Gestor de Corpora – Um ambiente Web integrado para Linguística baseada em Corpora

Belinda Maia
Faculdade de Letras de Universidade do Porto
bmaia@mail.telepac.pt

Luís Sarmento
Linguateca - Pólo CLUP
las@letras.up.pt

Resumo

Neste artigo apresenta-se o sistema "Gestor de Corpora", um ambiente Web integrado para suporte a investigação Linguística baseada em corpora. Será feita uma análise dos problemas frequentemente associados a este tipo de investigações e serão discutidos alguns dos requisitos básicos para uma ferramenta deste género. Será também apresentado o estado de desenvolvimento actual do Gestor de Corpora e as linha de evolução futura.

Palavras-Chave

Corpora, Linguística Computacional, Ferramentas Web, Gestor de Corpora.

1. Introdução

A utilização de Corpora tem vindo progressivamente generalizar-se nas áreas da Linguística, do Processamento de Linguagem Natural e da Tradução. Nos últimos anos tem sido realizado um grande esforço na criação de grandes corpora para fins genéricos (ex: CETEMPúblico[3], BNC[5]), constituídos por vários milhões de palavras, e na sua posterior disponibilização, tanto em formato CD-ROM como através de ferramentas de acesso via Web. Actualmente, pode considerar-se que existe já um número satisfatório de corpus genéricos disponíveis, facto que muito tem contribuído para o desenvolvimento da utilização de corpora.

Por outro lado, é também reconhecido o valor prático dos corpora mais específicos, em particular para suporte a pesquisa terminológica, actividades de tradução e mesmo actividades de avaliação. Contudo, essa mesma especificidade torna-se um obstáculo para a existência de uma oferta abundante deste tipo de corpra que, por esse motivo, necessitam quase sempre de ser construídos à medida para utilização em tarefas específicas. Muito frequentemente, estes corpora são utilizados apenas por um período de tempo muito reduzido (ex.: apoio a uma tradução) o que inviabiliza a possibilidade de investir demasiado esforço na sua construção (noção de *"Do it yourself Corpora"* [2]). Igualmente frequente é o facto de o esforço efectivamente dispendido não ser posteriormente utilizado por outros utilizadores (ex: resultados da pesquisa terminológica sobre o corpus construído) o que representa uma infeliz perda de recursos.

2. O Gestor de Corpora

No sentido de colmatar esta reduzida oferta e ao mesmo tempo permitir a reutilização máxima de recursos, o Pólo do Porto da Linguateca tem vindo a desenvolver uma ferramenta Web denominada "Gestor de Corpora" (GC). O GC tem como objectivo auxiliar a rápida construção de corpora específico assim como a sua posterior pesquisa e disponibilização.

O GC abrevia e simplifica todo um conjunto de actividades associadas à construção de corpora (extracção e pré-processamento de texto, codificação do corpus) permitindo a utilizadores sem conhecimentos de programação a criação transparente do seu próprio corpus pessoal. Os corpora criados utilizando GC tornam-se automaticamente pesquisáveis por expressões regulares sendo imediatamente possível realizar vários tipos de estudos de frequência e colocação.

À data desta comunicação, o GC encontra-se já equipado com um pequeno conjunto de ferramentas específicas para além das obrigatoriamente necessárias para a preparação e pesquisa de corpus: i) um módulo para suporte a estudos estatísticos; e ii) um módulo destinado a à pesquisa e armazenamento de candidatos terminológicos.

A arquitectura do GC permite a adição de novas funcionalidades de uma forma modular pelo que encontram-se já em estudo e desenvolvimento novos módulos para tarefas de anotação corpora bem como para auxiliar a construção de corpus paralelos e comparáveis e a respectiva pesquisa. O GC encontra-se em rápido desenvolvimento e espera-se que possa vir a tornar-se uma ferramenta útil para o maior número de utilizadores possível.

3. Utilizadores e Problemas Frequentes

Trabalhar com corpora é, para grande parte dos seus utilizadores, um actividade onde diversos pequenos problemas dificultam um fluxo de trabalho rápido e eficiente. A construção de uma ferramenta integrada para a utilização de corpora pressupõe desde logo o conhecimento e compreensão destes problemas, cuja natureza é radicalmente diferente para os diversos segmentos de utilizadores. Comecemos então por diferenciar os eventuais utilizadores de uma ferramenta de corpora em 3 segmentos bastante abrangentes em função da sua proveniência e objectivos. Esta divisão não pretende ser completa já que assumidamente não inclui todos os tipos de utilizadores existentes. Os 3 segmentos que iremos descrever de seguida referem-se apenas aos utilizadores com os quais o Pólo CLUP da Linguateca tem contactado com mais intensidade tendo constituindo o ponto de partida para a posterior análise de requisitos.

3.1 Utilizadores da Área Linguística

Num primeiro segmento podemos incluir utilizadores com formação na área Linguística. Estes utilizadores têm, evidentemente, conhecimentos de Linguística profundos mas nem sempre têm experiência da utilização de corpora nem das respectivas ferramentas, apesar de compreenderem o seu princípio de funcionamento. Na prática, à excepção dos utilizadores mais próximos da Linguística Computacional, estes utilizadores esbarram frequentemente em barreiras tecnológicas e logísticas que dificultam seriamente a utilização de corpora e desmotivam futuras tentativas. Problemas como a falta de conhecimento genéricos a nível da informática, dificuldades na obtenção e utilização de software apropriado acabam por gerar alguma fobia à utilização de corpora. Ainda assim, há neste segmento quem vá compilando corpora e utilize software como o

WordSmith Tools[4] para estudos de frequência e colocação. O principal objectivo deste utilizadores na utilização de corpora consiste no estudo de um determinado fenómeno linguístico pelo que as funcionalidades mais pretendidas numa ferramenta de acesso a corpora são a pesquisa de concordâncias e o respectivo tratamento estatístico.[1] Na maior parte dos casos as pesquisas pretendidas exigem algum tipo de anotação pelo que é também importante para estes utilizadores que uma ferramenta disponibilize suporte à anotação dos corpora a pesquisar.

3.2 Utilizadores da Área Tradução

Num segundo segmento iremos incluir utilizadores com formação na área da Tradução. Estes utilizadores, possuem normalmente bons conhecimentos informáticos na óptica do utilizador já que tradicionalmente trabalham com vários pacotes de software especializado na área da tradução. No entanto, este utilizadores nem sempre estão por dentro dos conceitos associados a corpora nem aos procedimentos relativos à sua compilação/preparação e às respectivas ferramentas. Há contudo uma fracção significativa de utilizadores que faz uso de corpora para a pesquisa terminológica e a construção de glossários pessoais. Esta é sem dúvida uma das principais funcionalidades que utilizadores deste segmento procuram em ferramentas de acesso a corpora.

Uma situação problemática que se verifica com muita frequência neste segmento de utilizadores é a dificuldade na preparação de corpora, nomeadamente na extracção de texto dos documentos compilados (a maioria dos quais em formato PDF, PostScript ou HTML) para a sua utilização em ferramentas que normalmente apenas lidam com texto em formato ASCII. Esta contrariedade, apesar de tecnicamente ser bastante simples de resolver, reduz significativamente a quantidade de texto que estes utilizadores empregam na criação de corpora reduzindo consequentemente a eficácia da pesquisa terminológica.

Na área da tradução revela-se também particularmente importante o acesso a corpora bilingue paralelo (e também comparável) cuja utilização permite resolver vários problemas práticos de tradução de uma forma semelhante às *memorias de tradução*. No entanto, a construção deste tipo de recursos exige normalmente um esforço muito elevado que se encontra para lá da disponibilidade da maior parte dos utilizadores. Uma ferramenta que auxilie os utilizadores em tarefas de pré-processamento e alinhamento revela-se de grande utilidade.

3.3 Utilizadores na área do PLN

Decidimos incluir nesta segmentação utilizadores provenientes da área do PLN. Sem dúvida que os interesses na utilização de corpora para os utilizadores deste segmento são abundantes e reconhecidos pelo que não iremos listar extensivamente as possibilidades. Achamos importante realçar que, cada vez mais, a investigação nestas áreas é feita por equipas multidisciplinares cujos elementos provêm não só das áreas mais tradicionais do PLN (Engenharia, Ciências de Computação e Linguística) mas também por elementos provenientes das mais variadas áreas e que trazem conhecimento pericial sobre domínios específicos. Neste contexto, e pensando nas necessidades próprias do trabalho em equipa torna-se necessário que as ferramentas de acesso a corpora potenciem o trabalho em grupo e permitam o intercambio e exportação de dados em formatos estruturados e estandardizados, quer dentro da equipa quer entre aplicações.

4. Lista de Requisitos

Depois desta breve análise das dificuldades encontradas por diferentes grupos de utilizadores e das respectivas expectativas relativamente a uma ferramenta para corpora, iremos agora realizar uma breve listagem dos requisitos que serviram de base à especificação do Gestor de Corpora.

4.1 A Arquitectura Global

Há vários factores em jogo na escolha de uma Arquitectura apropriada para o Gestor de Corpora. Em primeiro lugar devem ser preferidas soluções que não exijam por parte do utilizador grandes conhecimentos de administração informática. A experiência tem verificado que a instalação de software, em particular em grandes organizações como Faculdades, por muito simples que possa parecer, arrasta consigo demasiados problemas (permissões de instalação, dificuldade de manutenção, restrições de segurança, etc.) para os quais o utilizador comum nem sempre sabe ou tem possibilidades de encontrar a solução apropriada. Adicionalmente, a frequente necessidade de trabalhar em equipa em certas tarefas de preparação dos corpora, tais como a anotação ou alinhamento de corpora (ou apenas a sua verificação), sugere uma solução que envolva desde logo a utilização de uma infra-estrutura de comunicação em rede.

Esta duas condições apontam para que uma ferramenta de corpora seja construída recorrendo a uma Arquitectura centralizada do tipo Cliente-Servidor, preferencialmente através de uma aplicação de acesso Web. Desta forma, resolvem-se todos os problemas de administração garantindo-se automaticamente a possibilidade de comunicação e partilha de dados entre os utilizadores. Por outro lado, soluções centralizadas trazem outro tipo de problemas nomeadamente problemas de tolerância a falhas e escalabilidade. Relativamente ao primeiro, as soluções habituais envolvem normalmente a inclusão de redundância no sistema (ex: RAID/clustering). Relativamente aos problemas de escalabilidade, consideramos que só a longo prazo se poderá revelar crítico, i.e. quando o número de utilizadores simultâneo for muito grande, já que o tamanho do corpora pessoal será quase sempre relativamente reduzido (<< 1 Milhão de palavras). A solução para os eventuais problemas de escalabilidade consistirá transferir parte do processamento desde o servidor Web para o cliente, o que poderá ser realizado convenientemente utilizando a tecnologia Java.

4.2 Requisitos de Pré-processamento

Tendo em vista as tarefas de preparação e pré-processamento de corpora, a ferramenta implementada deverá possuir as seguintes funcionalidades básicas:

1. importação de Ficheiros de Texto em vários formatos – PDF, Ms-Word, RTF, PS, HTML,TXT;
2. edição do texto importado para correcções e "limpeza";
3. agrupamento e divisão do texto em frases;
4. codificação de texto num formato apropriado para a realização de pesquisas;

[1] Exclui-se aqui grande parte de outras funcionalidades que são bastante úteis a investigadores mais próximos da Linguística Computacional.

5. organização de Corpus mediante uma taxinomia de domínios, géneros e registos.

Esta podem ser consideradas as funcionalidades mínimas e que dão suporte a pesquisas simples e estudos de frequência. Seria também interessante disponibilizar algumas outras funcionalidades associadas tais como a exportação de dados em formatos diversos para utilização noutras ferramentas (XML, txt, Excel).

No sentido de permitir pesquisas mais complexas é também importante fornecer ao utilizador mecanismos de anotação morfo-sintáctica, quer do tipo automático quer do tipo semi-automático. Consideramos que a utilização de mecanismos semi-automáticos poderá ser preferível já que a dimensão dos corpora pessoais não será certamente demasiado grande ao ponto de tornar a intervenção do utilizador excessivamente trabalhosa. Por outro lado recorrendo a mecanismos semi-automáticos obtém-se normalmente uma maior precisão no processo.

A utilização de corpora alinhado irá também exigir a disponibilização de ferramentas de alinhamento. Novamente aqui se considera que a adopção de mecanismos semi-automáticos poderá ser mais apropriada, pelas mesma razões referidas anteriormente.

Quanto à utilização de corpora comparável, o processo deverá ser muito mais simples por parte do utilizador cuja única intervenção se prenderá com a associação dos textos comparáveis.

4.3 Requisitos de Pesquisa e Estudo

Relativamente às funcionalidades de pesquisa poderemos considerar que os requisitos mínimos para uma ferramenta deste género incluem, pesquisa por palavra chave, pesquisa por expressão regular e a pesquisa de colocações em corpus simples. Adicionalmente deverá ser possível fazer estudos estatísticos de frequência de concordâncias, colocações e N-gramas. Estas funcionalidades encontram-se ao nível de ferramentas como o WordSmith e na nossa opinião devem ser consideradas como requisitos mínimos para qualquer ferramenta de corpora.

Tendo em conta as necessidades dos segmentos de utilizadores que apresentamos a ferramenta deverá também permitir:

- pesquisas por estruturas morfo-sintácticas sobre os corpora anotados;
- pesquisa em corpora paralelo e comparável;
- extracção de terminologia monolingue sobre corpora simples e a pesquisa de terminologia bilingue em corpora comparável.

Além destes pesquisas torna-se cada vez mais indispensável conjugar informação local com o mega-corpus que é a Web. Por isso uma ferramenta de pesquisa deverá ser capaz de estabelecer ligação à Web para recolha de corpora em bruto ou de informação mais específica (ex: testes frequência relativa de termos).

4.4 Outros Recursos Associados

A quase totalidade das ferramentas que operam sobre corpus e que se encontram mais facilmente disponíveis não incluem recursos que permitam optimizar a pesquisa sobre corpora em Língua Portuguesa. Esta limitação advém do facto de as referidas ferramentas serem construídas por investigadores cuja Língua nativa não é o Português. Por esta razão parece-nos importante incluir numa ferramenta cujo público alvo é essencialmente Lusófono recursos de Língua Portuguesa.

A este nível o recurso elementar é sem dúvida um dicionário de Português. Um outro recurso útil será por exemplo uma lista contendo o nome das localidades Portuguesas e Brasileiras ou a algumas localidades mundiais expressas na Língua original e em Português. Estes recursos representam conhecimento importante que pode ser utilizado na optimização de pesquisas sobre corpora. Adicionalmente, a utilização de dicionários bilingue centrados no Português torna-se de grande valor em pesquisa terminológica bilingues e em pesquisas sobre corpora comparáveis bilingues.

Finalmente, como forma de garantir a reutilização da informação recolhida pelos utilizadores a partir as pesquisas efectuadas, verifica-se ser de grande utilidade associar à ferramenta de pesquisa um sistema de bases de dados. A ligação de uma base de dados à ferramenta permite a recolher de uma forma organizada elementos tais como entradas terminológicas e as respectivas definições, associações entre termos, nomes de entidades, etc.

5. O Gestor de Corpora

Após esta breve análise do público alvo ao qual se pretende fornecer uma ferramenta de corpora e da listagem de alguns dos requisitos essenciais aos quais deve obedecer tal ferramenta passamos a apresentação concreta do Gestor de Corpora.

O Gestor de Corpora surge como o resultado da compilação de algumas ferramentas que tem vindo a ser desenvolvidas no Pólo do Porto da Linguateca para dar resposta aos problemas que os utilizadores ai localizados (maioritariamente alunos e professores de Tradução e de Linguística) encontram durante as seus trabalhos e pesquisas, sendo por isso um sistema orientado para a resolução de problemas práticos e frequentes.

A arquitectura geral do GC encontra-se apresentada na Figura 1 e tem como elemento central o sistema a que decidimos denominar por ToolPool.

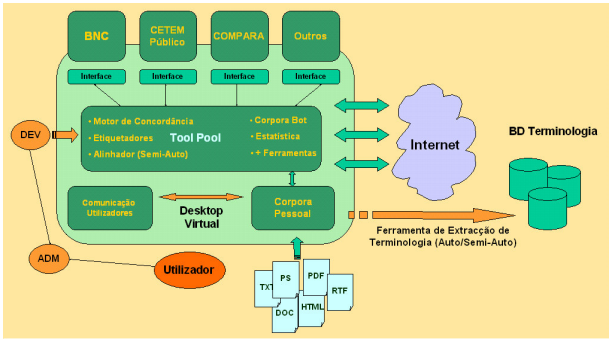

Figura 1 – A arquitectura geral do Gestor de Corpora

O ToolPool consiste num conjunto de módulos que executam tarefas específicas sobre os corpora dos utilizadores e acompanham o utilizador ao longo de todo o processo de utilização de corpora, ou seja desde a preparação dos corpora até à respectiva pesquisa e geração de resultados.

Parte dos módulos do ToolPool são desenvolvidos de raíz (em Perl e Java) enquanto que outros representam interfaces que desenvolvemos para permitir a interacção com outras

ferramentas já existentes e cuja integração é normalmente difícil ou exige conhecimentos informáticos ao nível de programação. Desta forma o ToolPool garante um fluxo de trabalho integrado permitindo a interacção do utilizador com várias ferramentas de uma forma transparente. Por exemplo, actualmente estamos a utilizar no GC as ferramentas IMS-CWB[1] como motor de pesquisa sobre corpora sendo que todo o processo de codificação necessário é feito por módulos próprios de integração e é totalmente transparente ao utilizador.

O ToolPool garante igualmente o desenvolvimento modular do GC. Á medida que surgem novas necessidades, novas aplicações ou algoritmos, poderão ser adicionados novos módulos ao ToolPool

6. Três Perspectivas sobre o GC

O GC pode ser visto de 3 perspectivas diferentes relativamente às pessoas que com ele interagem: a do utilizador/pesquisador, a do Administrador do Sistema e a do Programador.

6.1 O Utilizador

Comecemos por analisar a perspectiva do utilizador/pesquisador. Para o utilizador o Gestor de Corpora pode ser considerado um "desktop" virtual ou uma mesa de trabalho onde pode utilizar certas ferramentas sobre a sua matéria prima: ficheiros de texto em vários formatos.

Toda a interacção se faz através da Web e os corpora do utilizador ficam armazenados no servidor (sujeitos a uma determinada limitação de quota). Através do interface Web do GC, e após a devida autenticação, o utilizador tem possibilidade de carregar os ficheiros em formato *PDF*, *PostScript*, *html*, *rtf*, *Ms-Word* e *txt* que possui para a sua área pessoal no servidor. O texto existente nesses ficheiros é imediatamente extraído (usando o módulo Perl ExTex.pm desenvolvido para o efeito e que serve de interface a vários utilitários de extracção de texto) e fica armazenado no servidor. Nesta versão do GC o documento original não é guardado embora se reconheça que essa seja uma funcionalidade interessante para a organização do trabalho do próprio utilizador.

O texto extraído dos ficheiro encontra-se frequentemente fragmentado sendo por isso impróprio para pesquisas. O GC proporciona aos utilizadores uma ferramentas de edição onde o utilizador pode executar operações de "limpeza" do texto, assim como executar semi-automaticamente a sua divisão em frases. Após este processo de pré-processamento assistido, o ficheiro de texto encontra-se num formato apropriado para a codificação no formato XML que será depois codificado pelas ferramentas IMS-CWB. Contudo, a codificação do texto não é feita ainda neste momento. De facto, no GC existe a noção de Selecção de Ficheiros que consiste num agrupamento de ficheiros definido pelo utilizador que constituem a base mínima para a codificação. Assim, o utilizador tem a possibilidade de definir as Selecções de Ficheiros que de facto irão constituir o seu corpus pesquisável, sendo o processo de codificação do texto (em formato IMS-CWB) realizado só após a definição da Selecção. A gestão de toda a informação relativa aos corpora do utilizador é feita utilizado um módulo Perl que permite a criação de pequenas Bases de Dados em ficheiros de Texto. Foi tomada a opção a opção de este módulo em detrimento de um SGBD regular para evitar adicionais transtornos de administração.

Depois da codificação, que ocorre de uma forma totalmente transparente, o utilizador poderá fazer vários tipos de pesquisa sobre a Selecção usando um interface próprio. Á data desta comunicação o utilizador tem a possibilidade de executar pesquisa por expressões regulares, e fazer estudos estatísticos de palavras chave e colocações.

Estão a ser desenvolvidas novas funcionalidades que passarão, para além de pesquisas mais complexas suportadas pelo IMS-CWB, por módulos próprios de extracção de terminologia e por possibilidades de realizar pesquisas sobre corpus comparáveis. Está também planeado o desenvolvimento de uma interface (applet Java) destinada a permitir o alinhamento semi-automático do textos para dar suporte à construção de corpora paralelos.

Uma última nota para um outro módulo que se encontra já em fase inicial de implementação: Corpora Bot, um softbot destinado a facilitar a pesquisa de corpora temático. O Corpora Bot pesquisa a Web através de alguns portais e directórios Web usando palavras chave fornecidas pelo utilizador, e atravessa os sítios encontrados descarregando para a área do utilizador documentos de texto neles contidos. O Corpora Bot têm preferência por documentos do tipo *PDF*, *Ms-Word* e *PostScript*, e documentos *html* contendo grande quantidade de texto. As capacidade de pesquisa de Corpora Bot serão aumentadas assim que seja possível incluir a informação terminológica recolhida pelo utilizador.

6.2 O Administrador

O Gestor de Corpora possui um interface de administração que permite a um utilizador privilegiado a gestão de alto nível de utilizadores e dos recursos que lhe estão associados.

No que diz respeito à gestão de utilizadores, o GC permite a executar as tarefas básicas tradicionais tais como criação de novos utilizadores no sistema e a respectiva remoção e a definição dos respectivos privilégios relativamente a bases de dados, quota de disco e acesso a corpora restrito (ex.: BNC).

Uma outra funcionalidade interessante que fica a cargo do Administrador é a criação de grupos de utilizadores que desta forma ficarão habilitados a partilhar corpora e bases de dados, bem como a efectuar algumas tarefas em conjunto (anotação – ainda não implementada).

Finalmente o Administrador está também incumbido de gerir e construir a base de dados que contem a taxinomia de domínios, géneros e registos, com os quais os utilizadores poderão categorizar os seus corpora.

6.3 O Programador

Também o programador possui uma perspectiva própria sobre o GC. Apesar de não existir uma API no verdadeiro sentido do termo, o GC é implementado segundo uma metodologia OO, recorrendo a módulos Perl (e futuramente a classes Java) que fornecem ao programador uma forma modular de estender o GC. Com o alargamento do GC a novos programadores espera-se conseguir desenvolver uma API para a criação rápida de funcionalidades específicas.

7. Desenvolvimentos Futuros

O Gestor de Corpora está ainda em desenvolvimento. Apesar de já terem sido concluídas as infra-estruturas básicas, há ainda muito trabalho de fundo a realizar. Como pontos principais para desenvolvimento futuro destacamos:

1. Sofisticação dos Algoritmos de Extracção de Terminologia;
2. Criação de Módulos para a extracção Semi-Automática de definições;
3. Interface de pesquisa sobre corpora comparáveis;
4. Conclusão da Implementação do Corpora Bot.

Certamente que durante a aplicação prática do Gestor de Corpora, outras necessidades irão surgir para as quais pretendemos dar resposta. Neste momento, vários alunos de mestrado e doutoramento da FLUP estão a começar a utilizar o GC e esperamos poder recolher destes utilizadores informações práticas que nos permitam orientar o desenvolvimento futuro.

8. Conclusão

Nesta comunicação realizou-se a análise do contexto geral de utilização de corpora, em particular no ambiente da Faculdade de Letras da Universidade do Porto. Foram caracterizados alguns dos principais utilizadores de corpora, foram elencadas as suas necessidades típicas relativamente a corpora e o tipo de problemas com os quais esses utilizadores mais frequentemente se debatem. Foi também realizada uma breve análise de requisitos para uma ferramenta de apoio a estudos sobre corpora que pudesse dar reposta aos referidos problemas. Em seguida apresentamos o Gestor de Corpora, a nossa implementação de uma ferramenta com tais objectivos. Foi apresentado o estado actual do desenvolvimento do Gestor de Corpora e apresentadas algumas linhas para o desenvolvimento futuro.

Bibliografia

[1] Christ O.:"A modular and flexible architecture for an integrated corpus query system". *COMPLEX'94, Budapest, 1994*

[2] Maia, B "Do-it-yourself corpora ... with a little bit of help from your friends". In Lewandowska-Tomaszczyk, B, Melia, P.J (eds.) pp. 403-410, 1997

[3] Rocha P., Santos D. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", in Maria das Graças Volpe Nunes (ed.), Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000) (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), pp. 131-140.

[4] Scott, M. WordSmith Tools – Software Package Oxford University Press.

[5] The British National Corpus - http://www.natcorp.ox.ac.uk/