que considere irrelevante (ex.: caracteres estranhos, referências bibliográficas, etc.) e/ou corrigir erros (como palavras com letras trocadas, por exemplo, “porcesso” em vez de “processo”), basta clicar em qualquer parte do texto e editar. No caso do texto em questão, poderíamos eliminar o cabeçalho, o resumo em inglês, as tabelas, as datas, as referências bibliográficas, os links, os agradecimentos e também a informação sobre a autora.

Pode apagar ou acrescentar o que quiser, não esquecendo que, tal como num documento Word, é preciso gravar todas as alterações. Para isso, clique em **Gravar**. Poderá ainda verificar se a divisão do texto em frases foi feita correctamente, clicando em **Ver Fraseamento**.

Remover informação supérflua ou ruidosa

Dica:

Se o texto aparecer numa só linha ou não for visível, clique em **Frasear**. Se mesmo assim o texto não aparecer, provavelmente o ficheiro não poderá ser utilizado.

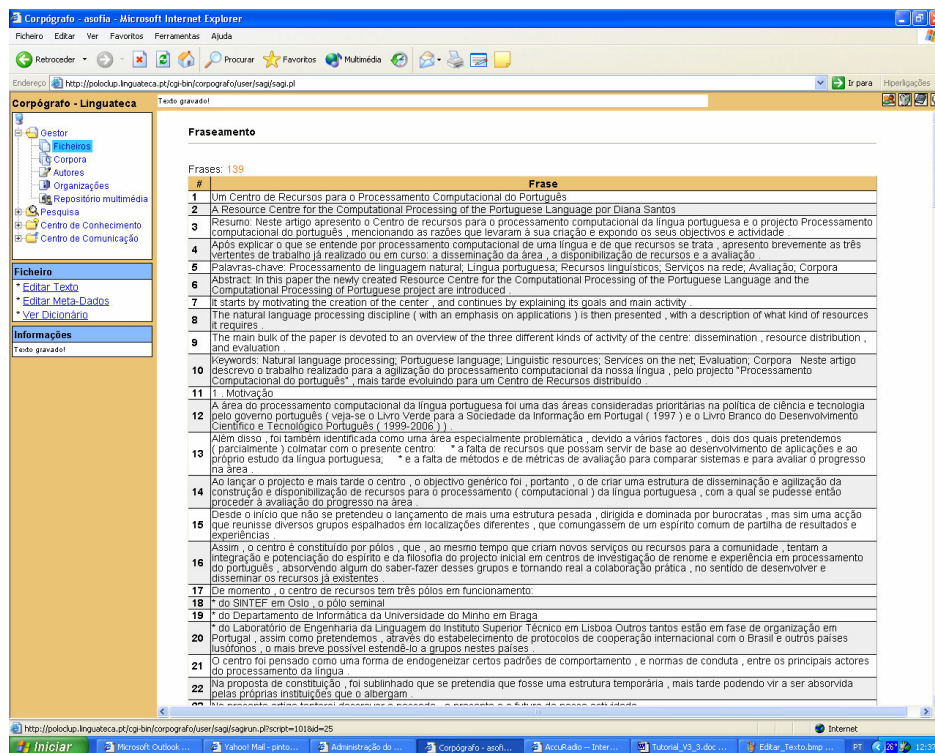


Fig. 4 - Ver Fraseamento

Caso detecte algum erro na divisão em frases poderá facilmente corrigi-lo, clicando em **Editar Texto** no menu **Ficheiro**.

Ainda a partir da área de edição de texto, clique em **Ver Dicionário** para poder ter uma panorâmica geral dos átomos que constituem o ficheiro que se encontra a tratar.

Uma outra funcionalidade que encontrará aqui será a possibilidade de estabelecer relações de comparabilidade entre ficheiros. Isto é, se tiver inserido dois (ou mais) ficheiros que considere comparáveis, poderá estabelecer essa relação entre os ficheiros no Corpógrafo.

Como neste momento só temos um ficheiro armazenado no Corpógrafo, vamos inserir o texto de exemplo nº 2. Clique em **Ficheiros** no menu superior e, seguidamente, em **Do meu computador**, no menu **Adicionar Ficheiros** e siga os passos acima descritos para inserção de ficheiros. Depois de inserir o ficheiro, no menu **Ficheiro**, clique em **Relacionar Ficheiros**, seleccione o ficheiro que pretende relacionar e escolha a opção **Comparáveis**. De imediato ambos os ficheiros serão considerados comparáveis. Para eliminar uma relação basta clicar em **Apagar**.

Clicando em **Ficheiros** no menu superior terá acesso à lista de todos os ficheiros que tiver inserido no Corpógrafo e, se desejar, poderá organizar esses ficheiros em pastas. Para fazê-lo clique em **Criar Pasta**, no menu

Operações, e digite o nome que pretende dar a essa pasta na caixa de diálogo que aparece e clique **OK**. Em seguida, seleccione os ficheiros que pretende que sejam movidos para a pasta que criou e clique em **Mover para....** Na janela de diálogo que lhe vai aparecer, clique na pasta para onde pretende mover os ficheiros e depois no botão **Mover para**

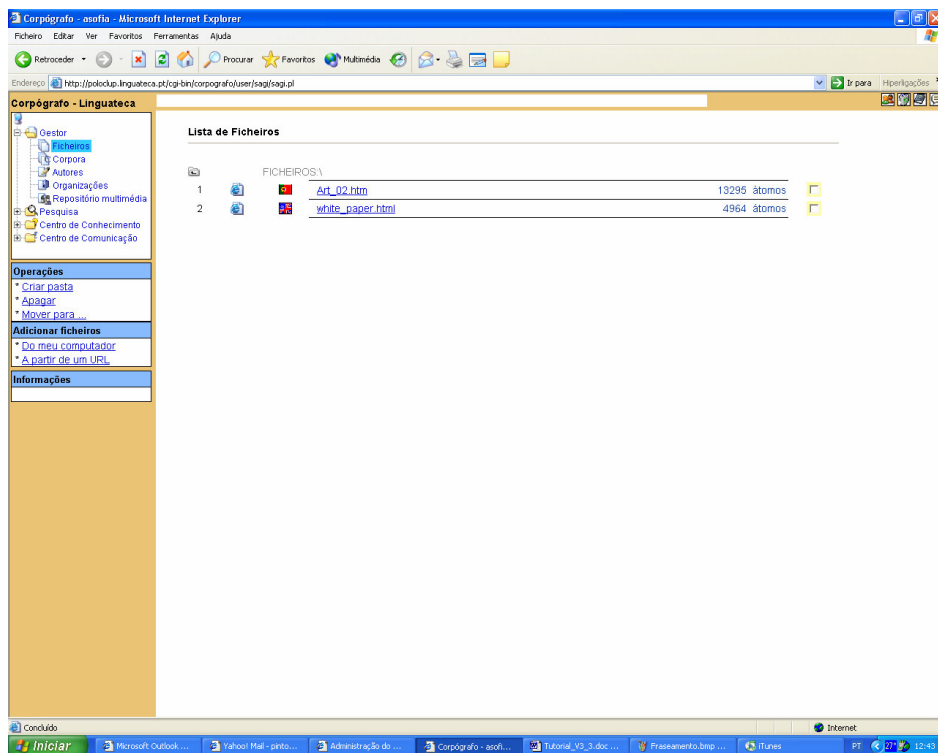


Fig. 5 - Lista de Ficheiros

Depois de realizar todas as tarefas de tratamento de texto, o seu ficheiro estará pronto para fazer parte de um *corpus*. O *Corpus* é a unidade de texto que é efectivamente pesquisável, podendo ser composto por vários textos individuais que tenha disponíveis na sua área pessoal.

Passemos então à criação de *corpora* para podermos efectuar pesquisas.

Criar Corpora para pesquisa

No **Gestor** clique em **Corpora** e depois em **Novo Corpus**. Dê um nome ao *corpus* que pretende criar, defina o idioma e preencha os campos que achar relevantes. Clique em **Criar** e depois em **Anexar Ficheiros**, no menu **Corpus**. Ser-lhe-á apresentada a lista de todos os ficheiros que se encontram armazenados no Corpógrafo. Seleccione os ficheiros que pretende que façam parte do *corpus* (neste caso, um dos ficheiros de exemplo) e clique em **Incluir Ficheiros**, no menu lateral **Corpus**.

Os *corpora* podem ser editados, isto é, pode acrescentar-lhes ou retirar-lhes ficheiros já depois de os ter criado. Para adicionar/remover ficheiros clique em **Corpora** no **Gestor** e, em seguida, no *corpus* a editar. No menu lateral **Corpus** clique em **Anexar Ficheiros**. Os ficheiros a cinza correspondem aos ficheiros que já fazem parte do corpus que se encontra a tratar. Seleccione quais os ficheiros que pretende acrescentar ao corpus e clique em **Incluir Ficheiros**.

Para remover ficheiros, escolha os ficheiros a remover do corpus e clique em **Excluir Ficheiros**, no menu lateral **Corpus**.

Se desejar alterar um ficheiro já depois de este ter sido adicionado a um *corpus*, pode fazê-lo, o *corpus* será actualizado automaticamente.

Tal como os ficheiros, também os corpora poderão ser organizados em pastas. Na lista de corpora, clique em **Nova Pasta** no menu lateral **Corpora** para criar a pasta para onde irá mover o corpus (ou corpora). Em seguida, no mesmo menu clique em **Mover para...** Clique sobre a pasta para onde quer mover o corpus (ou corpora) e depois no botão **Mover para...**

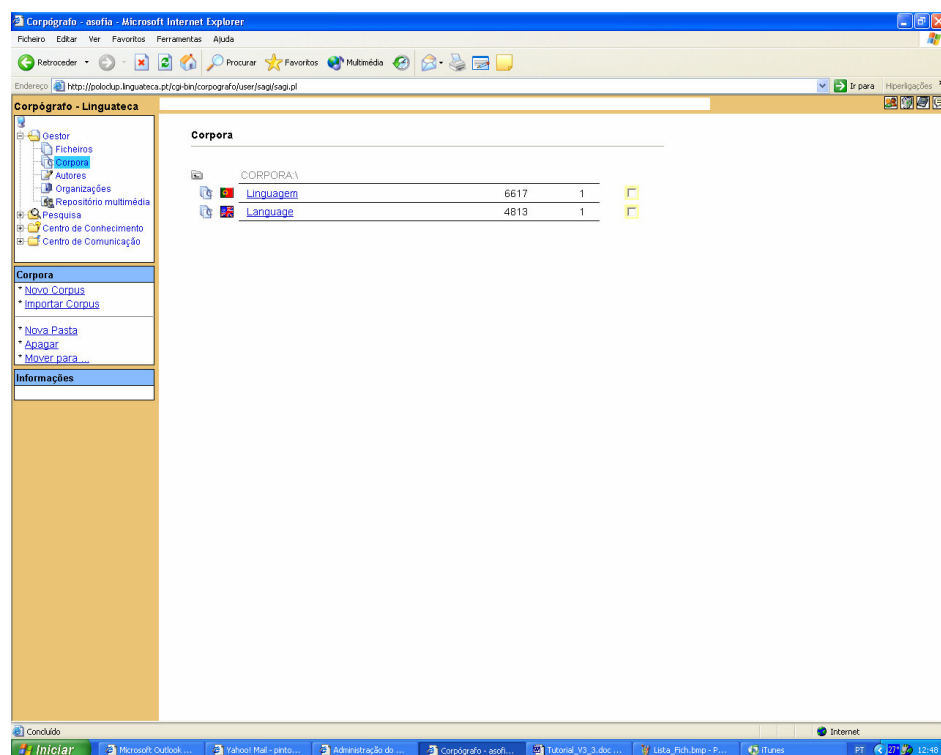


Fig. 6 - Lista dos Corpora

Uma vez realizadas estas tarefas, o seu *corpus* estará pronto para ser pesquisado. Para aceder à área onde poderá efectuar as pesquisas clique em **Pesquisa** no menu do lado esquerdo.

4. Pesquisa

Aqui encontram-se as ferramentas que lhe permitirão pesquisar e estudar o(s) seu(s) corpus, assim como extrair informação genérica. Poderá efectuar pesquisas por *expressões regulares*, se pretender verificar *concordâncias* ou realizar estudos de *N-Gramas*, no caso de pretender saber a frequência com que ocorrem as palavras.

Comecemos por efectuar um estudo de *N-Gramas* para termos uma panorâmica geral do que poderemos encontrar no texto.

O estudo de *N-Gramas* é um estudo tradicional no âmbito da Linguística Computacional. Este estudo permite inspeccionar sequências de N palavras consecutivas (por exemplo, sequências de duas palavras, ou bigramas) que ocorrem no texto, verificando a sua frequência. Esta análise possibilita a visualização de algumas das características que se destacam no texto, como, por exemplo, construções sistemáticas de uma ou mais palavras que permitem identificar ou dar pistas para a identificação de termos ou de estilos de escrita.

Para realizar o estudo de *N-Gramas* clique em **Pesquisa** e em **Estudo de N-Gramas**. Selecciona o corpus e o tamanho da sequência a pesquisar. Escolha a forma como pretende que os resultados lhe sejam apresentados (ordenação por frequência ou alfabética) e clique em **Pesquisar**. Ser-lhe-á apresentada uma tabela com os resultados.

Estudo de N-Gramas

#	n-grama	# oc.	freq. %
1	a	419	6.332
2	de	337	5.092
3	a	201	3.037
4	que	141	2.130
5		138	2.085
6	e	137	2.070
7		130	1.964
8	o	120	1.813
9	do	102	1.541
10	l	100	1.511
11	o	100	1.511
12	da	83	1.254
13	para	77	1.163
14	um	77	1.163
15	em	70	1.057
16	uma	56	0.846
17	na	43	0.649
18	a	42	0.634
19	como	41	0.619
20		41	0.619
21	as	40	0.604
22	recursos	40	0.604
23	ou	39	0.589
24	por	38	0.574
25	com	36	0.544
26	lingua	35	0.528
27	*	34	0.513
28	não	34	0.513
29	mas	31	0.468
30	avaliação	28	0.423
31	dos	28	0.423
32	area	27	0.408
33	se	27	0.408
34	la	25	0.377
35	projecto	25	0.377
36	português	24	0.362
37	ser	24	0.362
38	processamento	23	0.347
39	a	22	0.332
40	forma	21	0.317
41	através	18	0.272

Fig. 7 – Resultados de um estudo de *N-Gramas*

Agora que já temos uma ideia das palavras que poderemos encontrar neste texto, podemos efectuar uma pesquisa de *concordâncias*.

Poderá realizar três tipos de pesquisa de concordâncias:

- **Concordância Frase** – a concordância é apresentada no contexto da frase em que ocorre.
- **Concordância Janela** – poderá definir o número de átomos que pretende que antecedam e precedam a concordância, sendo que cada átomo será apresentado dentro de uma janela, permitindo-lhe ordenar a concordância por qualquer um desses átomos.
- **Concordância KWIC** – poderá definir o tamanho do contexto (em palavras ou em caracteres) que pretende que anteceda e preceda a concordância.

Imaginemos que pretendia pesquisar todas as palavras que tivessem o prefixo “ling-” (por exemplo, linguagem, linguística) inseridas no contexto da frase. Para efectuar esta pesquisa, clique em **Concordância Frase** no menu lateral. Defina o corpus a pesquisar e insira a seguinte expressão: **ling\w+**. Clique em **Pesquisar!**. Ser-lhe-á apresentada uma tabela com todas as palavras existentes no corpus que começam pelo prefixo “ling-”.

Pesquisar Concordâncias

Dica:

Pode fazer o mesmo tipo de pesquisa para sufixos. Se quisesse ver todas as palavras existentes no corpus que terminam em “-mento” (ex.: conhecimento), bastava inserir a seguinte expressão de pesquisa: **\w+mento**.

Corpógrafo - Linguateca

Expressão de pesquisa: ling\w+
Corpus pesquisado: Linguagem (6617 átomos : 1 ficheiros)

Nº de concordâncias obtidas: 23 distribuídas por 2 ficheiros
Frequência da totalidade das instâncias da concordância: 0.34 %

#	f	Frase onde ocorre a concordância	Info
1	-	Desde a comparação de vários produtos segundo um conjunto de parâmetros externos (como por exemplo Rocha (2000)) a uma análise da contribuição de alguns factores no desempenho de uma única ferramenta (Santos & Oksefjell , 2000) , até à análise de recursos (Santos & Rocha , 2001) e da repetibilidade da avaliação de hipóteses linguísticas (Santos & Oksefjell , 1999) , muitas destas análises podem ser feitas como um trabalho isolado.	
2	-	De momento , estamos a considerar qual a melhor forma de um seu prolongamento , após uma reflexão sobre a utilidade , a qualidade do produto obtido , e a possibilidade de vir a incorporar elementos de grupos diferentes e que sigam paradigmas linguísticos diferentes.	
3	-	Finalmente , o projecto da Floresta Sinta (c) fica , veja-se http://cgi.portugues.mct.pt/ptreebank/ , tem como objectivo criar um conjunto de unidades correctamente analisadas sintacticamente (revistas por linguistas) que possam servir tanto como instrumento de calibragem de analisadores sintácticos automáticos , como para melhorar a cobertura descritiva da análise sintáctica computacional , obrigando a um consenso ou pelo menos a uma formalização das escolhas e alternativas tomadas em áreas "cinzentas" da sintaxe.	
4	-	Um projecto paralelo focando a linguagem jornalística brasileira , desenvolvido em colaboração com o Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo , do Brasil , encontra-se em fase final de execução.	
5	-	Um projecto paralelo focando a linguagem jornalística brasileira , desenvolvido em colaboração com o Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo , do Brasil , encontra-se em fase final de execução.	
6	-	O texto , do jornal diário português PÚBLICO , foi dividido em extractos ordenados aleatoriamente , de forma a impedir a reconstrução das notícias integrais , e é distribuído em CD (versão 1.0) gratuitamente a todos quantos registarem o seu endereço postal , além de ser distribuído (na sua versão 1.7) pelo Linguistic Data Consortium , a maior agência de distribuição de recursos linguísticos .	
7	-	O texto , do jornal diário português PÚBLICO , foi dividido em extractos ordenados aleatoriamente , de forma a impedir a reconstrução das notícias integrais , e é distribuído em CD (versão 1.0) gratuitamente a todos quantos registarem o seu endereço postal , além de ser distribuído (na sua versão 1.7) pelo Linguistic Data Consortium , a maior agência de distribuição de recursos linguísticos .	
8	-	Tal permite a um linguista (ou a um engenheiro da linguagem) interrogar um corpus formulando questões que excedem largamente a simples procura de palavras ou formas , bastando para isso ter acesso à Internet.	
9	-	Tal permite a um linguista (ou a um engenheiro da linguagem) interrogar um corpus formulando questões que excedem largamente a simples procura de palavras ou formas , bastando para isso ter acesso à Internet.	
10	-	Nessa linha , estamos a iniciar trabalho na procura de informação na Web , tentando obter um modelo da procura real que os utilizadores efectuem de forma a poder estudar formas de melhorá-la usando o processamento de linguagem natural.	
11	-	Alguns dos problemas identificados foram : área muito heterogénea , com dois pólos (correspondentes a engenharia e à linguística) , com tradições diferentes e sem facilidade de comunicação entre elas , falta de identificação da área como interessante em ambos os campos (é uma área periférica tanto para a informática como para a linguística) , falta quase total de pessoas que soubessem de ambos os campos.	
12	-	Alguns dos problemas identificados foram : área muito heterogénea , com dois pólos (correspondentes a engenharia e à linguística) , com tradições diferentes e sem facilidade de comunicação entre elas , falta de identificação da área como interessante em ambos os campos (é uma área periférica tanto para a informática como para a linguística) , falta quase total de pessoas que soubessem de ambos os campos.	
13	-	Uma das principais hipóteses veiculadas por esse documento era a de que devíamos concentrar esforços no português e não no processamento da língua em geral (ou no linguística computacional) .	
14	-	Recursos linguísticos não se referem , neste âmbito , a pessoas qualificadas ou linguas "ricas" , mas a materiais relacionados com o processamento da língua.	
15	-	Como disciplina científica , a área do processamento da linguagem natural (também chamada «linguística computacional» , «engenharia da linguagem» ou «tecnologia das línguas humanas») tem aplicações múltiplas (veja-se Santos (2001) para uma panorâmica mais abrangente da engenharia da linguagem , precisamente através das suas aplicações) .	
16	-	Como disciplina científica , a área do processamento da linguagem natural (também chamada «linguística computacional» , «engenharia da linguagem» ou «tecnologia das línguas humanas») tem aplicações múltiplas (veja-se Santos (2001) para uma panorâmica mais abrangente	

Fig. 8 – Resultados de uma Pesquisa de Concordância Frase

Poderá também fazer uma pesquisa de *Concordância Janela*. Imaginemos que queria saber que palavras ocorrem à esquerda, ou à direita, da palavra “linguagem”. Para isso, em **Pesquisa** clique em **Concordância Janela** e escolha o corpus a pesquisar. Introduza a expressão a pesquisar, neste caso *linguagem*, e defina o número de janelas que deverão estar antes e depois da concordância. Clique em **Pesquisar!**.

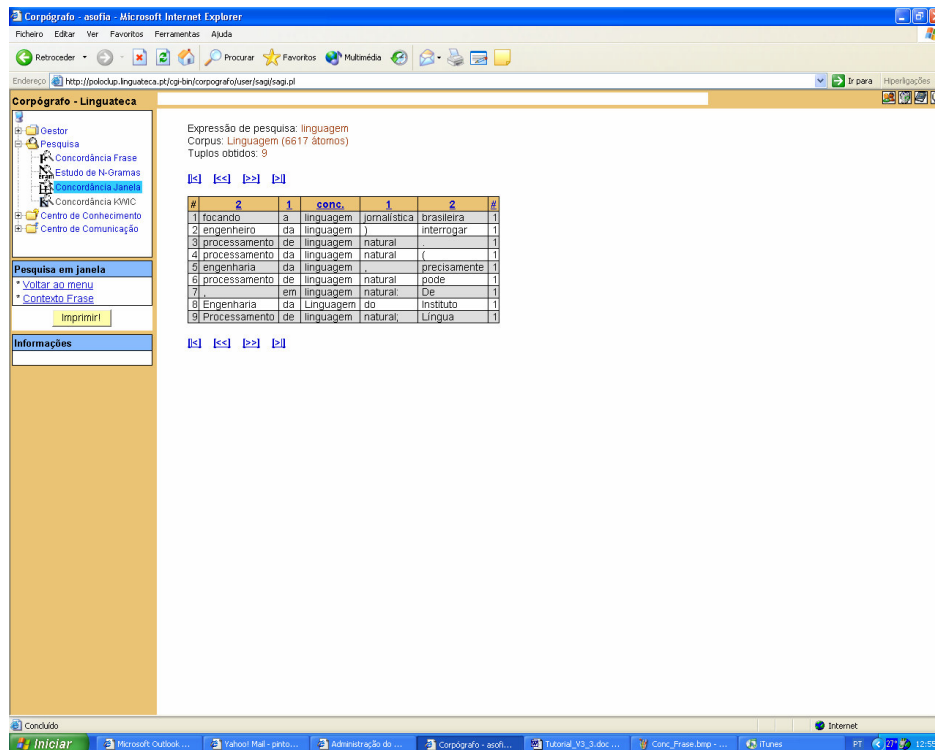


Fig. 9 - Resultados de uma Pesquisa de Concordância Janela

Depois de ficarmos com uma ideia geral do tipo de terminologia que poderemos encontrar no corpus exemplo, resta-nos fazer a *pesquisa terminológica* para que possamos extrair toda a terminologia relevante do texto (uma vez que é este o nosso objectivo). No entanto, abordaremos este tema numa secção mais à frente.

5. Centro de Conhecimento

O propósito do Corpógrafo é a construção de recursos informáticos reutilizáveis que representem uma mais valia em futuros processos de pesquisa de informação, produção de documentação, ou ainda de auxílio a actividades em torno da linguagem como a tradução.

O Corpógrafo possui uma zona onde é possível pesquisar e armazenar de forma sistemática toda a informação produzida através das pesquisas

nele realizadas. Essa zona foi denominada de **Centro de Conhecimento**. Fala-se em “conhecimento” e não em simples “informação”, uma vez que à informação base já está associado o contexto de utilização da mesma.

Por “conhecimento” entende-se toda a informação lexical, morfológica, sintáctica e semântica que permite tomar decisões em diferentes situações de processamento de informação textual (terminologia, listas de palavras, regras de pesquisa, padrões, relações semânticas, etc.).

Mais do que simples bases de dados ou listas de palavras, a informação que o Corpógrafo permite armazenar e gerir é uma teia complexa de elementos que representam a formalização do conhecimento linguístico e conceptual de um determinado domínio.

A terminologia é o elemento basilar desse conhecimento formal, pelo que neste momento o **Centro de Conhecimento** centra-se essencialmente na gestão das bases de dados terminológicas. O Corpógrafo permite armazenar sistematicamente em bases de dados criadas para esse efeito a informação relativa à terminologia que foi pesquisada usando as suas ferramentas.

Contudo, estas bases de dados terminológicas não são apenas simples listas de termos, pois permitem estabelecer relações entre os termos, criando uma malha conceptual multidimensional. Compreendendo a existência de diferentes tipos de relações entre conceitos que dependem fundamentalmente dos domínios de conhecimento em análise, o Corpógrafo permite a definição de novas relações semânticas, através das quais é possível associar termos de uma forma mais flexível. Isto é feito usando o *Gestor de Relações*, que será abordado mais à frente.

Presentemente, encontrará no Centro de Conhecimento as ferramentas necessárias para criar e editar bases de dados terminológicas, assim como para pesquisar e gerir relações semânticas entre termos.

Iremos agora demonstrar a utilização das bases de dados terminológicas e a forma como é possível relacionar os termos.

A extracção de terminologia é feita para uma determinada base de dados, pelo que começaremos por criar uma base de dados terminológica.

Clique em **Centro de Conhecimento** e, em seguida, em **BD Terminológicas**. No menu **Base de Dados** (lado esquerdo) clique em **Nova Base de Dados**. Preencha os campos relativos à base de dados a criar e clique em **Criar**.

Criar Bases de Dados

Será automaticamente direccionado para a base de dados que criou, onde terá um conjunto de ferramentas à sua disposição para a poder editar

(menu **Editar BD**). Será também aqui que poderá inserir termos manualmente. Imaginemos que queria inserir um termo que não foi detectado pelo Corpógrafo. Neste caso, basta digitar o termo na caixa **Novo Termo** e clicar em **Inserir**. O termo será imediatamente armazenado na base de dados.

Depois de criada a base de dados, estão reunidas todas as condições para iniciar a *pesquisa terminológica*.

5.1 Pesquisa Terminológica

Para efectuar a *pesquisa terminológica* era necessário criar primeiramente uma base de dados, uma vez que será nessa base de dados que terá acesso à ferramenta que lhe permite pesquisar termos e também, porque será nela que os termos que extrair dessa pesquisa serão armazenados.

Para iniciar a pesquisa terminológica, dentro da base de dados, clique em **Pesquisar termos** no menu **Editar BD**.

Seleccione o corpus que deseja pesquisar, defina se a pesquisa deve ser efectuada com restrições sobre o termo e o contexto (i. e., tendo em consideração o que pode aparecer antes de um termo e o que nunca pode estar dentro de um termo), ou sem qualquer tipo de restrição.

Defina ainda o tamanho mínimo da sequência terminológica a pesquisar, isto é, se os candidatos a termos deverão ser constituídos no mínimo por uma ou mais palavras (termos compostos). Se desejar, poderá tentar a singularização dos candidatos a termo (ou seja, o Corpógrafo tenta singularizar os candidatos que se encontram no plural), para que as ocorrências de um mesmo candidato a termo no singular e no plural sejam contabilizadas em conjunto (ex: imaginemos que “linguagem” ocorre 4 vezes e “linguagens” ocorre 2 vezes. Se tentarmos a singularização, estas duas palavras serão contabilizadas em conjunto, sendo apresentada como candidato a termo apenas “linguagem” com um total de 6 ocorrências). Poderá ainda optar por não visualizar os candidatos terminológicos que já tenham sido inseridos na base de dados (no caso de já ter efectuado uma pesquisa terminológica anteriormente). Por fim, clique em **Pesquisar!**.

Pesquisar Candidatos Terminológicos

Nota:

Note que as restrições só funcionarão correctamente se o idioma do corpus tiver sido definido aquando da compilação do mesmo. Se o idioma do corpus não tiver sido definido, o Corpógrafo não será capaz de fazer qualquer tipo de restrição na pesquisa terminológica, fazendo com que os resultados apresentados sejam iguais aos de um *Estudo de N-Gramas*.

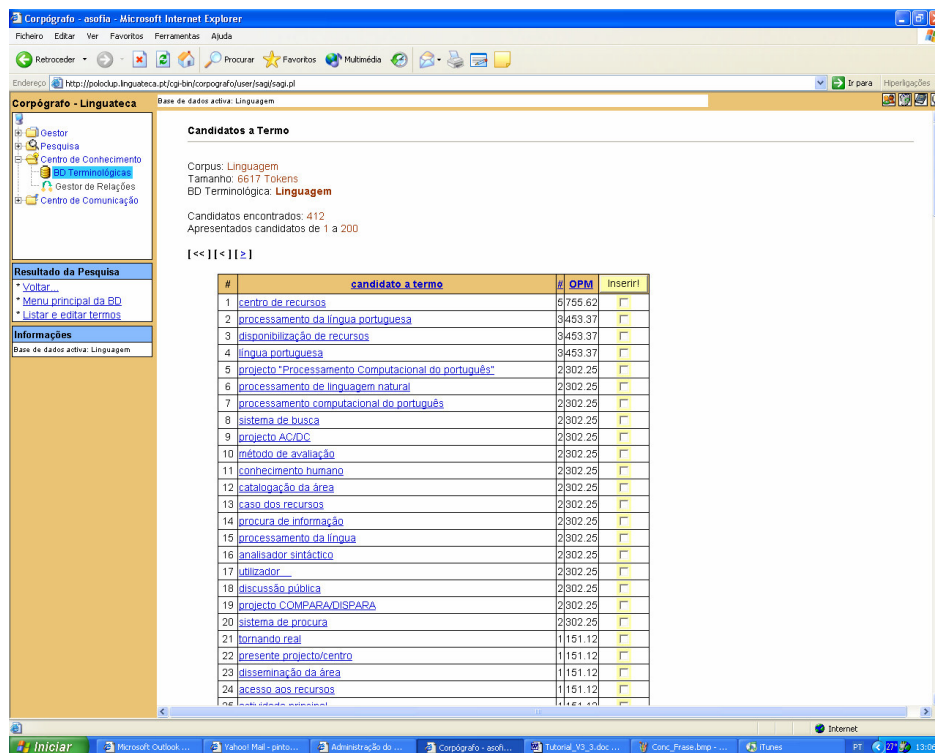


Fig. 10 – Resultados de uma Pesquisa Terminológica

Ser-lhe-á apresentada uma tabela com os candidatos terminológicos. Como pode verificar, não aparecem caracteres estranhos, palavras em inglês nem mesmo números, já que esses elementos foram eliminados aquando do pré-processamento do texto.

Selecione os termos a inserir na base de dados e clique em **Inserir!**.

Depois de extrair toda a terminologia e de a armazenar na base de dados, teremos de organizar e gerir o conhecimento que foi produzido.

5.2 Gestão da Informação Terminológica

Passemos então a demonstrar o que pode ser feito ao nível da organização e gestão das bases de dados terminológicas.

No menu do lado esquerdo, clique em **Listar e Editar Termos**. Aí poderá visualizar os termos que inseriu na base de dados de quatro formas diferentes: poderá ver a lista de todos os termos inseridos na base de dados (todos os termos em todas as línguas); poderá ver a lista de todos os termos inseridos numa só língua (ex.: todos os termos em português); todos os termos começados por determinada inicial em todas as línguas (ex.: todos os termos começados pela letra "c"); ou ainda todos os termos

Dica:

Se clicar no candidato a termo, poderá vê-lo em contexto para se certificar de que se trata realmente de um termo.

Dica:

Se quiser inserir manualmente um termo que não tenha sido identificado pelo Corpógrafo, poderá fazê-lo através do **Centro de Conhecimento** (ver **Criar BD** na secção anterior).

Listar Termos

começados por determinada inicial numa só língua (ex.: todos os termos começados pela letra “p” em inglês).

Escolha a opção que desejar e clique em **Listar**. Terá acesso à lista de termos que pretendia visualizar.

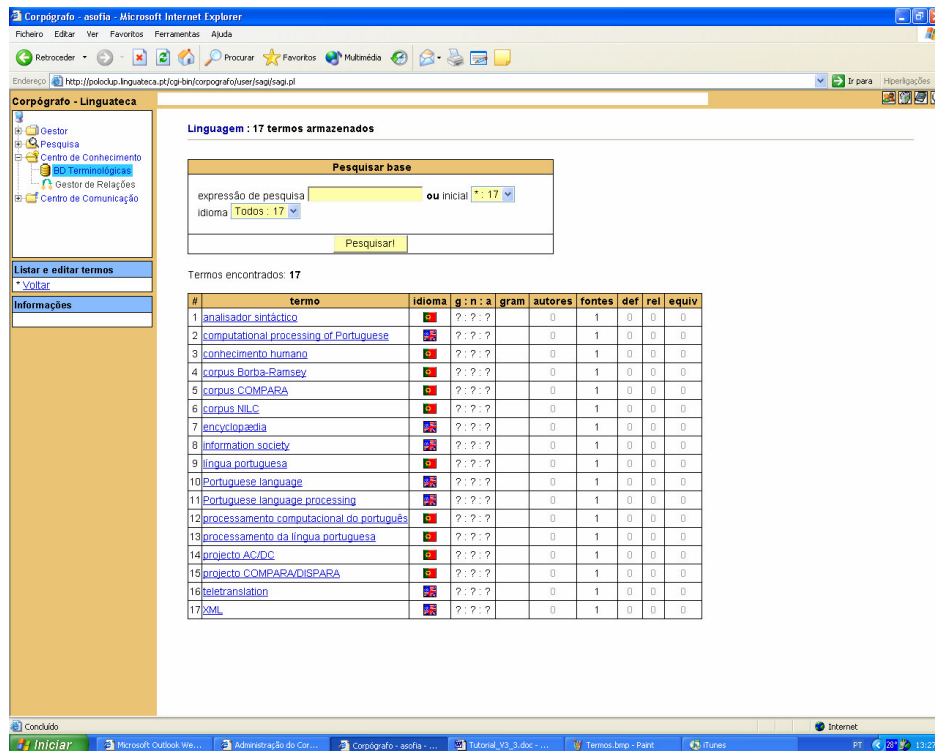


Fig.11 – Lista dos termos inseridos na base de dados

Para editar um termo basta clicar no termo que pretende editar. Encontrará a ficha do termo que contém toda a informação sobre esse mesmo termo que estiver armazenada na base de dados. No menu **Ficha do Termo** terá ao seu dispor um conjunto de ferramentas que lhe permitirão executar várias tarefas relacionadas com o termo. Poderá:

- editar os dados gerais do termo
- inserir/editar autores
- inserir/editar fontes,
- editar informação morfológica (ex: género e número),
- pesquisar definições do termo no corpus,
- pesquisar relações semânticas entre termos da base de dados (ex: merónimos, hipónimos, etc),
- pesquisar equivalentes de tradução,
- ver/associar objectos multimédia ao termo;

Editar Termos

- consultar estatísticas do termo num corpus;
- apagar o termo.

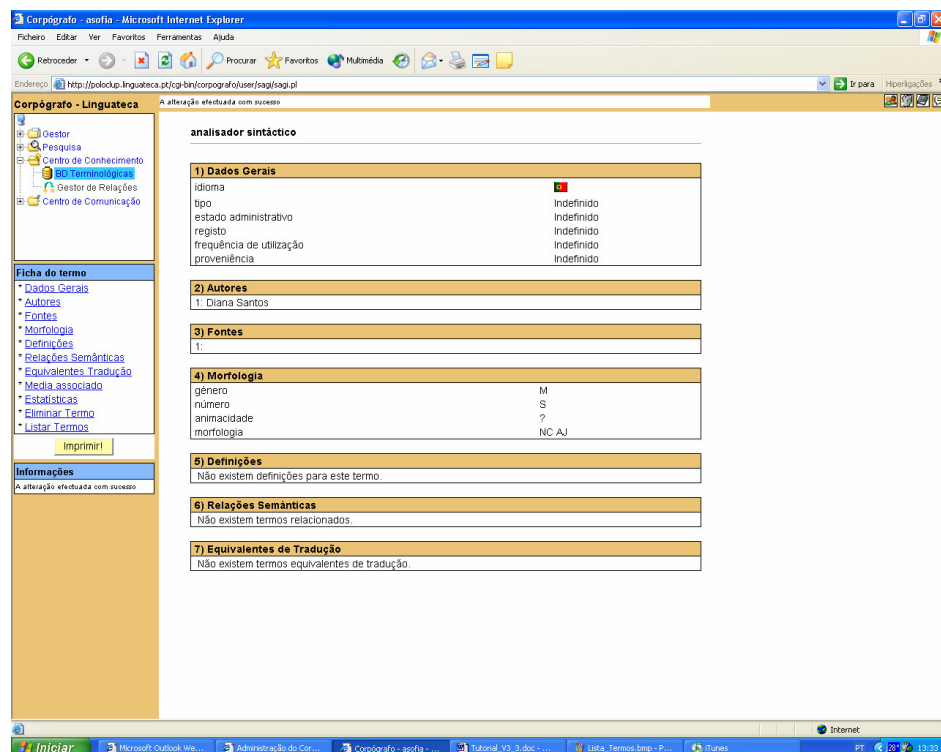


Fig. 12 – Ficha do Termo

Associar objectos multimédia a um termo:

Agora é possível associar objectos multimédia a um termo, isto é, se quiser associar um ficheiro de imagem, um ficheiro de vídeo, um ficheiro áudio, etc. a um termo poderá fazê-lo a partir do menu **Ficha do Termo**. Clique em **Media associado** e, no menu **Associar objecto media**, clique em **Do repositório**, se já tiver armazenado o ficheiro multimédia que pretende no **Repositório multimédia** do **Gestor**. Para associar o ficheiro ao termo, faça duplo clique sobre o nome do ficheiro.

Se ainda não tiver ficheiros armazenados no **Repositório multimédia**, no menu **Associar objecto media** clique em **Do meu PC**, seleccione o ficheiro a carregar e clique em **Carregar**. O ficheiro será imediatamente associado ao termo e armazenado no **Repositório multimédia**. Faça duplo clique no nome do ficheiro para o visualizar.

Depois de ter a sua base de dados “recheada”, terá a possibilidade de gerar uma pasta com toda a informação contida nessa base de dados. Ou seja, será possível gerar uma pasta, que poderá armazenar no seu computador e consultar sempre que quiser, com a lista de todos os termos que a sua base de dados contiver e toda a informação a eles associada (autores, fontes, definições, relações semânticas, equivalentes de tradução, etc.). Para tal, clique no nome da base de dados a partir da qual pretende gerar a pasta. No menu **Editar BD** clique em **Gerar Documentação** e o processo de geração da pasta terá início de imediato. **Note que este processo poderá demorar algum tempo!** Em seguida, aparecerá uma janela de diálogo. Clique em **Guardar** e seleccione o local onde quer que a pasta seja armazenada.

A partir da sua base de dados poderá ainda calcular as estatísticas globais dos termos num determinado corpus, se o desejar. Dentro da base de dados, clique em **Estatísticas dos termos** no menu **Editar BD**. Seleccione o corpus sobre o qual será feito o cálculo, o idioma e se pretende que o cálculo seja feito para todos os termos, ou para termos só de uma palavra, ou só de duas, etc. Por fim, clique em **Gerar Estatísticas!**.

Dica:

Se quiser, poderá exportar a sua base de dados para formato XML, de forma a torná-la utilizável por outros programas.

No menu **Editar BD** clique em **Exportar** e o processo de conversão será iniciado de imediato.

5.3 Pesquisa de Definições

O Corpógrafo permite pesquisar candidatos a definição de termo, ou seja, permite pesquisar um corpus para encontrar possíveis definições de um dado termo, desde que este já tenha sido introduzido na base de dados. As definições são pesquisadas com base em padrões lexicais que indiciam a presença de uma definição, como por exemplo:

- o TERMO é um
- o TERMO é definido como
- uma definição de TERMO é

Estes padrões são recolhidos manualmente através da verificação da sua ocorrência em corpora. Assim sendo, as listas de padrões não contêm todos os padrões lexicais existentes para extracção de definições, pelo que serão listas permanentemente em construção e permanentemente incompletas. Assim, se verificar que há uma definição de um determinado termo num corpus e que o Corpógrafo não a consegue detectar, por favor envie-nos a definição em questão, através do *Centro de Comunicação*, para que possamos recolher esse padrão e adicioná-lo ao Corpógrafo.

Para pesquisar candidatos a definição, na lista de termos clique no termo cuja definição pretende procurar. Clique em **Definições**, no menu **Ficha do Termo**, e depois em **Inserir ou Pesquisar**, no menu **Definições**. Selecciona o corpus a pesquisar e a técnica de pesquisa, isto é, se a pesquisa deverá ser feita usando restrições lexicais (disponível para inglês e português), ou se deverá ser uma pesquisa simples de *concordâncias*. Clique em **Pesquisar!**.

Se desejar guardar algum dos candidatos a definição, clique em **Guardar**. Para editar a definição basta clicar em qualquer parte da definição a editar, fazer as alterações que desejar e depois clicar em **Alterar!**.

Também poderá inserir definições manualmente. Clique em **Inserir ou Pesquisar**, digite a definição que pretende inserir e, por fim, clique em **Inserir**.

Pesquisar Definições

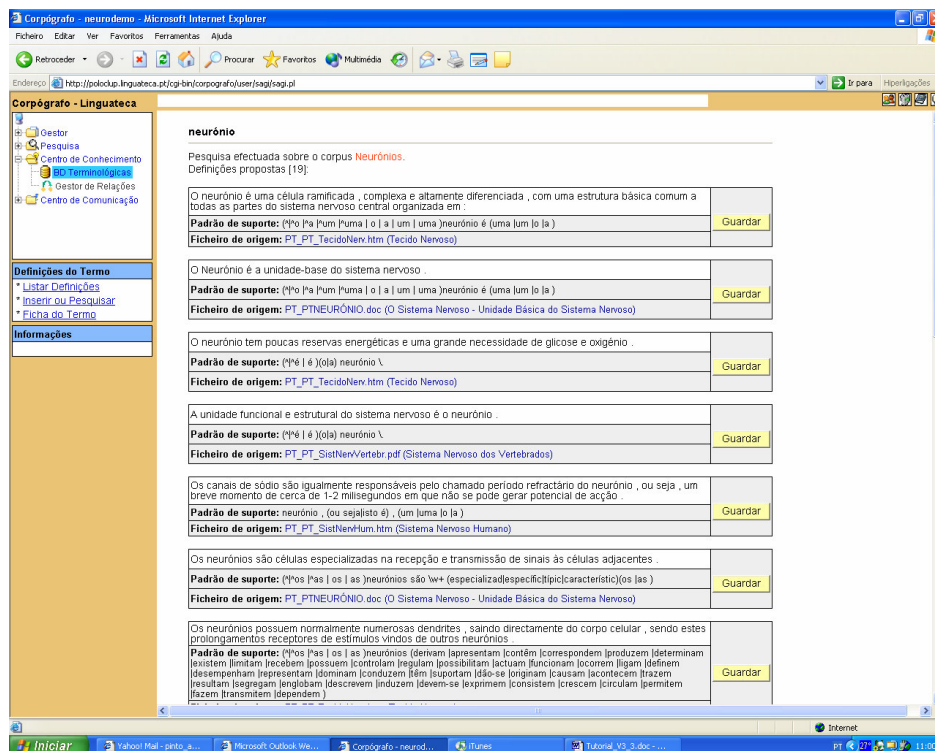


Fig. 13 – Resultados de uma Pesquisa de Definições com restrições

5.4 Pesquisa de Relações Semânticas

O Corpógrafo dá-lhe também a possibilidade de pesquisar algumas Relações Semânticas:

- **Hipónimo/Hiperónimo:**

Ex.: “um músico, principalmente um guitarrista.” (‘músico’ – hiperónimo; ‘guitarrista’ – hipónimo)

Ex.: “A amniocentese é um exemplo de um teste genético pré-natal” (‘amniocentese’ – hipónimo; ‘teste genético pré-natal’ – hiperónimo)

- **Merónimo/Holónimo:**

Ex.: “a terminação axónica do neurónio” (‘terminação axónica’ – merónimo; ‘neurónio’ – holónimo)

Ex.: “nos neurónios cujos axónios” (‘neurónios’ – holónimo; ‘axónio’ – merónimo)

- **Produtor/Produto:**

Ex.: “o retículo produz triglicéridos, fosfolípidos e esteróides”

Ex.: “a comunicação ocorre via Internet”

Ex.: “o campo magnético é gerado por duas bobines ligadas em série”

- **Processo/Objectivo:**

Ex.: “Produção de aminoácidos por fermentação”

Ex.: “Produção de Hidrocarbonetos por Pirólise de Resíduos Plásticos”

- **Causa/Efeito:**

Ex.: “O desenvolvimento físico provoca o crescimento dos neurónios”

Ex.: “uma câibra muscular é desencadeada por um espasmo muscular clónico”

- **Alterador/Alterado:**

Ex.: “O estímulo altera a permeabilidade da membrana”

- **Classe/Instância:**

Ex.: “O Porto é uma cidade”

Para pesquisar candidatos a *merónimos* ou a *hipónimos* (ou a outras relações previstas pelo Corpógrafo), na lista de termos clique no termo cujo *merónimo/hipónimo* pretende procurar e depois em **Relações Semânticas**. No menu **Pesquisar Relações**, clique em **Usando restrições lexicais**, seleccione a relação e o corpus a pesquisar e clique em **Pesquisar!**.

Se desejar guardar algum dos candidatos a *merónimo/hipónimo* (ou outra das relações previstas pelo Corpógrafo), clique em **Guardar!**.

Pesquisar Relações Semânticas

Poderá optar por pesquisar as co-ocorrências de um termo (i.e., uma pesquisa sem restrições) para verificar se esse termo e os termos com que co-ocorre estão ligados por alguma das relações semânticas previstas pelo Corpógrafo. Para tal, clique em **Verificando co-ocorrências**, no menu **Pesquisar Relações**, seleccione o corpus que pretende pesquisar e clique em **Pesquisar!**.

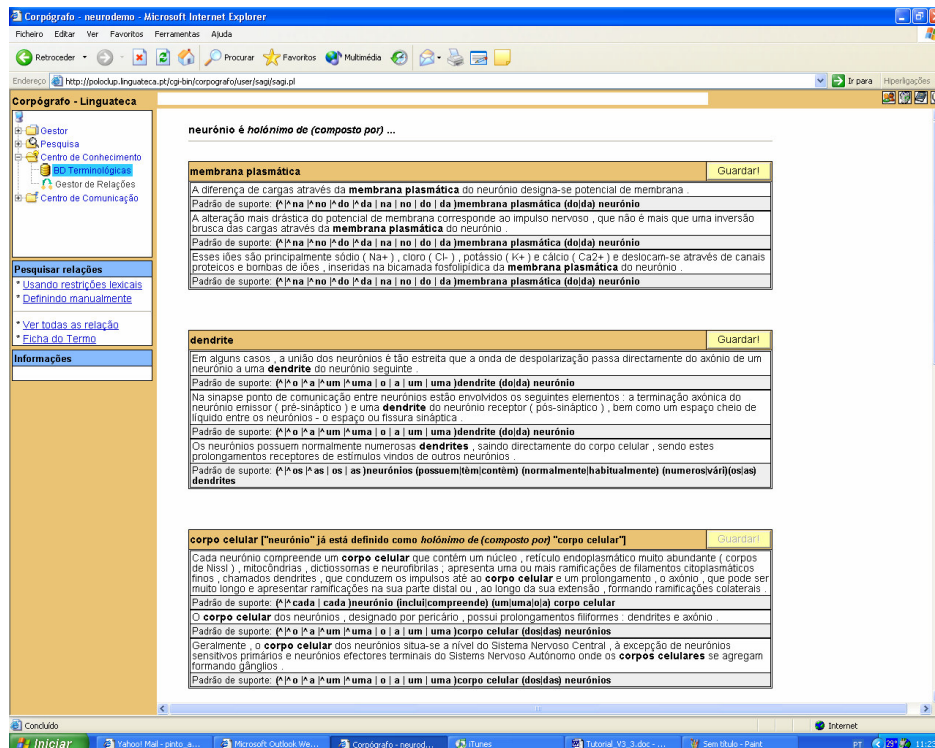


Fig. 14 – Resultados de uma Pesquisa de Relações Semânticas com restrições

No caso de saber *a priori* que dois termos presentes na base de dados estão relacionados por uma das relações semânticas previstas pelo Corpógrafo, terá também a possibilidade de associar ao termo que estiver a editar um termo associado. No menu **Pesquisar Relações**, clique em **Definindo manualmente**, defina o tipo de relação a estabelecer entre os termos e seleccione o termo a relacionar com o termo que se encontra a editar, clicando, em seguida, em **Relacionar!**.

Poderá ainda relacionar termos através de etiquetas semânticas definidas por si (ver **Gestor de Relações**), mas também aqui terá de saber *a priori* que os termos estão relacionados pela etiqueta que tiver definido.

5.5 Gestor de Relações

É no **Gestor de Relações** que poderá criar as suas próprias etiquetas semânticas para relacionar termos. Se quiser associar dois termos que estão ligados por uma relação semântica não prevista pelo Corpógrafo, pode fazê-lo criando no Gestor de Relações uma etiqueta para essa relação.

Imaginemos que o Corpógrafo não previa a relação de Hiperonímia/Hiponímia e que queria inserir essa etiqueta semântica. No Gestor de Relações, clique em **Criar nova Etiqueta** no menu **Etiquetas Semânticas**. Dê um nome à relação, neste caso “Hiperonímia/Hiponímia”, e digite uma sigla que represente essa relação (ex: Hiper/Hipo). Em seguida insira um descritivo e uma sigla para o papel no sentido directo e para o papel no sentido inverso. Neste caso, o papel no sentido directo seria o de Hiperónimo, pelo que o descritivo poderia ser “*é uma generalização de*”, e a sigla poderia ser HIPER. O papel no sentido inverso seria o de Hipónimo, pelo que o descritivo poderia ser “*é uma especialização de*”, e a sigla poderia ser HIPO. Se desejar, poderá ainda inserir uma pequena descrição da etiqueta semântica. Por fim, clique em **Criar Etiqueta**.

Criar Etiquetas Semânticas

Corpógrafo - Linguateca

Etiquetas Semânticas de Utilizador

A partir deste formulário poderá adicionar ao Corpógrafo a sua própria Etiqueta de Relação Semântica, que depois estará disponível para associar termos nas Bases de Dados Terminológicas.
Todos os campos são obrigatórios (excepto o campo Descrição).
O Corpógrafo só aceitará novas Etiquetas **se todos os dados obrigatórios forem preenchidos!**

Criar nova Etiqueta Semântica

1) Nome:

2) Sigla:

3) Papel no sentido directo:

3.1) Descritivo:

3.2) Sigla:

4) Papel no sentido inverso:

4.1) Descritivo:

4.2) Sigla:

5) Descrição:

Criar Etiqueta

Fig. 15 – Formulário para inserção de Etiquetas Semânticas

Depois de criadas as etiquetas, estas estarão à sua disposição nas ferramentas de Edição do Termo. Clique numa base de dados e em seguida liste os termos de acordo com os seus objectivos. Clique no termo que deseja relacionar através da etiqueta semântica que criou. No menu **Ficha do termo** clique em **Relações Semânticas** e depois em **Definindo manualmente**, no menu **Pesquisar relações**. Seleccione a etiqueta que criou e clique em **Relacionar!**.

6. Resumo

Com este tutorial foi possível acompanhar do início ao fim uma tarefa de extracção terminológica. Na realidade, este processo não tem verdadeiramente um fim, já que as bases de dados terminológicas poderão ser constantemente enriquecidas com novos termos e meta-informação mais completa, à medida que novos textos são adicionados ao corpus.

Actualmente, é já possível fazer um valioso trabalho de construção que pode melhorar significativamente sistemas de recolha e pesquisa de informação.