UAM CorpusTool

Guia do Usuário Versão 2.6

(Dezembro 2010)

Mick O'Donnell

michael.odonnell@uam.es

(Traduzido por Mário Martins)

Índice

Seção 1:

UAM CorpusTool - Visão Geral

1 Introdução

O UAM CorpusTool, ou simplesmente CorpusTool, é um conjunto de ferramentas criadas para a anotação linguística de textos. Com o CorpusTool:

- você pode constituir um *corpus*, que é um conjunto de arquivos de texto, e também definir quais esquemas de análise serão aplicados a esses arquivos.
- cada análise pode ser representada por uma camada distinta de anotação. O CorpusTool atualmente suporta dois tipos de anotação:
  1. **Anotação do Documento**: em que determinadas características de anotação são aplicadas ao texto como um todo. Essas características, por exemplo, podem se referir ao registro do texto (campo, relações e modo) ou ao tipo do texto (narrativo, argumentativo, etc.).
  2. **Anotação de Segmentos**: em que determinadas características de anotação são aplicadas a segmentos dentro do texto. Para especificar a extensão de um segmento, basta posicionar o cursor do mouse em um ponto inicial, segurá-lo e arrastá-lo até um ponto final. A partir daí, as características de anotação previamente definidas tornam-se disponíveis.

Outros tipos de anotação serão possíveis em versões futuras, o que permitirá a anotação de estruturas retóricas (RST), estruturas genológicas (GSP), cadeias de participantes, estrutura oracional (Sujeito, Predicador, Modo Oracional, Adjunto, etc.), textos falados, etc.

O CorpusTool substitui plenamente o programa Systemic Coder, do mesmo autor. O Systemic Coder foi criado para a anotação de documentos simples utilizando uma única camada de anotação. O CorpusTool é uma tentativa de superar as várias limitações sentidas pelos usuários daquele primeiro programa. Agradeço aos muitos usuários do Systemic Coder que me enviaram comentários ao longo desses anos. Agradeço também a todos aqueles que me enviem comentários sobre esta nova ferramenta.

Para saber como importar projetos criados no Systemic Coder para o CorpusTool, consulte o Apêndice I.

O CorpusTool está disponível em:

http://www.wagsoft.com/CorpusTool/

Nesse site, encontram-se as instruções para instalar o CorpusTool no seu computador.

Seção 2:

Projeto

2 Iniciando um Novo Projeto

2.1 Abrindo o CorpusTool

Depois de instalado, você já pode começar a trabalhar com o CorpusTool. O primeiro passo é criar um novo projeto. Para isso, abra o programa:

Windows:

- Faça duplo clique no ícone do CorpusTool que está localizado na área de trabalho, criado durante a instalação.
- Alternativamente, você pode clicar no ícone do CorpusTool que está na lista de programas do menu Iniciar (Iniciar > Programas > CorpusTool).

Macintosh:

- Faça duplo clique no ícone do CorpusTool que está localizado na pasta Aplicações.
- Para facilitar o acesso ao programa, você pode colocar um ícone do CorpusTool na sua *Dock*.

Para abrir um projeto já criado, basta fazer duplo clique no arquivo .cptr que está localizado na pasta *Project*. Este arquivo é representado por um dos ícones abaixo:

MacOSX: 

Windows: 

A Janela de Abertura

Uma janela, como na Figura 2.1, se abre. Essa janela apresenta, entre outras informações, o número da versão do CorpusTool em uso, o que pode ser bastante útil para comunicar falhas (*bugs*). Além disso, essa janela oferece duas opções de ação: *Start New Project*, para criar um novo projeto, ou *Open Project*, para continuar a trabalhar num projeto anteriormente criado. Neste último caso, e se você estiver trabalhando no mesmo computador, aparece uma terceira opção: *Open Last Project*, que abre o projeto trabalhado mais recentemente.

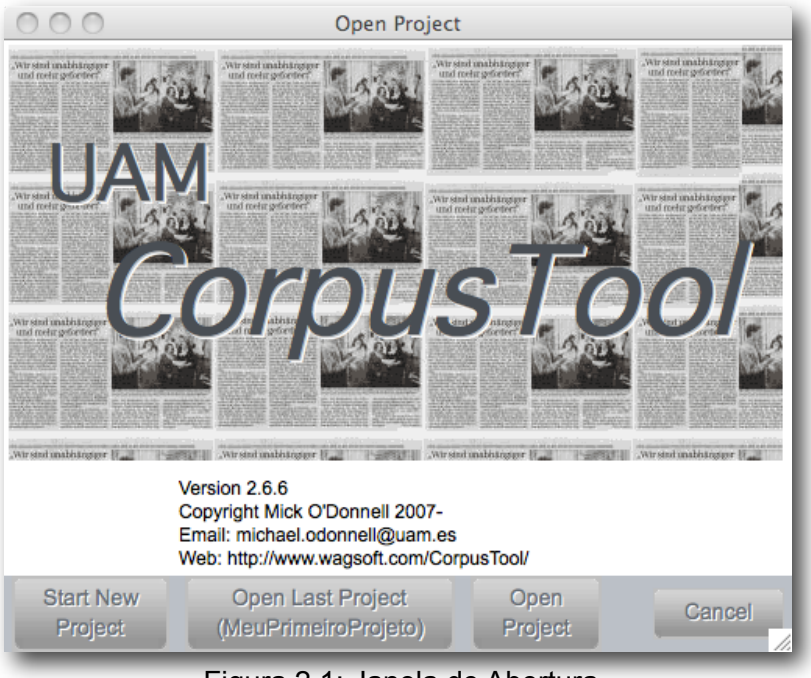


Figura 2.1: Janela de Abertura

2.2 *Iniciando um Novo Projeto*

Ao clicar em *Start New Project*, um assistente de criação de projeto (*Create Project Wizard*) aparece para conduzi-lo através dos passos necessários à criação de um novo projeto:

1. Atribua um nome ao novo projeto. Escolha, por exemplo, “MeuPrimeiroProjeto”.
2. Especifique o destino onde a pasta com o novo projeto deve ser armazenada. Escolha, por exemplo, a área de trabalho do seu computador.

Ao clicar no botão *Finalise*, o CorpusTool cria o projeto, que se constitui de uma pasta (com o mesmo nome do projeto) contendo o *corpus* e os arquivos de anotação. Nessa pasta, também se encontra um ícone do CorpusTool (extensão .cptr), o qual pode ser usado para abrir seu projeto diretamente.

Tendo sido criado o projeto, a janela principal do CorpusTool se abre, mostrando o painel de gerenciamento de projetos (Figura 2.2). A partir desse painel, você pode controlar os detalhes do seu projeto, tais como os arquivos a serem incluídos e o tipo de análise envolvida.

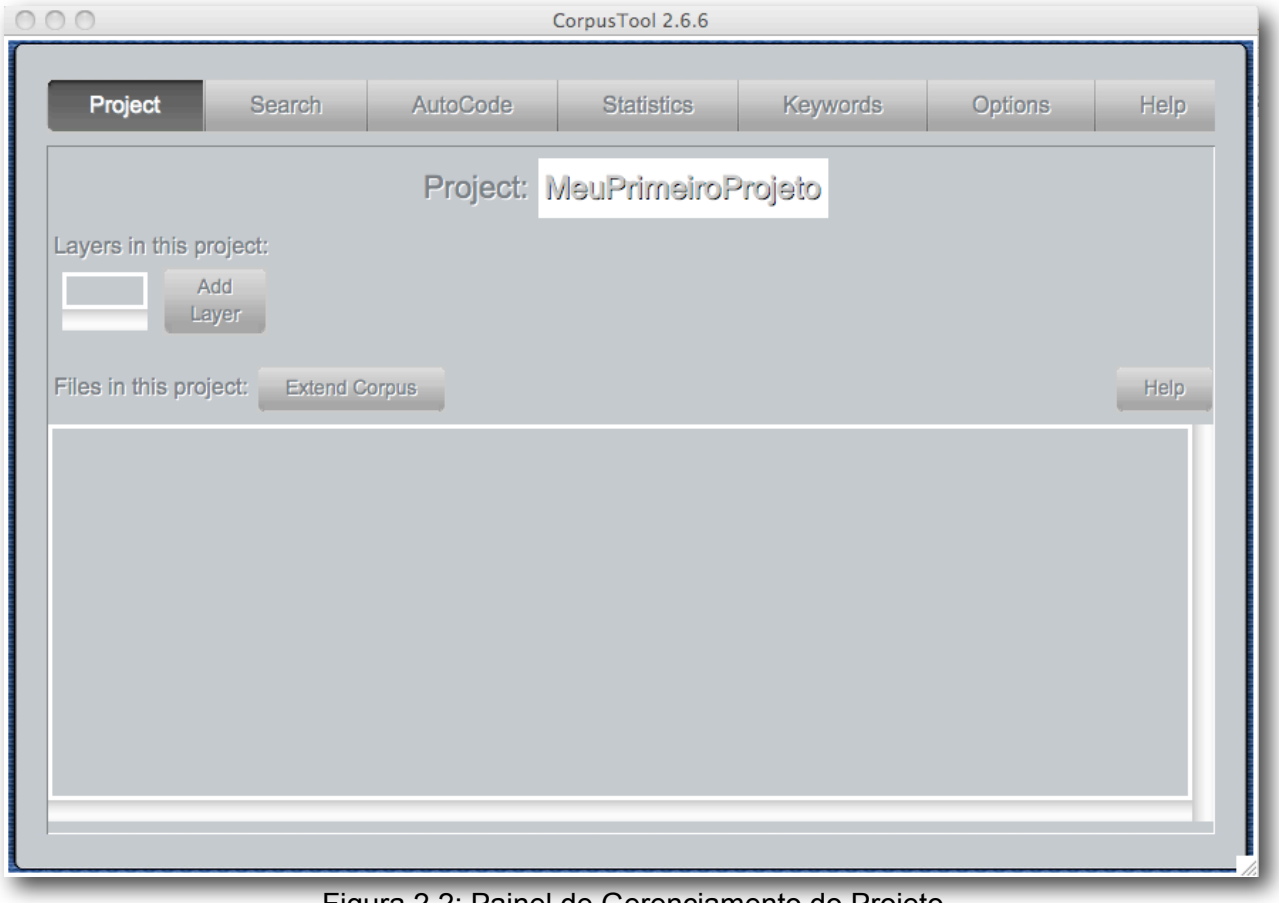


Figura 2.2: Painel de Gerenciamento do Projeto

Os botões no topo do painel permitem ao usuário movimentar-se entre os diferentes painéis do CorpusTool: ***Project*** (Projeto - esta seção), ***Search*** (Pesquisa - seção 5), ***Autocode*** (Anotação Automática - seção 6), **Statistics** (Informações Estatísticas - seção 7), ***Keywords*** (Palavras-chave - seção 8) e ***Options*** (Opções) e ***Help*** (Ajuda).

As letras grandes em fundo branco, no topo, exibem o nome do projeto. Logo abaixo, está a área de camadas de anotação (*Layers in this project*), que por omissão está vazia. Abaixo da área de camadas, está a área em que são mostrados todos os arquivos no projeto (*Files in this project*) e que inicialmente também se encontra vazia.

Vamos primeiramente adicionar uma camada de anotação ao projeto.

3 Adicionando uma Camada de Anotação

Para adicionar uma camada de anotação, clique em *Add Layer*. Uma camada corresponde a um tipo específico de anotação aplicada a um texto. É possível adicionar camadas para anotar orações, grupos, registro e gênero de texto, etc.

Vamos começar por adicionar uma camada para anotar o registro (característica do texto como um todo).

Ao clicar em *Add Layer*, uma nova janela se abre. Responda às perguntas, clicando em *Next* para a próxima pergunta:

- *Layer Name*: é o nome que você atribui à camada de anotação. Para este tutorial, coloque “Registro”.
- *Coding Object*: aqui você indica se quer atribuir características de análise ao texto na sua totalidade (por exemplo, registro) ou se quer

atribuir características a segmentos dentro do texto (por exemplo, orações). Selecione a primeira opção (*Annotate Document*).

- *Coding Scheme*: aqui você define o esquema de anotação, que é a descrição das características de anotação que você quer atribuir ao texto. Há duas opções:
    i. *Create New Scheme*: em muitos casos, o usuário está interessado em criar seu próprio esquema de anotação, definindo as características de análise que quer ver aplicadas ao texto e organizando-as do modo que considere mais conveniente. O CorpusTool possui uma interface bastante simples e intuitiva que permite criar e modificar esses esquemas (Seção 3).
    ii. *Copy Existing Scheme*: em outros casos, o usuário pode querer utilizar um esquema de anotação previamente criado, seja pelo próprio usuário, seja por outros investigadores. O CorpusTool contém alguns esquemas predefinidos, como o esquema de avaliação (*Appraisal*), desenvolvido por Peter White, e o esquema de anotação de erros (*Error Annotation*), desenvolvido por Sylviane Granger.

Para acompanhar este guia, selecione *Create New Scheme*. Depois, clique em *Finalise*. Uma nova camada de anotação é adicionada ao projeto.

A Figura 2.3 apresenta a janela *Project* com uma camada de anotação adicionada. Nesse espaço, há algumas informações sobre a camada de anotação: nome (Registro), tipo (*code-document*) e o nome do esquema associado à camada de anotação (Registro.xml). No painel de controle da camada de anotação, há dois botões:

- *Delete*: ao clicar neste botão, a camada e todas as anotações realizadas com base nesta camada são apagadas. Essa ação é irreversível, portanto, antes de apertar em *Delete*, você deve ter a certeza de que é isso mesmo que pretende fazer.
- *Edit*: ao clicar neste botão, uma nova janela se abre. Nela, você pode editar o esquema de anotação. Esse tópico será desenvolvido na próxima seção.

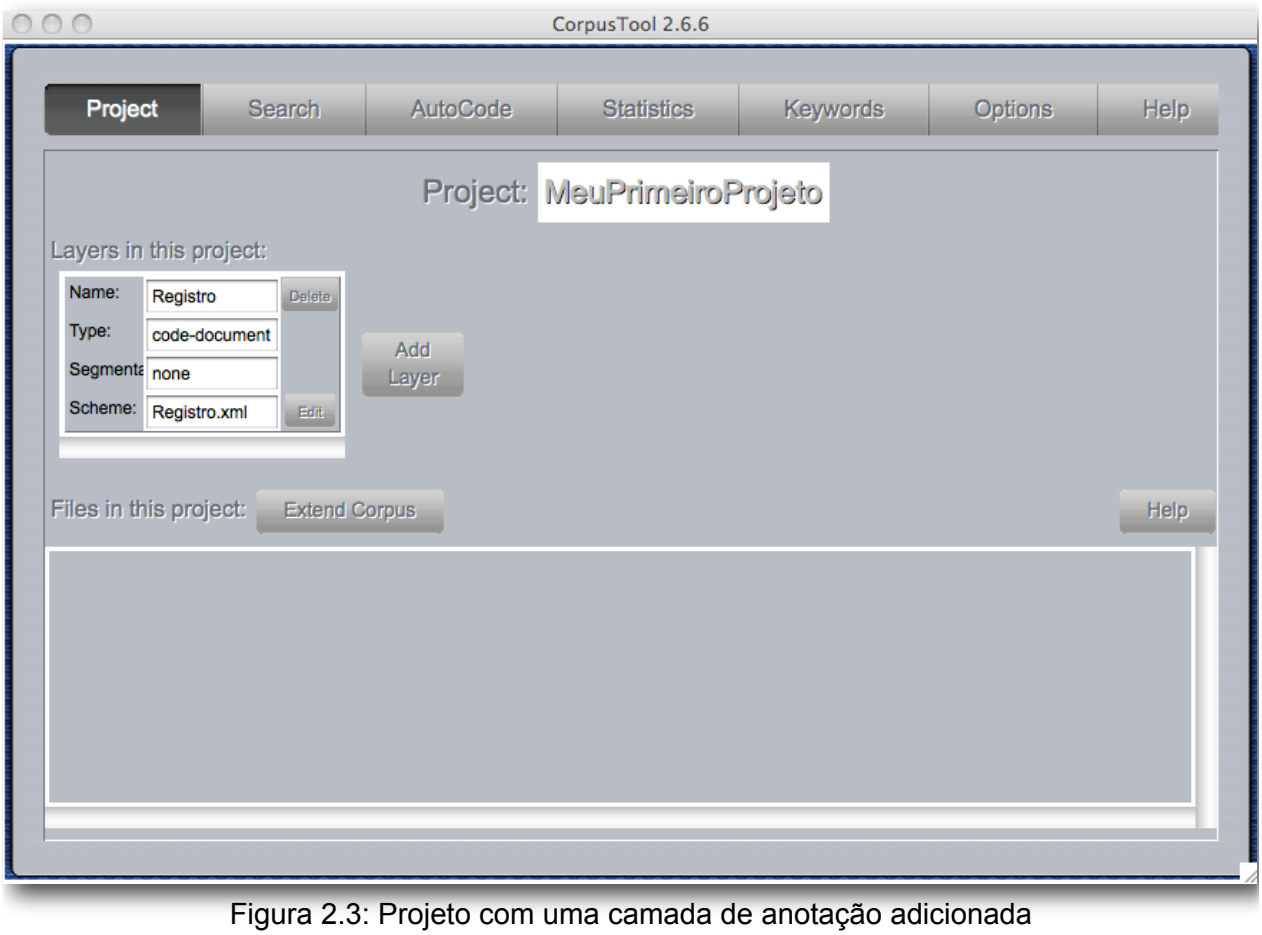


Figura 2.3: Projeto com uma camada de anotação adicionada

As camadas de anotação criadas no Systemic Coder podem ser importadas para o CorpusTool. Para mais detalhes, consulte o Apêndice I.

4 Adicionando Arquivos de Texto

O próximo passo é adicionar alguns arquivos de texto ao seu projeto. Durante o processo de criação de um projeto, há a possibilidade de indicar os arquivos de texto com que você vai trabalhar. Neste caso, eles devem aparecer na área de arquivos de texto. Para este guia, vamos assumir que você não selecionou arquivos de texto, assim a área de arquivos de texto deve estar em branco, como na Figura 2.3.

4.1 Aumentando o Corpus

Para adicionar arquivos de texto ao *corpus*:

1. Clique no botão *Extend Corpus*: um assistente o guiará nesse processo. Você pode acrescentar tanto um único arquivo de texto como uma pasta com vários arquivos de texto. Se deseja adicionar apenas um arquivo de texto, você pode fazê-lo de dois modos: adicionando-o como parte de um *subcorpus* já existente (a pasta *Corpus* dentro da pasta *Project*); ou adicionando-o como um novo *subcorpus* (uma nova pasta será criada com o nome que você definir). Em ambos os casos, o arquivo selecionado será copiado da pasta original para a pasta *subcorpus*. Se deseja adicionar uma pasta de arquivos de texto, basta escolher a pasta que quer copiar e ela será copiada para a pasta *subcorpus*.
2. Depois de atribuir um nome à pasta, clique em *Next* e depois em *Finalise*.

Os arquivos de texto serão exibidos na área de arquivos (Figura 2.4). Os arquivos recentemente adicionados aparecem localizados logo abaixo da caixa de texto *Files in*

corpus but not incorporated in project. O CorpusTool faz uma distinção entre os arquivos incorporados (*incorporated*) no projeto, que estão disponíveis para qualquer anotação, e os arquivos não incorporados (*unincorporated*) no projeto, que não estão disponíveis para anotação. Essa distinção é útil para que você visualize separadamente os arquivos em que já começou a trabalhar daqueles arquivos que pode querer incluir no projeto mais tarde. Se você tem 100 arquivos no *corpus*, mas só fez anotações em cinco, é importante que esses cinco arquivos anotados estejam claramente discriminados.

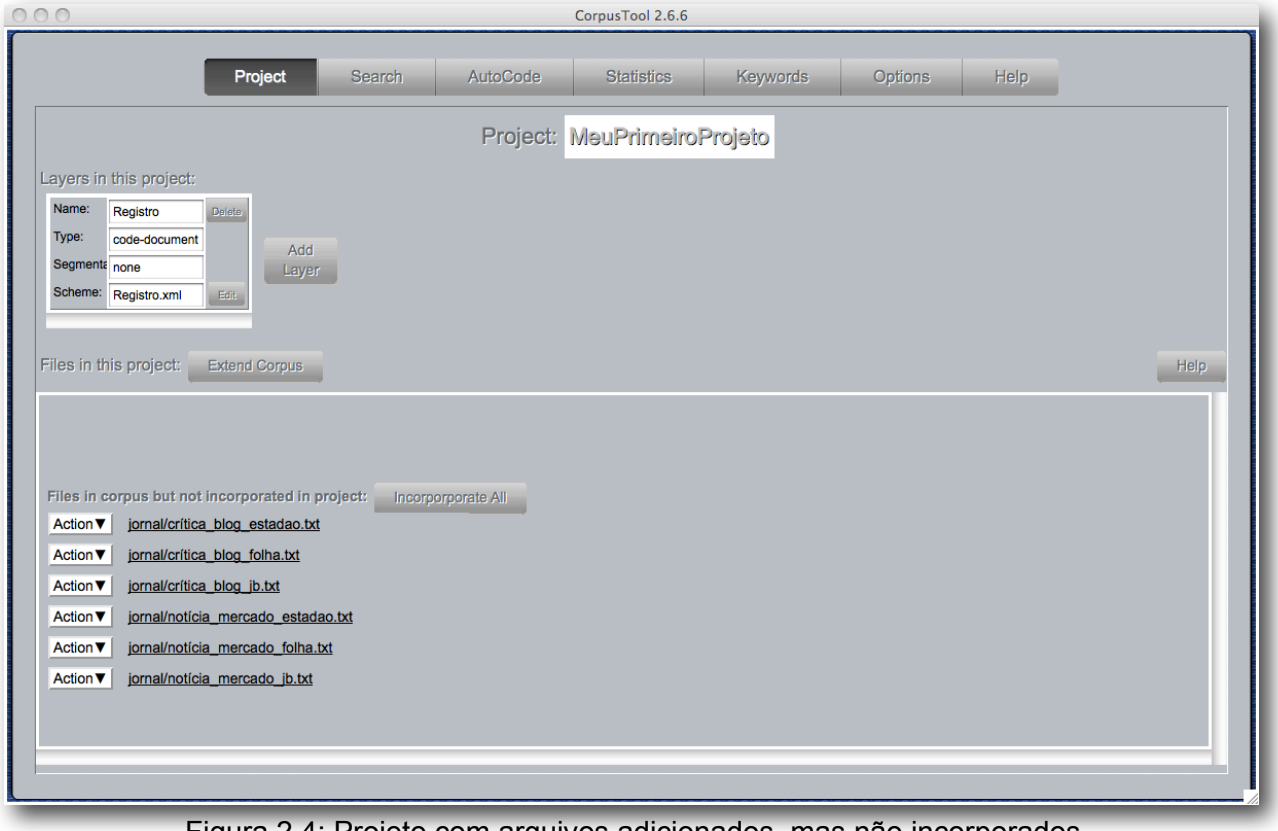


Figura 2.4: Projeto com arquivos adicionados, mas não incorporados

4.2 *Incorporando Arquivos de Texto*

Para incorporar arquivos de texto no projeto, tornando-os disponíveis para a anotação, clique no botão *Incorporate All,* se quiser que todos os arquivos sejam incorporados de uma vez só, ou abra a caixa de texto *Action* e selecione a opção *Incorporate* para incorporar um único texto.

Definindo a língua, a codificação de caracteres e o tipo de letra: quando um arquivo está sendo incorporado, algumas informações sobre o modo de armazenamento desse arquivo (*metadata*) serão pedidas (Figura 2.5). São elas:

- ***Language***: em que língua o texto está escrito? Essa informação define quais recursos de extração de dados poderão ser usados nesse arquivo. Esses recursos incluem léxicos (para buscas de concordância, cálculo da densidade lexical, etc.), *parsers* (para segmentação automáticas) e *taggers*. Atualmente, esses recursos estão disponíveis apenas para o Inglês, mas brevemente estarão disponíveis para outras línguas.

- ***Encoding***: qual é o tipo de codificação de caracteres do texto? Você pode informar ao CorpusTool qual é esse tipo de codificação, bastando escolher da lista de tipos. Por omissão, o CorpusTool faz uma suposição sobre qual o tipo de codificação do texto a ser incorporado. Mas nem sempre é possível identificá-lo corretamente. Desse modo, é necessário que você o faça manualmente. Para saber qual o tipo de codificação do seu texto, vá à pasta onde se encontra o texto original (.txt) e dê um clique direito. Selecione *Abrir com*, ou o equivalente em MacOSX, e escolha MS Word, que o ajudará a descobrir qual a codificação. Alternativamente, em *Abrir com*, pode escolher o navegador Firefox. Nele, você poderá encontrar que tipo de codificação foi atribuída ao seu arquivo de texto (submenu *Character Enconding*, logo abaixo de *View*).
- ***Display Font***: qual o tipo e o tamanho da fonte com que pretende ver o seu texto? Selecione o tipo mais apropriado à codificação de caracteres previamente escolhida (há sistemas mais adequados às codificações ocidentais; outros mais adequados aos sistemas asiáticos). De qualquer modo, a maioria de tipos modernos de fontes se ajustam a qualquer tipo de codificação de caracteres.

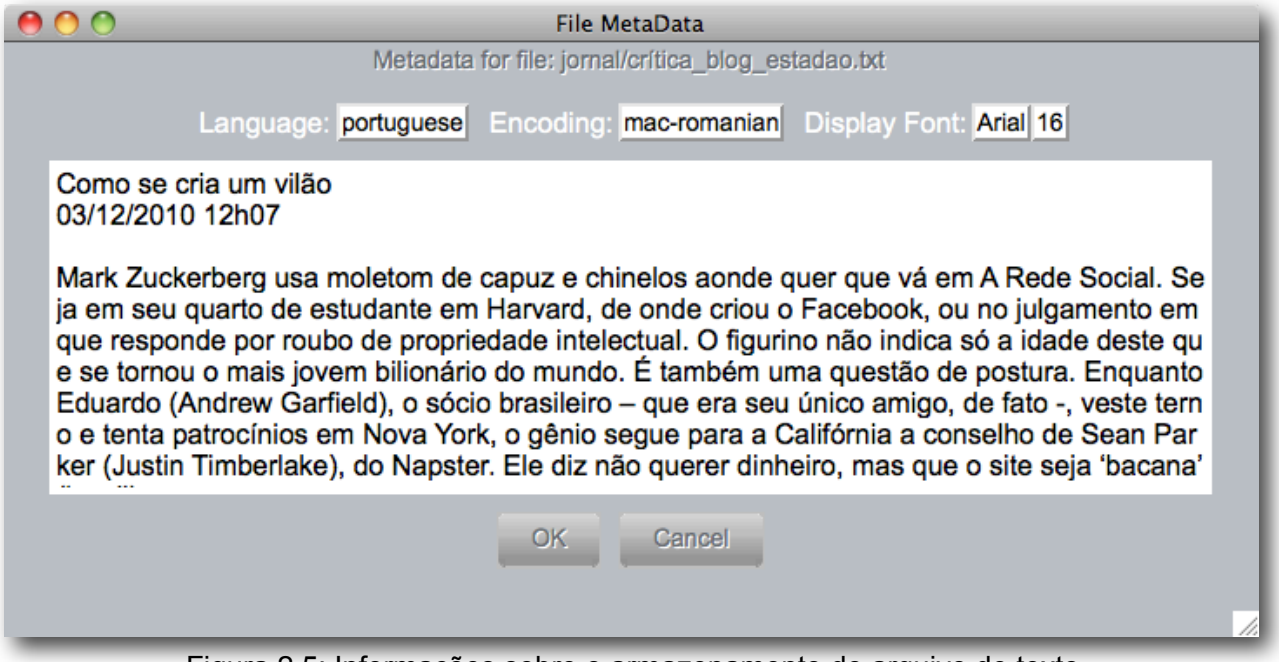


Figura 2.5: Informações sobre o armazenamento do arquivo de texto

Como exemplo, a Figura 2.6 abaixo apresenta o Projeto com dois arquivos de textos incorporados (na parte superior) e outros por incorporar (na parte inferior).

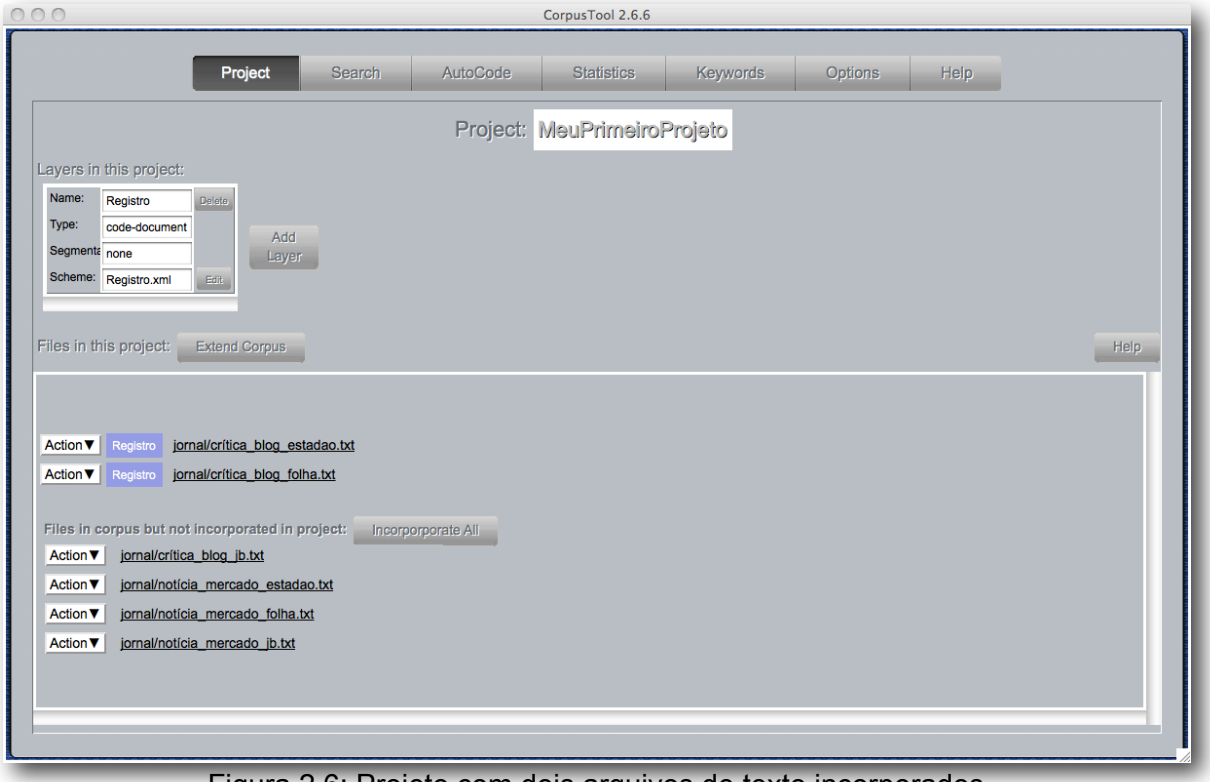


Figura 2.6: Projeto com dois arquivos de texto incorporados.

4.3 *Arquivos de Textos não Incorporados*

Ha outras ações disponíveis aos arquivos de textos não incorporados no projeto:

- ***Info***: fornece algumas informações estatísticas sobre o texto, como número de palavras, de períodos, média de ocorrências pronominais.
- ***Delete***: permite remover o arquivo de texto do *corpus*. Ao selecionar esta opção, o arquivo de texto na pasta *Corpus* do seu projeto também será apagado.
- ***Filename***: ao clicar em *Filename*, o texto será apresentado.

5 Trabalhando com Arquivos de Texto Incorporados

Uma vez que haja arquivos de textos incorporados ao Projeto, outras opções de ação ficam disponíveis:

5.1 *Mudando Metadados do Arquivo*

(Somente para anotação de texto) Já vimos acima que, quando um arquivo é incorporado, você será levado a especificar a língua, a codificação de caracteres e o tipo de fonte. Você pode mudar essas escolhas a qualquer momento, bastando para isso selecionar o botão *Change File Metadata*, que se encontra disponível no menu *Action*.

5.2 *Obtendo Informações Estatísticas sobre um Arquivo de Texto*

Para ver informações estatísticas gerais sobre cada arquivo de texto, selecione *View Basic Text Stats*, no menu *Action*. Ao selecionar essa opção, serão exibidas informações estatísticas gerais sobre o texto (essas informações não dependem de qualquer anotação aplicada ao arquivo) como se pode ver na Figura 2.7:

- número de palavras (*Words in text*);

- extensão média das palavras (*Average Word Length*);
- número de períodos (*Sentences in Text*) (somente para línguas européias).
- extensão média em quantidade de palavras por períodos (*Average Sentence Length*) (somente para línguas europeias).

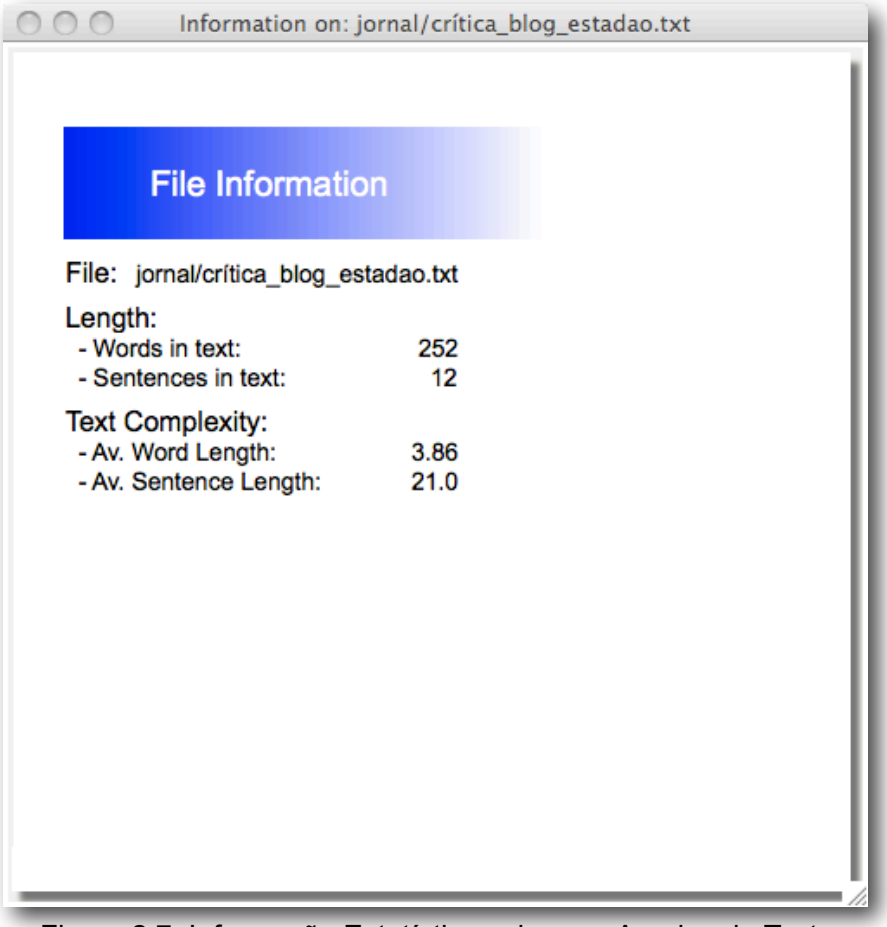


Figura 2.7: Informação Estatística sobre um Arquivo de Texto

Para textos escritos em Inglês, mais informações estatísticas estão disponíveis:

- densidade lexical (*Lexical Density*): média de palavras lexicais por período ou percentagem de palavras lexicais por texto.
- densidade de Referência Pronominal (*Reference Density*): percentagem de uso dos pronomes pessoais no texto.

Nota: À medida que dicionários de outras línguas forem adicionados ao CorpusTool, essas informações estatísticas se tornarão disponíveis para essas línguas.

5.3 Desincorporando um Arquivo do Corpus

O botão *Unincorp* remove o arquivo de texto do seu estudo.

AVISO: Qualquer anotação feita nesse arquivo será apagada. O arquivo de texto será incluído na lista de arquivos não incorporados. Você pode reincorporar esse arquivo mais tarde, porém sem nenhuma anotação.

5.4 Abrindo uma Janela de Anotação

Além do botão *Action*, os outros botões correspondem às camadas de anotação definidas no seu projeto.

Cores do Botão de Anotação: As cores dos botões de anotação de cada camada de um documento correspondem aos graus da anotação que se tenha realizado:

- Branco: totalmente anotado
- Azul claro: parcialmente anotado
- Azul escuro: anotado num grau avançado

Note que essas cores são somente indicativas.

6 Saindo do CorpusTool

Quaisquer mudanças no projeto são salvas automaticamente. Se você sair da janela de gerenciamento do projeto (clicando no X do canto superior direito), encerrando o CorpusTool, todas as mudanças serão salvas.

7 Continuando um Projeto

Uma vez criado o seu projeto, a forma mais fácil de abrir o CorpusTool é:

1. abrir a pasta do projeto na Área de Trabalho
2. fazer duplo clique no arquivo .cptr (representado por um globo azul).

O CorpusTool abrirá diretamente no painel de gerenciamento do seu projeto.

DESFAZER: Atualmente, o CorpusTool não permite desfazer ações. Em versões futuras, essa opção será incluída.

Seção 3:

Esquemas de Anotação

8 Abrindo o Editor de Esquemas

Antes de começar a anotação de textos, você deve definir um esquema para cada camada. Abra o editor de esquemas. Clique no botão *Edit*, que está dentro da área de trabalho da camada a ser editada (figura 3.1).

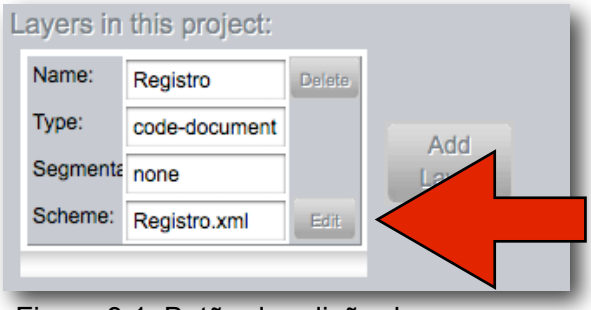


Figura 3.1: Botão de edição de esquemas

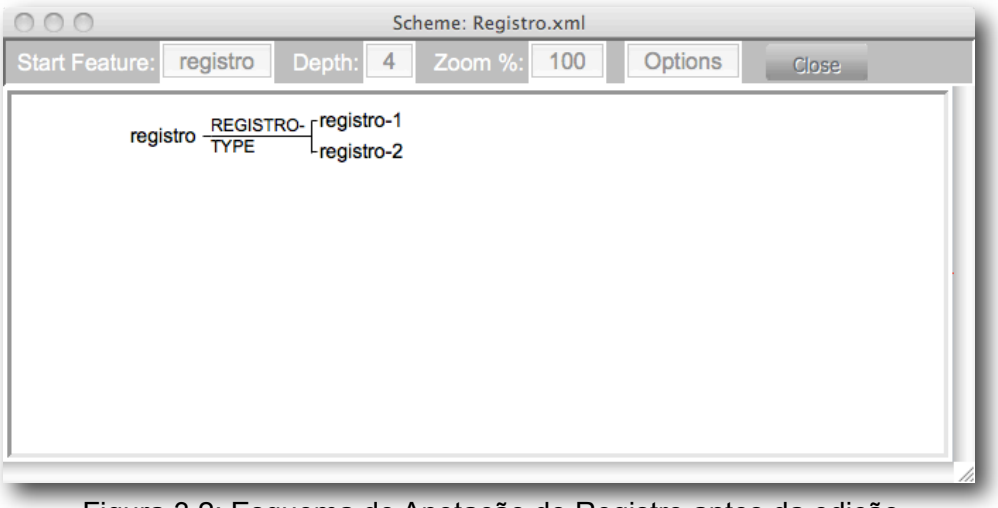


Figura 3.2: Esquema de Anotação do Registro antes da edição

Ao clicar em *Edit*, uma janela se abre (Figura 3.2), apresentando uma pequena rede de sistemas, sendo "registro" a base dessa rede, e "registro-1" e "registro-2" as características.

9 Editando o Esquema de Anotação

As características são geradas automaticamente e podem ser renomeadas. Clique em "registro-1". Um menu, como na Figura 3.3, se abre:

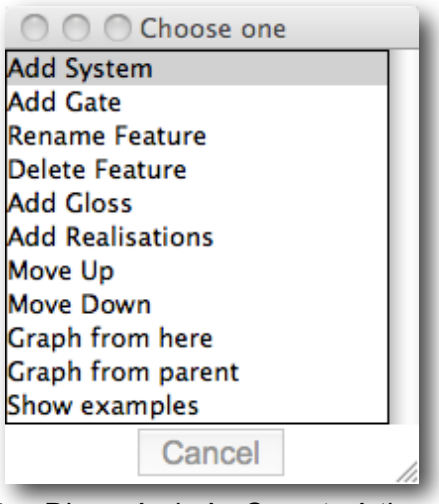


Figura 3.3: Opções Disponíveis às Características de um Sistema

Essas opções serão detalhadas mais adiante. Por agora, o objetivo é modificar o nome "registro-1" para algo mais informativo. Imagine que todos os textos incorporados no seu projeto são textos jornalísticos, ora notícias, ora críticas de cinema. Então, você deseja que "registro-1" seja "notícia" e "registro-2" seja "crítica".

Clique em *Rename Feature*. Forneça o novo nome. Digite: *notícia* e dê *enter*. Repita o mesmo processo para a característica "registro-2", renomeando-a para "crítica". Note que entre as características do sistema (agora, "notícia" e "crítica") e a base do sistema ("registro"), há outro nome automaticamente fornecido como "REGISTRO-TYPE". Renomeie para "JORNAL". Para isso, clique em "REGISTRO-TYPE". Escolha *Rename System*.

Os esquemas podem ser muito mais complexos. O esquema apresentado na Figura 3.4 é bastante mais complexo, podendo ser estendido para conter centenas de escolhas. Contudo, quanto menor o esquema, mais fácil a anotação.

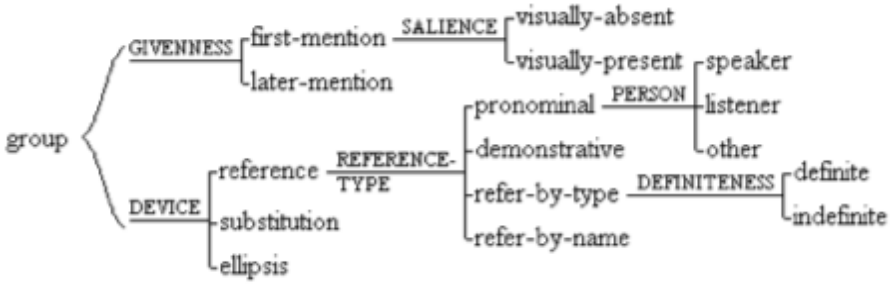


Figura 3.4: Um Esquema de Anotação Complexo

9.1 *Entendendo Redes de Sistemas*

O CorpusTool baseia-se na representação hierárquica da língua proposta pela Linguística Sistêmico-Funcional. Essa hierarquia é chamada de rede de sistemas, consistindo de escolhas interdependentes, chamadas de sistemas. O sistema na Figura 3.2 é um deles. A Figura 3.4 apresenta 6 sistemas organizados dentro de uma rede. Um sistema é composto por três partes:

- **nome do sistema**: o nome da escolha. Podem ser usados nomes típicos da gramática, como MODO, POLARIDADE, etc. Os nomes dos sistemas podem ser uma sequência de letras, números e hífen, mas sem espaços. Cada sistema deve ter um nome único. O CorpusTool não permite que dois sistemas dentro do

mesmo esquema tenham nomes iguais. Os nomes dos sistemas são mostrados em letras maiúsculas.

- **características**: alternativas de escolhas. No seu projeto, “crítica” e “notícia” são características do sistema. As características são mostradas em letras minúsculas. Não pode haver repetição de nomes dentro do sistema.
- **condição de entrada**: cada sistema tem uma condição de entrada (característica que forma o contexto onde uma escolha se torna relevante). No seu projeto, a condição de entrada para o sistema é “registro”, que também é a característica raiz da rede de sistemas.

Sistemas distintos podem ter a mesma condição de entrada, sendo chamados de sistemas simultâneos. Eles formam uma classificação cruzada da condição de entrada. Por exemplo, pode-se introduzir outro sistema a partir da condição de entrada “registro”, ao qual podemos atribuir as características “comercial”, “militar”, etc.

O conjunto de sistemas forma uma rede de sistemas, com características de um sistema sendo condições de entrada para outros sistemas mais específicos. A criação dessas redes de sistemas está descrita logo a seguir.

9.2 Criando e Modificando Sistemas

Ao clicar em uma característica (em letras minúsculas) ou em um sistema (em letras maiúsculas), um menu de ações será apresentado. Essas ações permitem estender ou modificar a rede.

9.2.1 Ações sobre o Sistema

- ***Add Feature***: permite adicionar uma característica a um sistema.
- ***Rename System***: permite mudar o nome de um sistema.
- ***Delete System***: permite apagar um sistema da rede de sistemas. **AVISO**: as características que pertencem a um sistema, e quaisquer sistemas que tenham essa características como sua condição de entrada também serão apagados. Lembre que ainda não existe a opção de desfazer ações no CorpusTool. Lembre também que qualquer anotação feita com base no sistema a ser apagado também será apagada.
- ***Change Entry Condition***: permite mudar a condição de entrada de um sistema, selecionando uma característica diferente da atual.
- ***Move Up***: permite mover um sistema para uma posição acima, reorganizando a rede de sistemas.
- ***Move Down***: permite mover um sistema para uma posição abaixo, reorganizando a rede de sistemas.

9.2.2 Ações sobre as Características

- ***Add System***: permite criar um novo sistema a partir de uma característica.
- ***Rename Feature***: permite mudar o nome de uma característica.
- ***Delete Feature***: permite apagar uma característica de um sistema. **AVISO**: quaisquer sistemas que dependam dessa característica também serão apagados. Lembre que ainda não existe a opção de desfazer ações no CorpusTool. Lembre também que qualquer anotação feita com base nessa característica também será apagada.

- ***Move Up***: permite mover uma característica para uma posição acima, reorganizando as características de um sistema.
- ***Move Down***: permite mover uma característica para uma posição abaixo, reorganizando as características de um sistema.
- ***Edit Realisations***: permite adicionar glosas a qualquer uma das características. O programa não reconhece essas glosas, mas elas podem ser úteis para anotar características, funcionando como indicação do contexto de realização.
- ***Show Examples***: se houver textos anotados, ao clicar nessa opção, o painel de pesquisa do CorpusTool se abre com exemplos anotados.

9.2.3 Mudando as Condições de Entrada

Para mudar a condição de entrada de um sistema, clique no sistema e selecione *Change Entry Condition*. Será apresentada uma caixa de diálogo como se vê na Figura 3.5 abaixo.

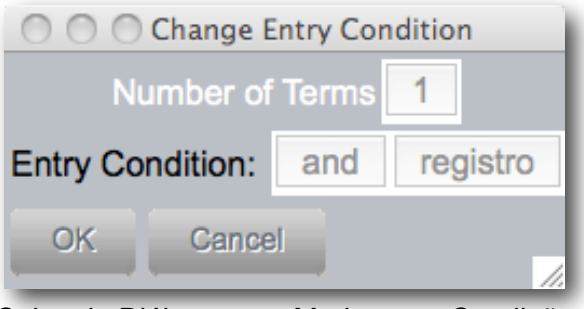


Figura 3.5: Caixa de Diálogo para Mudar uma Condição de Entrada

Simple Entry Condition: se você deseja uma condição de entrada simples (o sistema se expande a partir de uma única característica), então defina o número de entradas (*Number of Terms*) como uma entrada e depois escolha a característica que você quer que seja a condição de entrada. Aperte OK e o sistema será redefinido. O CorpusTool atualiza automaticamente as anotações afetadas por essa mudança.

*Complex Entry Condition***:** também é possível definir condições de entrada complexas a uma rede de sistemas. Uma condição de entrada complexa envolve uma conjunção (‘e’) ou uma disjunção (‘ou’) de características. Defina o número de entradas (*Number of Terms*) como 2 ou mais. A seguir, selecione as características que você quer que sejam as condições de entrada. Note que, na versão atual do CorpusTool, não é possível misturar conjunção (‘e’) com disjunção (‘ou’) na mesma condição de entrada. Mas é possível contornar a situação, criando um portão (*gate*, sistema com apenas uma característica). Por exemplo, para definir uma condição de entrada complexa ‘A e (B ou C), primeiramente crie um sistema com uma única característica (*gate*). Depois, use essa nova característica e mais a característica A em conjunção (‘e’) para o sistema que você deseja criar. A condição de entrada para esse sistema será 'A e (B ou C)'.

9.2.4 Movendo uma Característica para Outro Sistema

Se você deseja mover uma característica de um sistema para outro, clique sobre o sistema ao qual vai adicionar a característica e selecione *Add Feature*. A seguir, digite o nome da característica que você deseja mover. A característica será colocada no outro sistema. Todas as anotações serão ajustadas a essa mudança.

10 Adicionando Glosas

Você pode adicionar glosas a cada uma das características de anotação. Clique sobre a característica e selecione *Add Gloss*. Digite a descrição da característica, os critérios em que ela pode ser ativada, *etc*. A glosa pode ser visualizada no momento da anotação de um texto.

O menu *Options*

Para cada esquema de anotação, há o menu *Options*, com as seguintes ações disponíveis:

- *Save as*: salva o esquema numa pasta separada.
- *Show/Hide Glosses*: ao selecionar essa opção, as glosas de cada característica podem ser exibidas ou ocultas.
- *Show/Hide System Names*: ao selecionar essa opção, os nomes dos sistemas podem ser exibidos ou ocultos. Ocultar os nomes dos sistemas pode tornar a edição de esquemas mais difícil, já que você tem de clicar no nome do sistema para ter acesso a funções como adicionar característica (*add feature*).
- *Save Diagram as PDF*: salva a rede de sistemas como um arquivo pdf.
- *Save Diagram as SVG*: salva a rede de sistemas no formato SVG (*Scalable Vetor Graphics*). A seguir, veja mais informações sobre o formato SVG.
- *Copy to Clipboard*: (somente para *Windows*): permite copiar a rede de sistemas como está sendo exibida. Você pode colar no MS Word ou noutro programa. Note que o Word deve estar aberto no momento em que você copiar a rede. Se o Word não estiver aberto, você não consegue realizar essa operação.

11 Produzindo Imagens para Incluir em Documentos ou na Web

Embora o formato SVG não seja ainda amplamente suportado, é um excelente formato porque facilmente se converte em outros formatos, já que ele armazena a imagem geometricamente, diferentemente de um *bitmap*. Para isso, faça o download do *InkScape* (http://www.inkscape.org/) e instale-o. É um software livre, disponível para as plataformas Windows, Macintosh e Linux. Abra o InkScape e selecione *Open* no menu *File*. Escolha o seu arquivo svg.

Você pode editar o arquivo, se desejar. Para salvar noutro formato, há duas opções:

1. Selecione *Save as* para salvar como PDF, EPS, EMF, ou outro formato de arquivo baseado em vetor.
2. Selecione *Export Bitmap* para salvar como arquivo PNG, que é um formato de *bitmap* que pode ser incluído em páginas web ou documentos do Word. O diagrama na Figura 3.6 é um arquivo PNG produzido no InkScape:

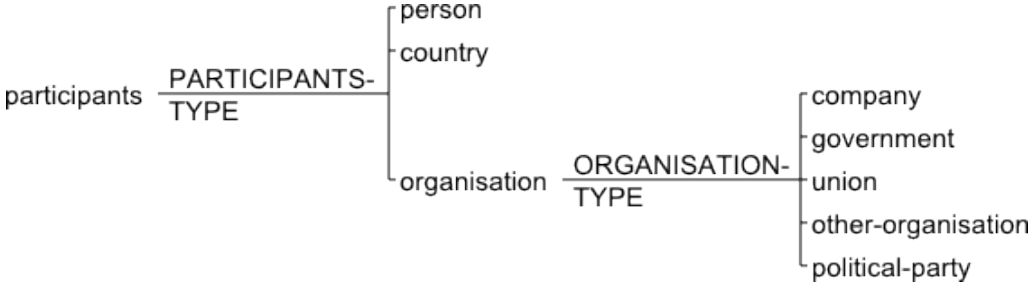


Figura 3.6: Resultado de arquivo PNG

Seção 4:

Anotação de Texto

12 Entendendo os Tipos de Anotação

O CorpusTool atualmente suporta dois tipos de anotação:

1. *Code-document*: em que são atribuídas características ao documento de texto como um todo. É bastante útil para definir tipo/gênero de texto, registro, etc. Também pode ser útil para atribuir características relativas ao escritor, como, por exemplo, nível de proficiência no uso da língua.
2. *Code*-segments: em que o usuário define segmentos dentro do texto e atribui características a esses segmentos apenas, e não ao texto. Por exemplo, oração, grupo nominal, palavras, etc., dentro de um texto.

A seguir, serão explicados os dois modos de anotação.

13 Anotando Documentos

Cada arquivo de texto incorporado no seu projeto tem um botão correspondente a cada camada de anotação. Ao clicar no botão de uma camada que previamente foi definida como *Code document*, uma janela, como na Figura 4.1, aparecerá.

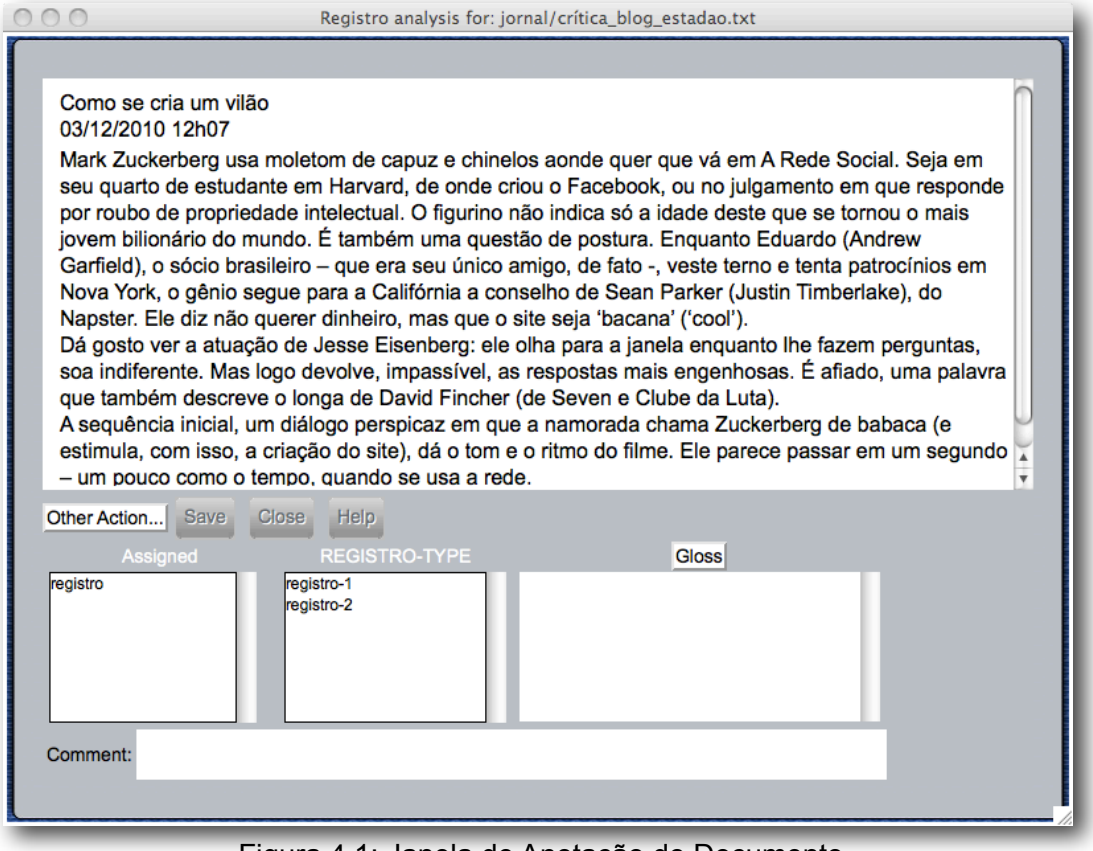


Figura 4.1: Janela de Anotação do Documento

A janela de Anotação do documento está dividida em quatro partes:

1. Quadro do texto: que apresenta o texto em si. Use as setas para cima e para baixo para ver o texto completo.
2. Barra de tarefas: que permite várias ações, como salvar (*Save*), fechar (*Close*) e ajuda (*Help*), bem como outras ações (*Other Action*).
3. Quadro de anotação: que contém três outros quadros:
    a. quadro de características selecionadas (*Assigned*): em que estão as características já escolhidas para o documento em estudo. Por omissão, este quadro já apresenta uma característica, que é a característica raiz da rede de sistemas. Todas as outras serão atribuídas pelo usuário. É possível excluir características do quadro *Assigned*, fazendo um clique duplo em cima da característica que se deseja excluir. A característica raiz não pode ser apagada.
    b. quadro de características possíveis (*Current Choice*): em que se encontram as características possíveis de serem atribuídas ao documento. Ao fazer duplo clique em uma dessas características, ela se move para o quadro das características escolhidas. Se houver mais características possíveis, elas serão apresentadas imediatamente depois.
    c. quadro de glosas (*Gloss Box)*: em que são apresentadas glosas sobre as características de anotação. As glosas são definidas pelo usuário no momento da edição da rede de sistemas. Ao dar um clique simples sobre uma característica do quadro de características possíveis (*Current Choice Box*), a glosa será exibida no quadro de glosas (*Gloss Box*).
4. Quadro de Comentários: em que você pode escrever comentários sobre um dado segmento. Esse comentário pode servir tanto como registro pessoal ou como forma de comunicação com outros investigadores que trabalhem no mesmo projeto. Por exemplo, você pode ter alguma dúvida e escrever o seguinte comentário: “Este é um processo material ou comportamental? Consultar Halliday (2004)”.

Em síntese, para anotar um documento:

1. Selecione características do quadro de características possíveis (*Current Choice Box*) até que não haja mais características disponíveis.
2. Se cometer um erro, faça duplo clique na característica incorreta no quadro de características selecionadas (*Selected Feature Box*) para desfazer o erro.
3. Salve a anotação.

14 Anotando Segmentos

Quando se quer anotar segmentos dentro de um texto, o procedimento é um pouco mais complexo.

Para continuar este guia, devemos adicionar uma nova camada de anotação.

1. Certifique-se de que está no painel principal do projeto (*Project*).
2. Clique em *Add Layer*
3. Nomeie a nova camada de “Participantes”
4. Selecione *Annotate Segments*
5. Selecione *No automatic segmentation*

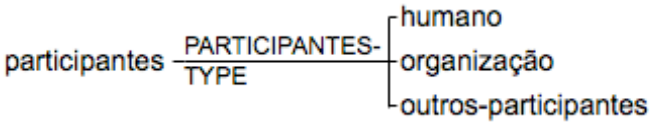
6. Selecione *Create New Scheme*
7. Clique em *Finalise*

Note que, ao adicionar uma nova camada de anotação, surge um novo botão ao lado de cada arquivo incorporado em seu projeto.

Agora, vamos definir o esquema de anotação desta camada:

1. Clique em *Edit*, que está na caixa da nova camada Participantes.
2. Quando a janela de esquema de anotação se abrir, mude o nome "participantes-1" para "humano" e o nome "participantes-2" para "organização".
3. Clique em PARTICIPANTES-TYPE e selecione *add feature*. Digite "outros-participantes".

A sua rede de sistemas deve ficar assim:



Feche a janela de edição e volte ao painel do projeto.

Clique no botão Participante que está ao lado de cada arquivo incorporado no projeto.

Será aberta uma janela de anotação para essa camada (Figura 4.2).

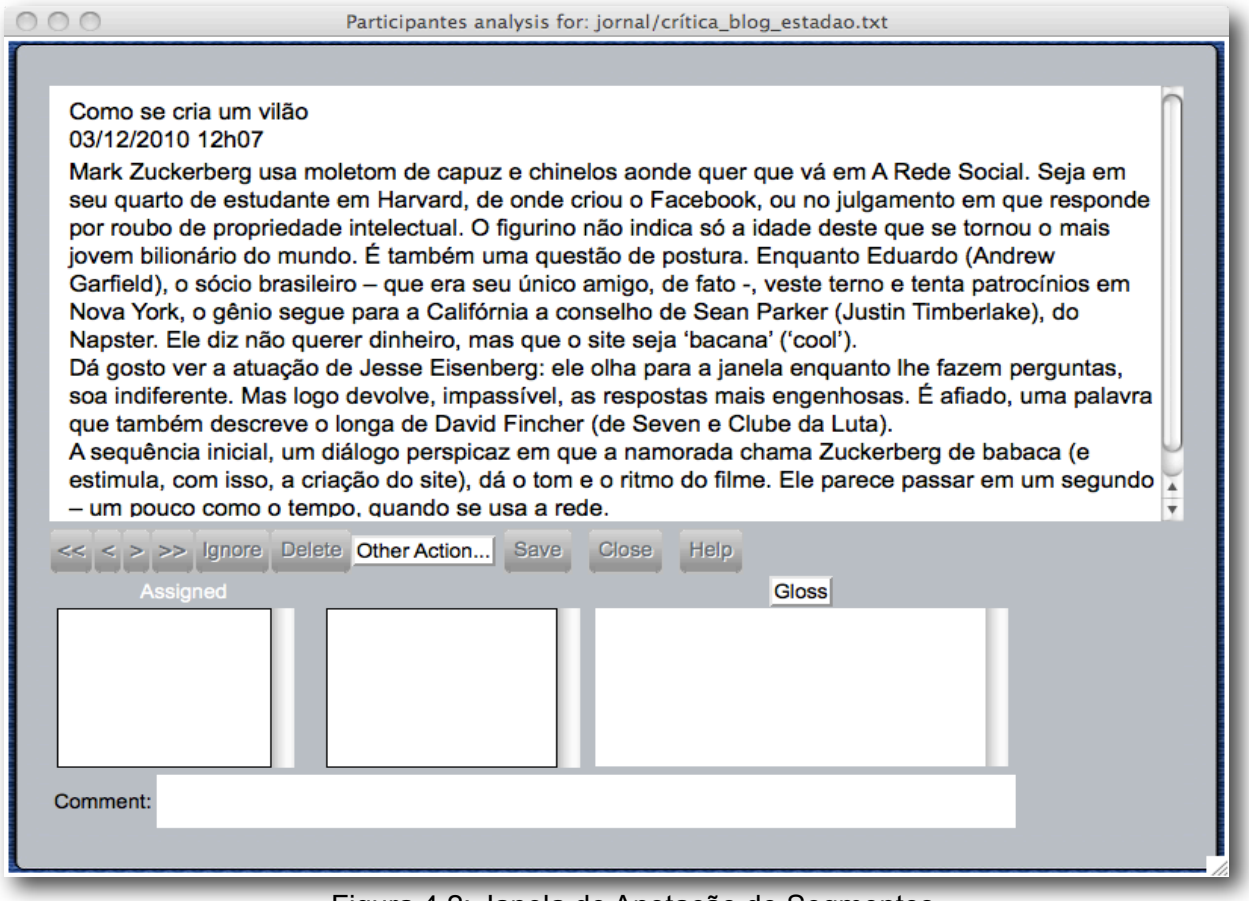


Figura 4.2: Janela de Anotação de Segmentos

A diferença entre esta janela de anotação de segmentos e a janela de anotação de documentos é que, para anotar segmentos, mais opções estão disponíveis na barra de tarefas. Essas opções permitem ao usuário mover-se de um segmento a outro.

14.1 Criando, Movendo e Selecionando Segmentos

- **Crie um segmento**: clique em um ponto do texto que você quer que seja o início do segmento e arraste o cursor até o ponto que você quer que seja o final do segmento e libere o mouse.
- **Selecione um segmento:** clique na linha abaixo de cada segmento (ao passar o cursor por cima de um segmento, o sublinhado se destaca).
- **Selecione um segmento anterior/posterior:** use os botões < e > na barra de tarefas.
- **Selecione um segmento anterior/posterior de anotação incompleta:** use os botões << e >>.
- **Redimensione um segmento:** Passe o cursor sobre uma das extremidades de um segmento, procurando um pequeno marcador (uma linha vertical) até que ele fique vermelho, o que indica que você está sobre ele. Clique e arraste até o ponto que você deseja.
- **Apagando um segmento:** selecione o segmento que quer apagar e clique no botão *Delete* na barra de tarefas ou na tecla delete do seu teclado.

14.2 Ignorando Segmentos

Se você não quiser que um dado segmento seja considerado na análise estatística ou na pesquisa, selecione-o e clique em *Ignore*, na barra de tarefas. O segmento ignorado se apresenta cinza no quadro de texto. Para anular essa ação, clique no mesmo botão (se o segmento selecionado estiver ignorado, o botão terá outro nome: *Unignore*).

15 Explorando o Menu *Other Actions*

Se você estiver anotando segmentos dentro de um texto, esse menu mostra as seguintes opções:

- ***Edit Scheme***: abre a janela de edição de esquemas daquela camada para que você possa editar o esquema.
- ***Add New Feature***: abre uma nova janela onde é possível inserir uma nova característica ao sistema de escolhas em que você esteja trabalhando.
- ***Copy Features***: copia as características atribuídas aos segmentos.
- ***Paste Features***: atribui as características previamente copiadas a outro segmento.
- ***Resegment Document***: apaga toda a segmentação atribuída ao texto. **AVISO**: não é possível desfazer essa ação.
- ***Show XML***: mostra o arquivo XML correspondente ao arquivo de texto.
- ***Show Structure***: altera a forma de apresentação das anotações atribuídas ao texto (Figura 4.3).

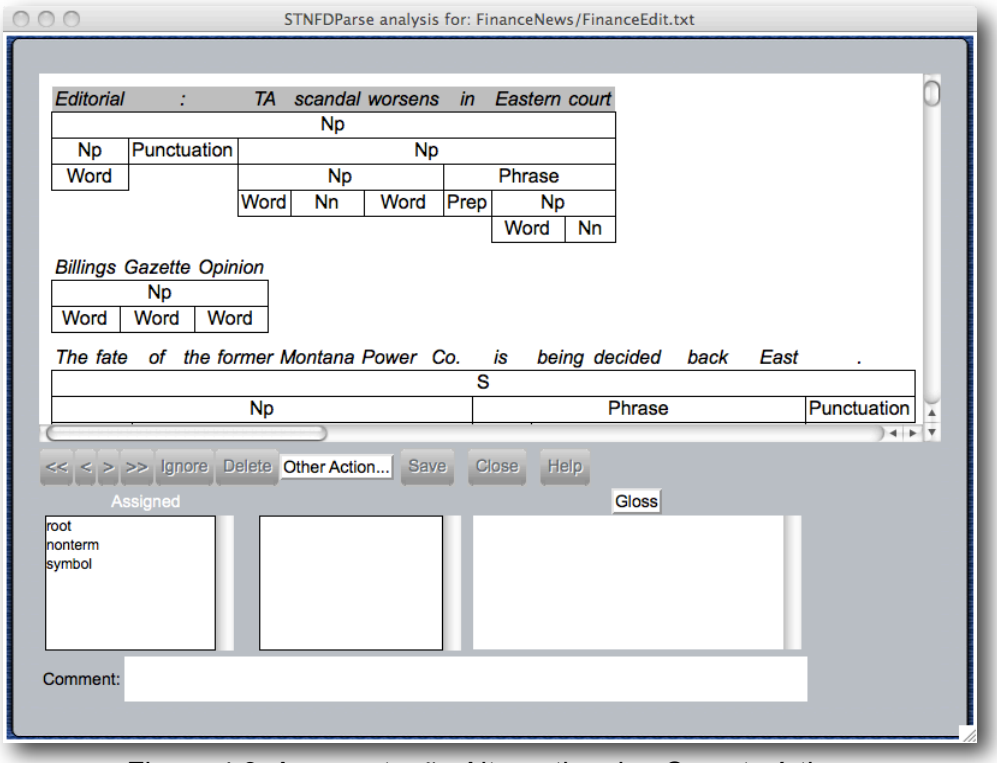


Figura 4.3: Apresentação Alternativa das Características

- **Show Text Stream**: abre uma nova janela que permite visualizar graficamente como as características se distribuem ao longo de um texto. Em *System to Graph*, selecione qual sistema deseja visualizar. Em *Smoking*, selecione o nível de contraste. Ao selecionar o nível de contraste 0, cada característica é mostrada na sequência em que ocorre. Aplique níveis mais elevados de contraste para visualizar melhor como as características se distribuem pelas fases de um texto. Por exemplo, na Figura 4.4, o gráfico mostra que os participantes do tipo organização ocorrem mais fortemente no início do texto e que ocorrem com menor frequência no fim do texto.

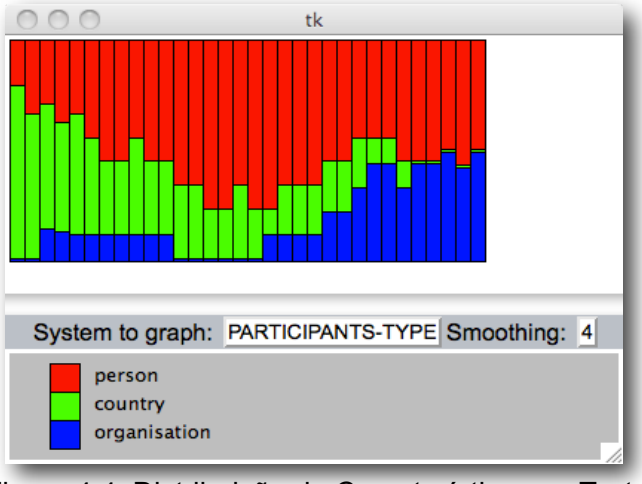


Figura 4.4: Distribuição de Características no Texto

Seção 5:

Pesquisas

16 Introdução

Para abrir o painel de pesquisas, basta clicar em *Corpus Search*, no painel principal do CorpusTool (Figura 5.1).

NOTA: Você pode abrir o painel de pesquisas:

- através do editor de esquemas, clicando em *Show Examples*. O CorpusTool abre uma nova janela em que os segmentos anotados com dada característica estarão destacados.
- Através do painel de informação estatísticas, clicando no número que indica o total de ocorrências de uma dada característica.

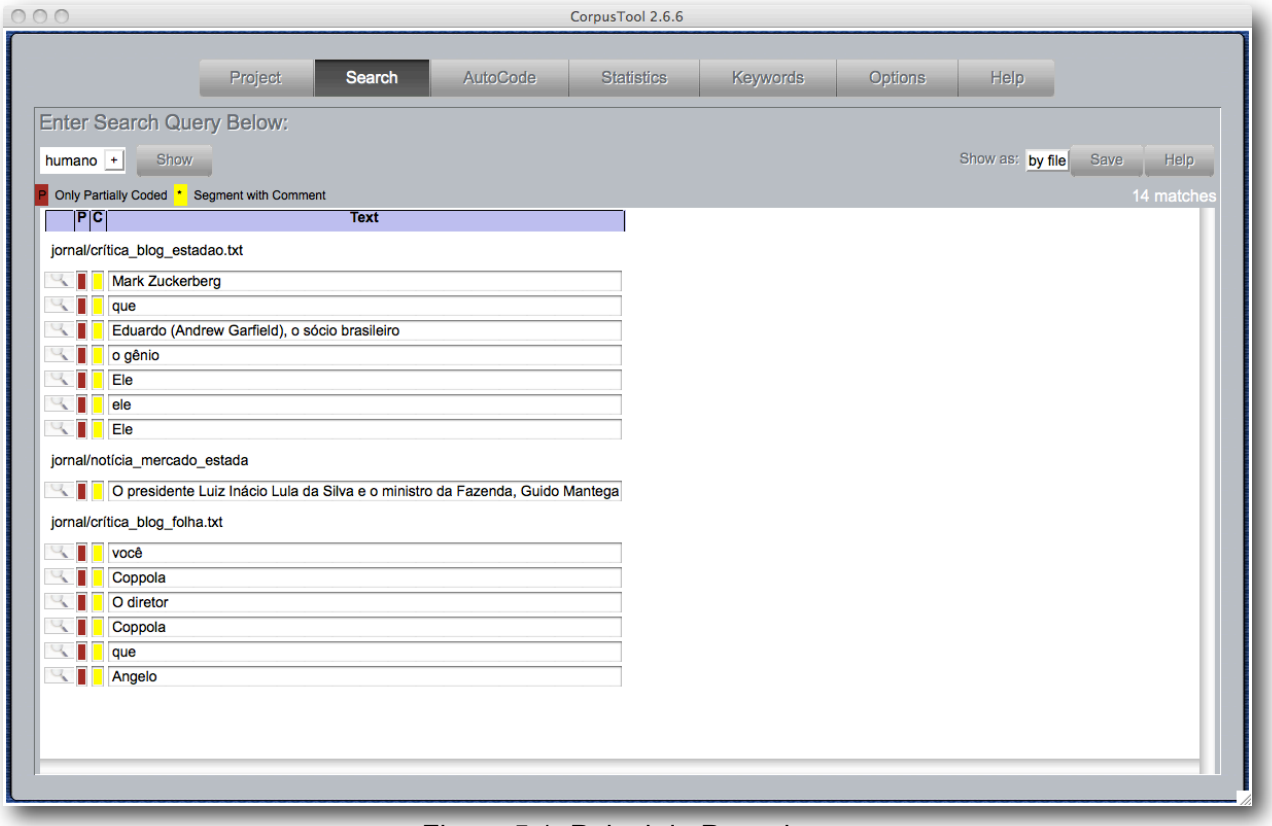


Figura 5.1: Painel de Pesquisas

17 Especificando Critérios de Pesquisa

Para fazer pesquisas, abra o menu de opções que se encontra logo abaixo de *Enter Search Query Below*, clicando no nome da camada. Para seguir este guia, vamos usar o projeto *Finance*, um pequeno projeto em Inglês, disponível para download na página do UAM CorpusTool (http://www.wagsoft.com/CorpusTool/download.html).

1. **Pesquisa simples**: para encontrar todos os segmentos com determinada característica, selecione uma camada e, de seguida, clique em *Show* para ver todas as instâncias. Clique em *Save* para salvar o resultado da pesquisa em um arquivo separado.
2. **Pesquisas complexas**: Clique no sinal "+" para ampliar a sua pesquisa.
   - ***and*:** permite adicionar outra característica. O resultado apresenta os segmentos que estejam anotados com as duas características ao mesmo tempo.
   - ***or*:** permite adicionar outra característica. O resultado apresenta os segmentos que contenham uma das duas características indicadas.

     NOTA: *and* e *or* não podem ser combinados!
   - ***and not*:** permite adicionar outra característica que funciona como critério de exclusão. O resultado apresenta todos os segmentos anotados somente com a primeira característica, e nunca com a segunda.
   - ***containing segment***: permite fazer pesquisas cruzadas, inclusive entre camadas: o resultado apresenta todos os segmentos anotados com a primeira característica que contêm segmentos anotados com a segunda característica. Por exemplo, pode-se fazer uma pesquisa por segmentos de orações com a característica "finite" que contenham segmentos de grupos nominais com a característica "person".
   - ***containing string***: permite encontrar qualquer segmento que contém uma *string* (sequência de carateres) definida pelo usuário.

     NOTA: esse tipo de pesquisa também funciona para pesquisa baseada em léxico, em caracteres especiais (*wildcards*), etc. Mais detalhes a seguir.
   - ***in segment***: permite fazer pesquisas cruzadas, inclusive entre camadas: o resultado apresenta todos os segmentos anotados com a primeira característica que estão contidos nos segmentos anotados com a segunda característica. Por exemplo, pode-se fazer uma pesquisa por segmentos com a característica "person" que estejam dentro de um documento com a característica "news".

     Critérios de inclusão: para pesquisas com *containing segment*, *containing string* ou *in segment*, você pode escolher entre inclusão do tipo *immediate* ou do tipo *anywhere*. Isso se dá porque o CorpusTool permite criar segmentos encaixados em outros segmentos. A diferença entre um tipo e outro é a seguinte:
     - *anywhere*: Considere, por exemplo, os segmentos "[They left because [she was tired]]", com uma oração encaixada em outra. Em uma pesquisa por "clause" `containing anywhere 'was'`, o resultado apresenta os dois segmentos: tanto o segmento completo como o segmento encaixado.
     - *immediately*: Considerando o mesmo exemplo, com os segmentos "[They left because [she was tired]]", ao especificar *immediately* com uma

pesquisa por “clause” `containing immediately 'was',` o resultado só apresenta a oração encaixada.

3. **Pesquisas complexas combinadas**: veja o exemplo:

   person containing immediately “bush” in finite-clause in editorial&english

18 Pesquisando por Concordância

O CorpusTool permite fazer pesquisas usando padrões lexicais (função disponível essencialmente para o Inglês, com poucas possibilidades para outras línguas).

18.1 Especificando um Padrão

Ao selecionar *containing string*, você pode especificar um padrão lexical em vez de uma *string*. Por exemplo, para encontrar orações passivas, o padrão "be% @participle" apresentará todos os segmentos que contenham uma forma do verbo 'be' seguido de um verbo no particípio (verbos terminados em -*en*).

Note que o *corpus* não está anotado com etiquetas POS (*part of speech*). O CorpusTool inclui um dicionário de Inglês e, para identificar ocorrências, procura todas as correspondências nesse dicionário. Por exemplo, uma pesquisa por “be%” terá como resultados todas as formas do verbo, inclusive “being”, mesmo quando não pertence à classe de verbos, e sim de nomes, como em “the being”.

É importante ainda referir que qualquer pesquisa é insensível a maiúsculas ou minúsculas (*case insensitive*), portanto a busca por ‘Birch’ resultará em ‘Birch’ e ‘birch’ e “BIRCH’.

A estrutura da pesquisa consiste em *tokens*, conjunto de caracteres separado de outro por espaço. Os *tokens* podem ser os seguintes:

1) *token* literal: um *token* não acompanhado de *, #, @ ou % terá como resultado ele mesmo.
2) *Wildcard token*: se o *token* incluir o símbolo ‘*’ (um dos caracteres especiais conhecidos por *wildcards*), esse símbolo substituirá quaisquer caracteres. Portanto:
   - ca* encontrará ‘cat’, ‘carburettor’, etc.
   - *ed encontrará ‘weed’, ‘lived’, etc.
   - bro*en encontrará ‘broken’, ‘Brollerglen’, etc.
3) O símbolo ‘#’ unicamente identifica qualquer *token*.

(As três categorias de pesquisa acima funcionam para qualquer língua cujas palavras se dividam por espaços em branco ou pontuação).

4) Restrição por classe: um *wildcard* pode estar associado com ‘@’ seguido de uma classe gramatical. Nesse caso, serão apresentadas todas as palavras que, de acordo com o dicionário, possam ser consideradas parte dessa classe:
   - ca*@noun encontrará nomes começando com ‘ca’.
   - *ing@mental-projecting encontrará verbos de projeção mental terminados em ‘ing’.

   O símbolo ‘*’ não funciona sozinho, ou seja, deve estar acompanhado de um texto antes ou depois dele.

Uma lista completa das características lexicais que podem ser usadas constam do Apêndice II (em Inglês) e também podem ser encontradas na opção *Show Wordclass Networ*k, no menu *Misc* do CorpusTool.

5) Classe geral: Se não houver nenhum *token* antes do símbolo '@', serão apresentadas todas as formas representadas pela classe indicada. Por exemplo:
   - @noun encontra qualquer nome
   - @verb encontra qualquer verbo
   - @adverb encontra qualquer advérbio
   - @mental-projecting encontra qualquer verbo anotado como mental
   - @human-noun encontra qualquer nome anotado como humano

6) Formas flexionadas: o símbolo '%', ao final de um *token* que representa a raiz da palavra, indica que qualquer forma daquele *token* deve ser encontrada. Portanto:
   - break% encontra 'break', 'broken', 'broke', 'breaking', 'breaks'
   - red% encontra 'red', 'reds' (noun), 'redder' (adj), 'reddest'
   - be% encontra 'be', 'is', 'are', 'was', 'were', 'been', 'being'
   - is% não encontra nada. Só a forma raiz da palavra pode ser usada.

   Para restringir a correspondência de formas flexionadas, pode-se acrescentar *noun*, *verb*, *adjetive* ou *pronoun* depois do símbolo '%'. Por exemplo:
   - red%noun encontra 'red', 'reds'
   - red%adjetive encontra 'red', 'redder', 'reddest'

   Note que *wildcards* não podem ser usados como símbolo %.

19 Modificando uma Pesquisa

Para mudar a característica, basta clicar na característica e selecionar outra.

Para apagar qualquer extensão da pesquisa, basta clicar em "&", "/", "containing", "in") e depois clicar em *remove*.

20 Resultados

O espaço logo abaixo da barra de ferramentas de pesquisa apresenta os resultados. Para ver o segmento no texto anotado, basta clicar na lupa à esquerda de cada linha de resultado.

As colunas à esquerda de cada linha de resultado indicam o estado da anotação:

- P/- se está completa ou parcialmente (P) anotado.
- */- se o segmento tem um comentário associado. Clique para ver o comentário.

Seção 6:

Anotação Automática

1 Introdução

No painel *Autocode*, é possível atribuir automaticamente características a segmentos, usando para isso regras de ocorrências. Por exemplo, é possível identificar as orações em voz passiva usando o seguinte padrão:

```
'clause' containing 'be% @participle'
```

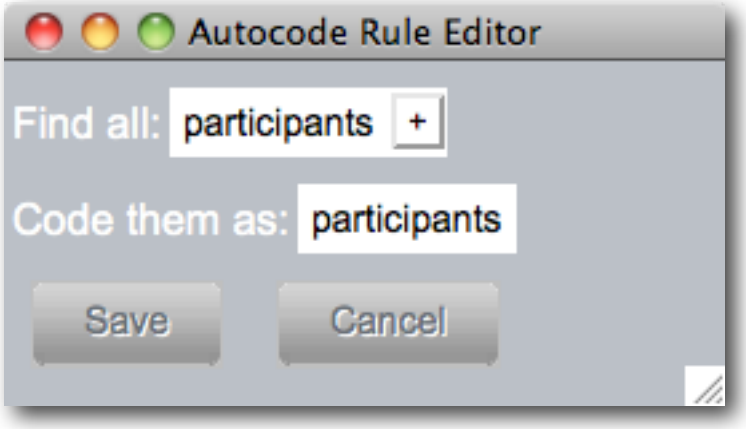
No editor de regras (*Rule*), defina uma regra como se mostra abaixo:

(**NOTA**: a anotação automática baseada em léxico atualmente só está disponível para textos em Inglês)

Ao clicar em *Show*, todas as instâncias que correspondem à regra definida são apresentadas. Ao lado de cada instância, há uma caixa de seleção que está previamente habilitada. Você pode desabilitar todas as instâncias que não corresponderem exatamente à regra de ocorrência que você definiu. Clique em *Code Selected* para que uma característica seja atribuída automaticamente a todos os segmentos habilitados.

Desse modo, poderá anotar rapidamente os principais padrões gramaticais que ocorrem num texto. Se quiser ver alguns exemplos de regras de ocorrência, adicione uma nova camada ao seu projeto e use o esquema de anotação *clauses.xml*, incluído no CorpusTool. Nas glosas desse sistema, há regras de ocorrência para a anotação de processos (mentais, verbais, etc.), voz (passiva, ativa), modalidade, etc.

1. ***Abrindo o painel Autocode***: clique em *Autocode*, no painel principal do CorpusTool.
2. ***Adicionando uma nova regra***: clique em *Add*. Uma janela se abrirá:



Selecione a característica que você deseja atribuir automaticamente. Especifique uma regra de ocorrência. As regras de ocorrência de anotação são semelhantes às perguntas de pesquisas (consulte a Seção 5 para saber como fazer pesquisas).

3. ***Editando uma regra***: Clique em *Edit*.
4. ***Apagando uma regra***: Clique em *Delete*.
5. **Anotando com a regra**: depois de definir uma regra, clique em *Show*. Todas as correspondências possíveis serão apresentadas. Uma nova barra de tarefas com mais opções se torna disponível:

- ***Display All/Agreements/Conflicting/Nonconflicting***: permite filtrar as correspondências:
  - *All*: mostra todas as correspondências.
  - *Agreements*: mostra todos os segmentos já anotados com a característica em questão.
  - *Conflicting*: mostra todos os segmentos que têm uma anotação que está em conflito com a anotação a ser aplicada. Por exemplo, se você quer anotar orações com a característica “passive”, a opção *Conflicting* permite ver todas as orações que já estão anotadas com a característica “active”.
  - *Nonconficting*: mostra todos os segmentos que nem estão anotados (*Agreements*) e nem em conflito com outras características (*Conflicting*).
- ***Select All/None***: permite habilitar ou desabilitar as caixas de seleção de uma única vez.
- ***Code Selected***: permite anotar automaticamente todos os segmentos selecionados.

Dicas

- Alguns fenômenos gramaticais são mais facilmente identificados se duas regras de ocorrência forem usadas em combinação. Por exemplo: primeiro, anote automaticamente todas as orações que contenham 'be% @participle'; depois, anote automaticamente com a característica “active” todas as orações que não contenham a característica “passive”;.
- Ao editar uma regra de ocorrência, você pode inserir o sinal # entre dois termos, como em: *select passive if contains 'be% **#** @participle'*. Esse sinal permite encontrar correspondências onde o sinal # substitui outro termo, como um advérbio de negação, por exemplo.

Seção 7:

Informações Estatísticas

1 Introdução

O painel *Statistics* permite extrair várias informações estatísticas de um *corpus* anotado. Clique em *Statistics* para ver o painel de informações estatísticas (Figura 7.1).

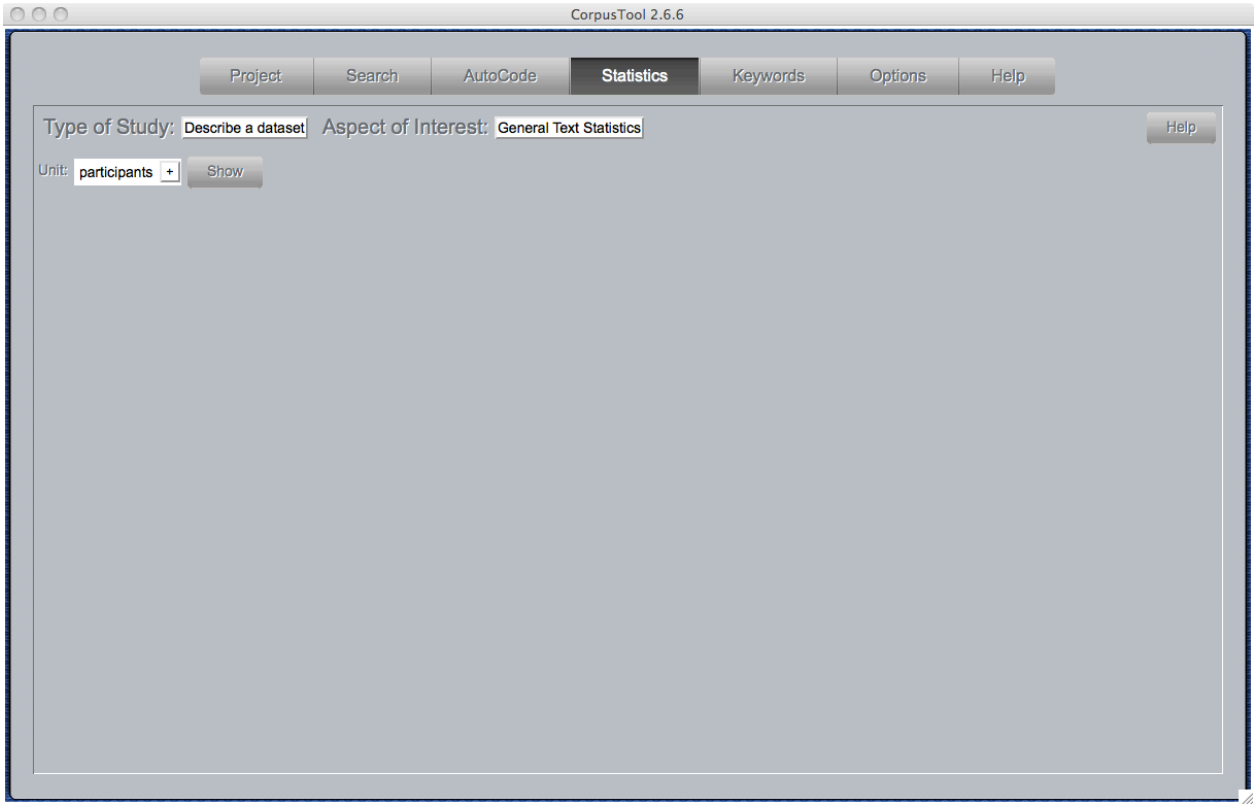


Figura 7.1: Painel de Informações Estatísticas

Você pode usar esse painel para executar tipos distintos de estudos.

Quanto ao aspeto de interesse (*Aspect of Interest*), você pode optar por:

1. ***General Text Statistics***: que oferece informações estatísticas gerais sobre o *corpus*, tais como número de segmentos, número de palavras por segmento, densidade lexical, uso pronominal (estas duas últimas informações somente estão disponíveis para textos em Inglês).
2. ***Feature Coding***: que permite especificar uma característica de uma camada de anotação (tipicamente, a característica raiz de uma rede de sistemas) para se obter valores totais, média e desvio padrão.

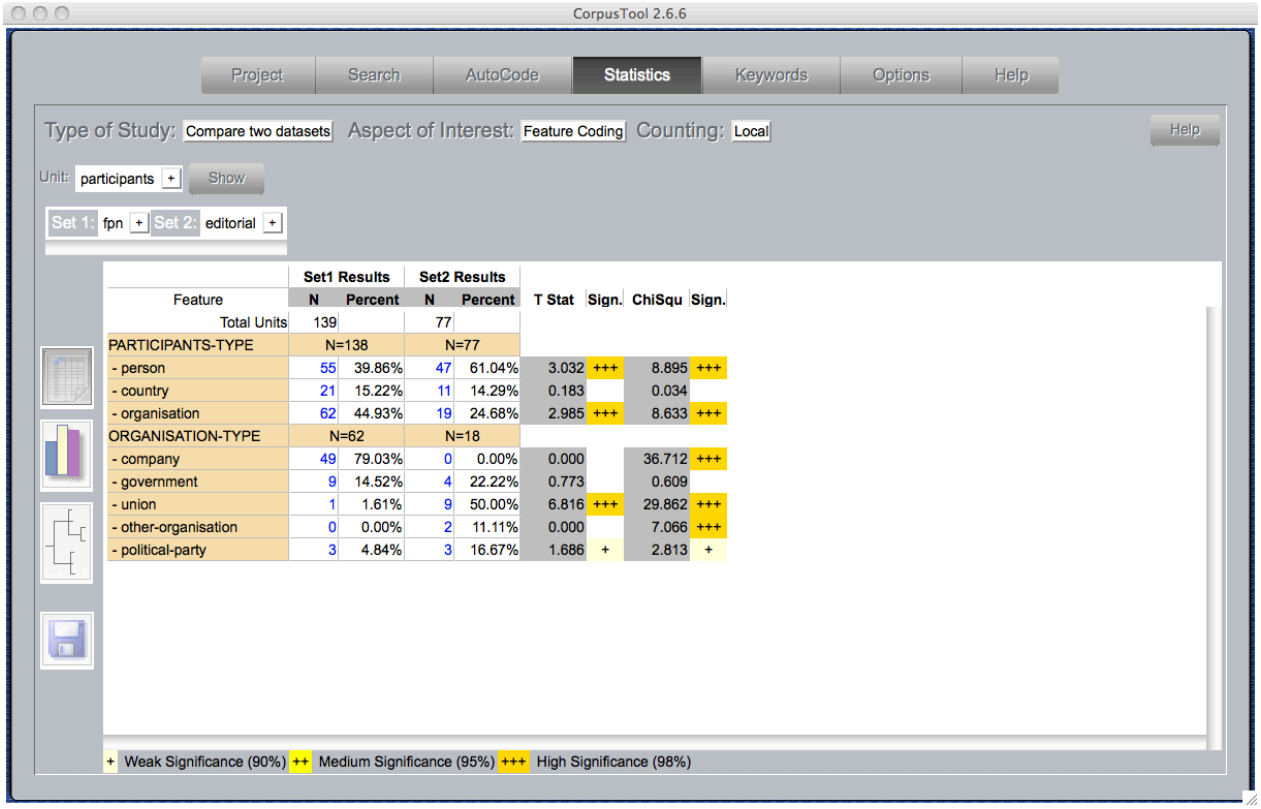
Quanto ao tipo de estudo (*Types of Study*), você pode optar por:

1. ***Describe a dataset***: que oferece informações sobre o *corpus*, ou um *subcorpus*.

2. ***Compare two datasets***: que permite executar uma comparação entre dois *corpora*, ou *subcorpora*. Se você combinar a opção *Compare two datasets* com a opção *Feature Coding*, obtém uma comparação entre as ocorrências das características selecionadas. Além disso, para cada característica comparada, níveis de significância estatística são exibidos, tanto baseados em *T-Stat* (Teste-T), como em *Chi-Squared* (Qui-Quadrado).
3. ***Compare Multiple Files***: que fornece detalhes sobre cada arquivo: um arquivo por coluna.

2 Comparando Características

A Figura 7.2 mostra uma comparação entre duas características do projeto *Finance* (disponível para download em http://www.wagsoft.com/CorpusTool/download.html). Note que, quanto à característica "participants", por exemplo, muito pouco do texto está anotado, por isso o resultado apresenta valores totais muito baixos. Para se obter resultados mais confiáveis, seria necessário anotar mil ou mais segmentos com a característica "participants", tanto em "fpn" como "editorial".



Feature	Set1 Results N	Set1 Results Percent	Set2 Results N	Set2 Results Percent	T Stat	Sign.	ChiSqu	Sign.
Total Units	139		77					
PARTICIPANTS-TYPE	N=138		N=77					
- person	55	39.86%	47	61.04%	3.032	+++	8.895	+++
- country	21	15.22%	11	14.29%	0.183		0.034	
- organisation	62	44.93%	19	24.68%	2.985	+++	8.633	+++
ORGANISATION-TYPE	N=62		N=18					
- company	49	79.03%	0	0.00%	0.000		36.712	+++
- government	9	14.52%	4	22.22%	0.773		0.609	
- union	1	1.61%	9	50.00%	6.816	+++	29.862	+++
- other-organisation	0	0.00%	2	11.11%	0.000		7.066	+++
- political-party	3	4.84%	3	16.67%	1.686	+	2.813	+

Figura 7.2: Um Estudo Estatístico Comparativo

3 Realizando um Estudo Estatístico

Para realizar um dos estudos acima apresentados:

1. **Escolha uma das opções no menu *Type of Study***: *describe a dataset*, *compare two datasets* ou *compare multiple files*.
2. **Escolha uma das opções do menu *Aspect of Interest***: *Feature Coding* ou *General Text Statistics*.

3. **Defina uma característica** (para mais informação, consulte a Seção 5, parte 2: Especificando Perguntas de Pesquisa*)*. Escolha a característica cujas diferenças você gostaria de explorar. Tanto pode ser a característica raiz da rede de sistemas (como na Figura 7.2) como pode ser uma característica de um nível mais específico.
4. Se você selecionou *Compare two datasets*, então **escolha uma característica em Set 1** e outra **em Set 2**. Selecione a característica que contém a unidade de interesse.
5. Clique em ***Show****.*

4 Interpretando Resultados: Estudo Baseado em Características

Somente os sistemas relevantes são mostrados. Por exemplo, se você especificou como unidade de interesse “person”, então o resultado envolve apenas os segmentos com a característica “person”. Por isso, não deve haver resultados para “person”, já que representa 100% de todas as ocorrências.

Valores Absolutos e Percentagens: Para cada característica, são mostrados tanto valores absolutos de ocorrências, como valores percentuais. A percentagem mostra a proporção dos segmentos estudados em relação a todos os outros segmentos do sistema. Observe que a soma dos valores percentuais dentro de um sistema (um conjunto de escolhas) deve sempre ser igual a 100%. Assim sendo, o que se está medindo é a tendência para escolher uma característica em oposição a outras do mesmo sistema.

Significância Estatística: quando um estudo comparativo é realizado, é possível medir se as diferenças entre dois conjuntos de dados são estatisticamente significantes, ou seja, se essas diferenças representam diferenças reais ou se são aleatórias.

O CorpusTool usa dois testes de significância estatística:

- ***T-Stats*** (Teste-T): *T-Stats* são os valores dos quais o nível de significância dos resultados deriva. Quanto maiores esses valores, maior é o nível de significância. Porém, isso depende da quantidade de dados no seu projeto. Para alguns estudos científicos, pode ser necessário apresentar estudos baseados em *T-Stat*, mas talisocorrências ainda são bastante raras em Linguística.
- ***Chi-Squared*** (Qui-Quadrado): recentemente, em particular na Linguística, o teste *Chi-Squared* tem-se tornado o teste de significância preferido. No CorpusTool, para cada comparação, é possível obter os valores resultantes do *Chi-Squared*, bem como o nível de significância correspondente.

Para cada linha de comparação, sempre à direita de cada valor estatístico, há um espaço em branco ou a presença de uma a três cruzes (+). Esse sinal indica quão significante é a diferença em questão:

(nenhum)	Não apresenta diferença significante.
+	Representa significância de 90% (10% possibilidade de erro).
++	Representa significância de 95% (5% possibilidade de erro).
+++	Representa significância de 98% (2% possibilidade de erro).

Por exemplo, numa comparação, se o valor de ocorrências for 10 e se for atribuída uma única cruz, isso quer dizer que é possível, e esperado, que uma ocorrência seja um falso resultado e nove sejam verdadeiros (90% de significância, 10% de possibilidade de erro).

O nível de significância estatística é importante para definir se os seus resultados podem ser reproduzíveis. Resultados sem significância podem ser arbitrários, e se o

estudo for repetido com outros textos, os resultados podem ser distintos. Se os resultados forem altamente significantes (98%), é provável que se mantenham mesmo com outro conjunto de textos.

5 Apresentando Resultados numa Rede de Sistema

Quando você realiza um estudo baseado em características, é possível ver os resultados dispostos na rede de sistemas, em vez de dispostos em tabela (Figura 7.3). Depois que um estudo é apresentado em forma de tabela, aparecerá a opção *View As*. Então, selecione *Network* para mudar a visualização de tabela para a visualização de rede.

Esse modo de visualização teve como base um recurso semelhante do SysFan[1]. Agradeço a Canzhong Wu, autor do SysFan, por me permitir usar esse recurso.

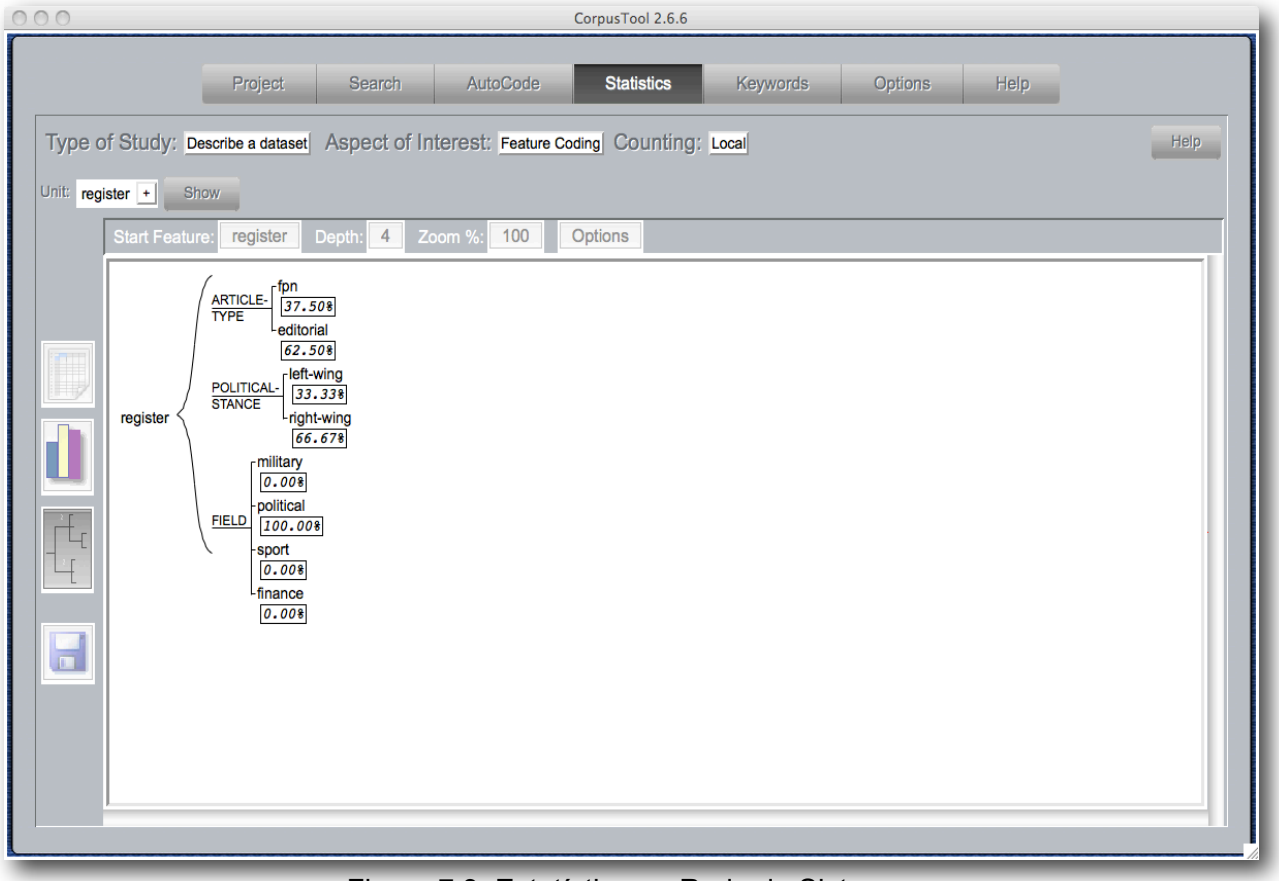


Figura 7.3: Estatística na Rede de Sistemas

6 Salvando Informações Estatísticas

O painel *Statistics* permite salvar os resultados como html, tabela delimitada por tabulação (*tab-delimited*) ou texto simples. Os resultados em html podem ser abertos no MS Word, e, para tê-los no seu artigo, basta copiar e colar.

[1] Disponível em http://minerva.ling.mq.edu.au/units/tools/index.htm

Os resultados em tabela delimitada por tabulação (*tab-delimited*) podem ser abertos no MS Excel (no Windows, dê duplo clique no arquivo .txt e especifique Abrir com (*Open with…*) Excel. Esse tipo de arquivo pode ser útil em outros programas, como SPSS.

Seção 8:

Palavras-chave

7 Identificando Palavras-chave

.

Considere o exemplo abaixo, que apresenta as palavras-chave de um *corpus* separado em três grupos. As palavras estão ordenadas pela frequência. O valor 100 indica que a palavra aparece 100 vezes mais neste *corpus* do que em outros corpora.

NOTA: para a identificação de palavras-chave, deve-se selecionar um *subcorpus*. Caso selecione o *corpus*, não haverá resultados.

Military		Economics		Crime	
troops	100.0	economy	121.38	crime	142.85
weapons	100.0	companies	116.52	detetive	50.0
engine	100.0	stock	100.0	police	49.16
mountains	100.0	tax	100.0	disappearance	40.0
smoke	90.0	cuts	85.0	criminal	39.86
gulf	85.0	profits	80.0	court	34.88
enemy	85.0	investment	75.0	justice	30.23
aircraft	80.0	billion	75.0	driver	30.23
force	70.0	returns	70.0	boy	29.06
civilians	70.0	sales	70.0	victims	18.6
civilian	70.0	earnings	65.0	family	17.56
guys	65.0	investors	65.0	child	13.95
military	62.47	jobs	65.0	car	12.81
squadron	60.0	package	65.0	lived	11.96
suicide	55.0	assets	65.0	officers	11.96
tanks	55.0	prices	60.0	legal	11.51
soldier	55.0	bill	60.0	children	10.57
jungle	55.0	corporate	60.0	kids	9.3
altitude	55.0	stocks	58.26	mercy	9.3
strikes	55.0	markets	55.0	investigators	9.3
trees	55.0	budget	55.0	woman	9.01
lieutenant	55.0	finance	50.0	murder	8.52
withdrawal	55.0	volatility	50.0	boys	8.52
missile	55.0	reforms	45.0	age	7.77
bomber	50.0	commercial	40.0	victim	6.64
invasion	50.0	temporary	40.0	street	6.27
combat	50.0	cent	37.87	body	6.22
rounds	50.0	analysts	32.04	incident	5.98
missions	45.0	growth	32.04		

8 Grupos

Em vez de se identificar palavras isoladas, a análise baseada em *n-gramas* identifica as sequências de palavras que são mais comuns em um *corpus*. Abaixo, está uma lista de 3-gramas (sequência de 3 palavras) identificadas em um *corpus* de artigos acadêmicos:

in terms of	12	ad hoc networks	6
a set of	11	we believe that	6
in this paper	10	of this paper	6
the performance of	7	terms of a	5
of the two	7	in section 4	5
be able to	7	some of the	5
a number of	7	in order to	5
the design of	7	large number of	5
which can be	7	that can be	5
the problem of	6	ad hoc network	5

Segundo Biber (Biber and Barbieri 2007), em um *corpus* de maior extensão, especialmente com mais de um milhão de palavras, os grupos de palavras mais comuns não são aqueles de valor lexical, mas sim grupos usados para enquadrar outros significados, como “in terms of”, “a set of”, etc.

Enquanto as palavras-chave nos indicam quais palavras devem ser ensinadas na produção de um determinado texto, n-gramas nos indicam quais grupos devem ser ensinados. Por exemplo, se tivermos de ensinar como escrever artigos acadêmicos, é possível montar um *corpus* de textos acadêmicos e extrair dele os principais n-gramas, de vários tamanhos. A partir daí, saberemos quais são os grupos mais frequentes, tais como “this paper reports on” ou “this paper/article is organized as follows”.

Seção 9:

Visualização de Textos Anotados

1 Visualizando um Texto Anotado

No CorpusTool, é possível visualizar as diferentes anotações aplicadas a um texto usando recursos estilísticos. Para isso, basta especificar quais segmentos devem estar em negrito, em itálico, sublinhado, numa fonte maior ou coloridos. A Figura 8.1 mostra um texto com diferentes segmentos, cada um destacado através desses recursos estilísticos.

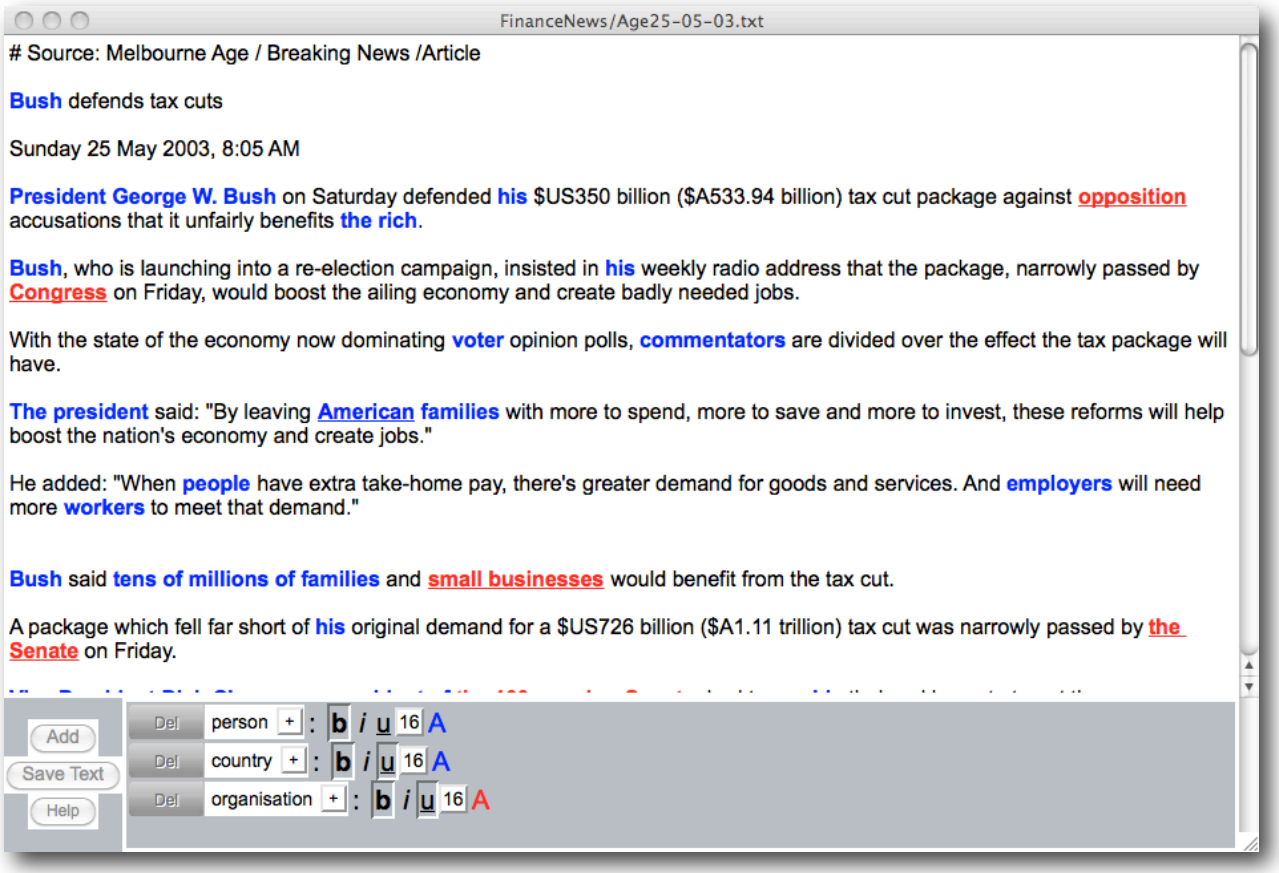


Figura 8.1: Visualização de Texto Anotado

2 Abrindo a Janela de Visualização

No painel *Project* (painel principal), clique no nome de um dos ficheiros. Note que o recurso de visualização só está disponível aos textos incorporados. Note também que é preciso que haja pelo menos uma camada de anotação já definida.

3 Selecionando Estilos

Você pode atribuir cores e/ou efeitos de fonte (negrito, itálico, sublinhado) ao texto anotado com uma única característica ou com várias caraterísticas. Esse procedimento permite visualizar padrões de seleção num texto. Por exemplo, você pode usar negrito/

itálico/sublinhado para os tipos de grupos e fontes coloridas para a oração. Assim você consegue observar como os grupos se distribuem em relação à oração.

4 Salvando uma Visualização

Você pode salvar o texto estilizado como arquivo html. Para incluir esse texto em um documento do MS Word, abra o arquivo html no MS Word, copie o texto anotado e cole no seu documento de trabalho.

Apêndice I:

Systemic Coder

1 Importando Estudos do Coder

Os arquivos de análise do Systemic Coder podem ser importados para o CorpusTool. Para isso, siga estas instruções:

Se você quer importar um único arquivo:

1. Salve o esquema de anotação como um arquivo externo (arquivo-mestre). Para isso, abra o arquivo no Coder e selecione *Scheme Storing* no menu *Options* . Selecione *Save to Master* e especifique o local para salvar.
2. Assegure-se de que as anotações estão salvas com a extensão .cd3, e não com .cd2; se o arquivo apresenta a extensão .cd2, abra-o e, no menu *File*, selecione *Save Codings As*. Você encontrará a opção para salvar como um arquivo .cd3.
3. Crie uma nova pasta e coloque nela o arquivo de esquema e o arquivo de anotações.
4. Abra o CorpusTool e crie um novo projeto.
5. Selecione *Import Layer* no *Project Menu*.
6. Indique a pasta criada em (3) acima.
7. O arquivo .cd3 se dividirá em um arquivo de texto simples (para ser colocado na pasta *Corpus*) e um arquivo de análise (para ser colocado na pasta *Analyses*). A janela seguinte perguntará a você em qual pasta de *subcorpus* o arquivo de texto deve ser guardado.
8. O esquema de análise será importado como uma nova camada.
9. No Coder, o único modo de evitar anotações é ignorá-lo. No CorpusTool, é possível selecionar apenas os segmentos de texto que se quer anotar. Se você quiser que os segmentos ignorados no Coder desapareçam, a próxima janela lhe permitirá fazer isso.
10. Clique em *Finalise*. O resultado desse procedimento será a criação de uma nova camada de análise e a importação do arquivo .cd3.

Se você tem um conjunto de textos, todos anotados sob o mesmo esquema:

1. Coloque todos os arquivos Coder numa pasta.
2. Assegure-se de que todos os arquivos estão no formato .cd3, e não .cd2.
3. Siga o passo (1) das instruções para um arquivo único, para pelo menos um dos arquivos (assegure-se de que há um arquivo com a extensão .scheme na pasta).
4. Continue do passo (4) das instruções para um arquivo único.

Se você tem um ou mais arquivos cujos textos estão anotados a partir de redes de sistemas distintos (uma espécie de anotação multicamada no Coder):

1. Para cada conjunto de arquivos anotados com o mesmo esquema, crie uma pasta e coloque nela os arquivos do Coder e o arquivo de esquema correspondente (assegure-se de que os arquivos estão no formato .cd3).
2. Assegure-se de que o arquivo de análise correspondente ao texto tem o mesmo nome. Por exemplo, se você tem dois arquivos de análise: Text1-Oração.cd3 e oText2-Grupo.cd3, ambos relacionados ao mesmo texto: Text1, renomeie os dois arquivos como Text1.cd3 (O CorpusTool só pode reconhecer que dois arquivos de análises se referem ao mesmo texto se eles tiverem o mesmo nome).
3. Crie um novo projeto e importe a camada, como descrito anteriormente.
4. Repita o passo (3) para as outras pastas.

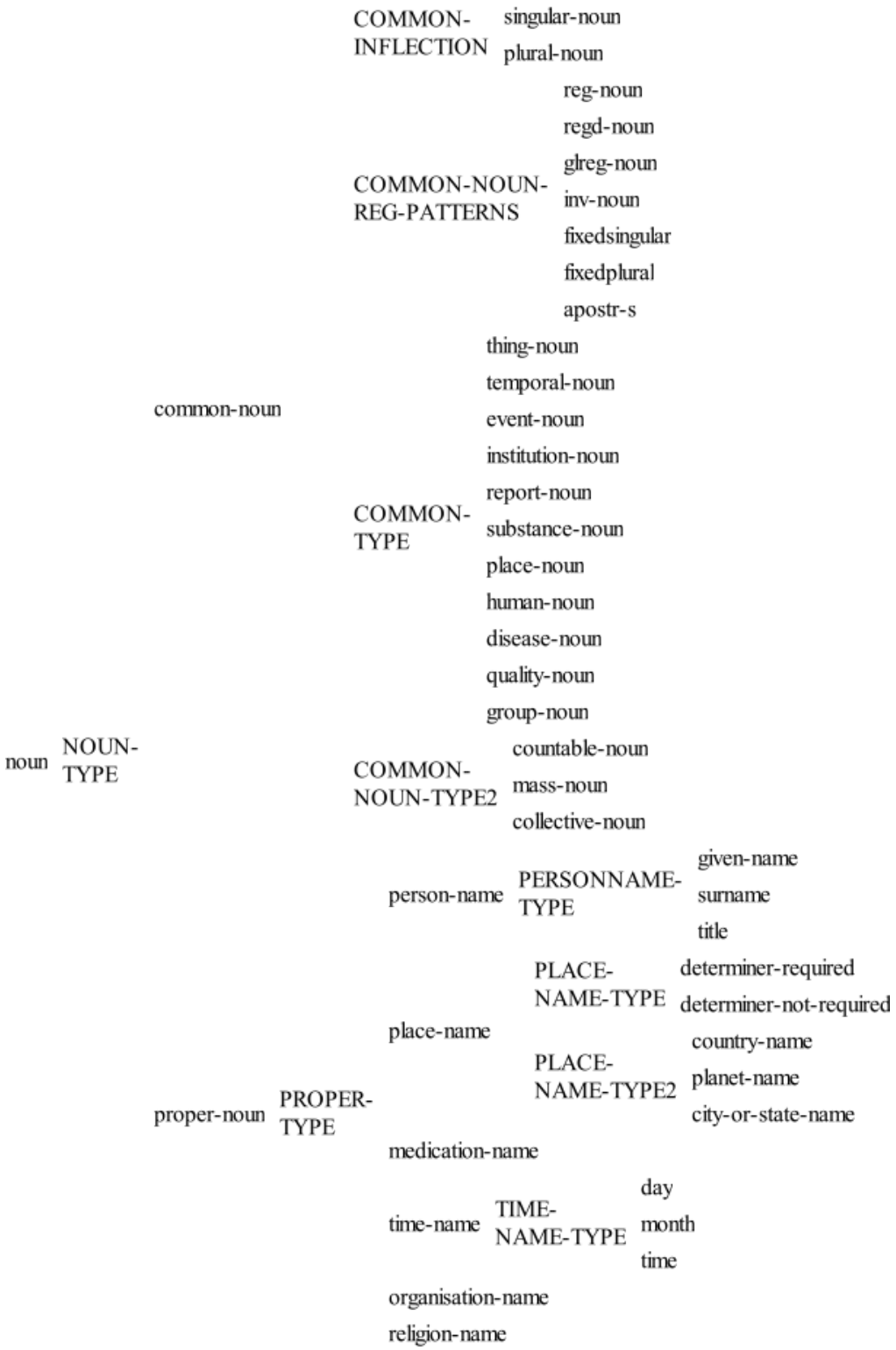
Alguns possíveis problemas:

- O CorpusTool indica que não consegue ler um ou mais arquivos .cd3: esse arquivo pode conter caracteres que não são reconhecidos pelo codificação ASCII. Esse problema ainda não pode ser resolvido pelo CorpusTool. Envie-me seus arquivos para que eu os importe para você.
- Se houver qualquer outro problema na importação dos arquivos .cd3, envie-os para mim (comprima a pasta com os arquivos) para que eu possa ajudar. Essa é uma boa maneira de estar informado dos problemas que os usuários estão tendo e, consequentemente, de os resolver.

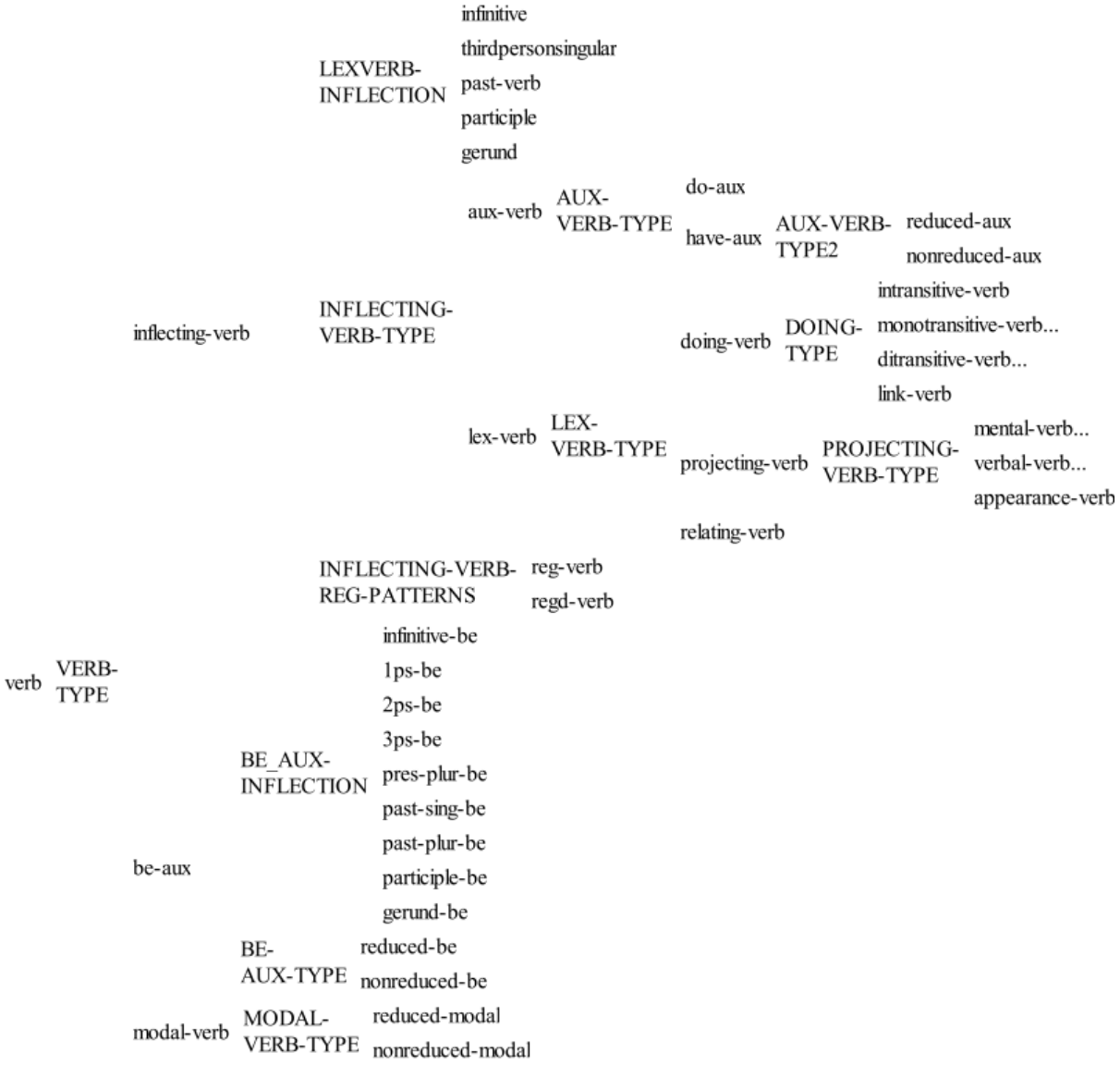
Apêndice II (em Inglês):

Recursos Lexicais para Pesquisa de Concordância

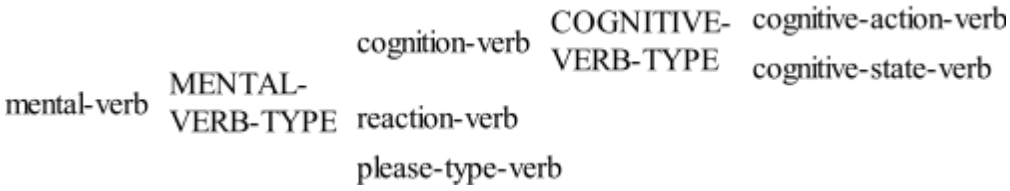
1 Nouns



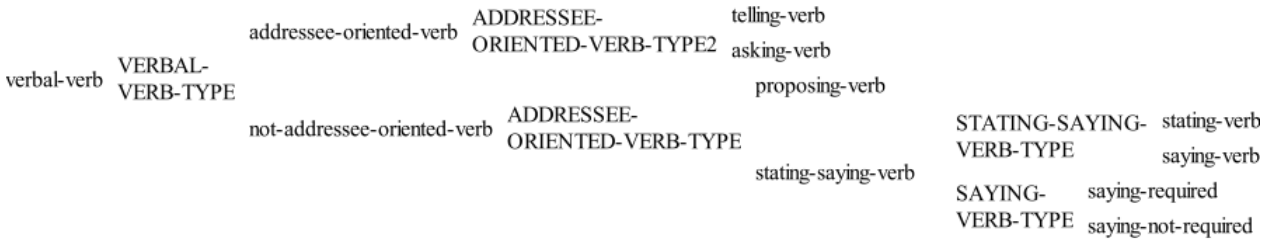
2 Verbs



Subclasses of mental verb



Subclasses of verbal verb



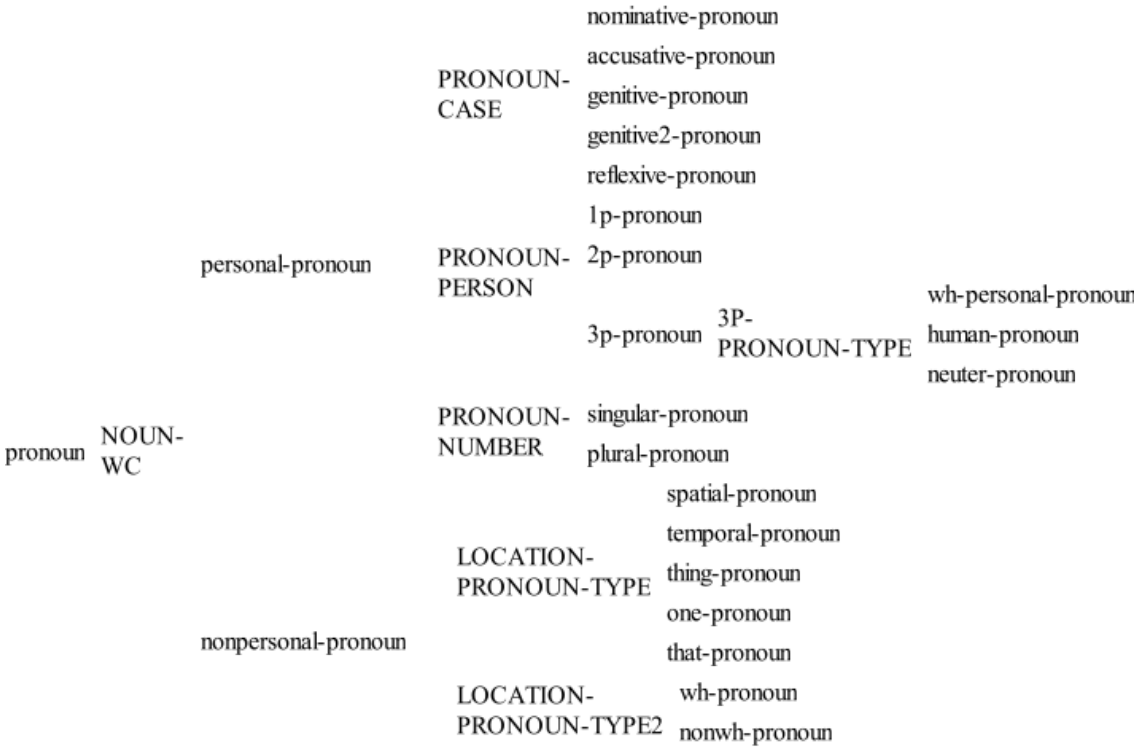
3 Adjetives

- adjective
  - ADJECTIVE-TYPE
    - er-est-adjective
      - ER-EST-ADJ-INFLECTION
        - absolute-adjective
        - comparative-adjective
        - superlative-adjective
      - ER-EST-ADJECTIVE-REG-PATTERNS
        - reg-adj
        - regd-adj
    - noninflecting-adjective
      - NONINFLECTING-ADJECTIVE-TYPE
        - comparable-adjective
        - noncomparable-adjective
  - ADJECTIVE-SEMTYPE
    - nationality-adjective
    - other-adjective

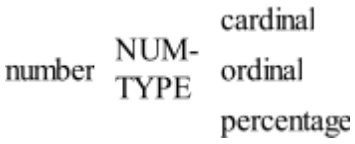
4 Adverbs

- adverb
  - ADV-TYPE
    - descriptive-adverb
      - DESCRIPTIVE-ADVERB-TYPE2
        - manner-adverb
        - temporal-descriptive
        - connective-adverb
        - modal-adverb
        - location-adverb
        - other-descriptive-adverb
    - intensifier
      - INTENSIFIER-TYPE
        - more-less
        - most-least
        - degree-intensifier
    - jussative
    - interrogative-adverb
    - ago-adverb

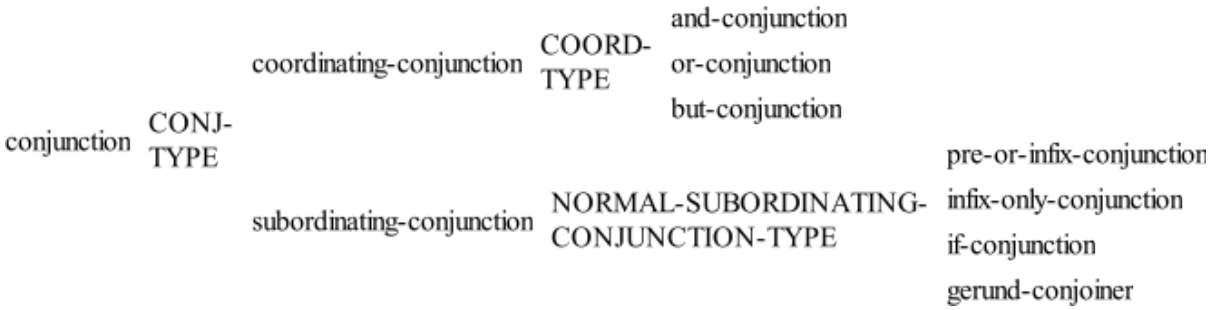
5 Pronouns



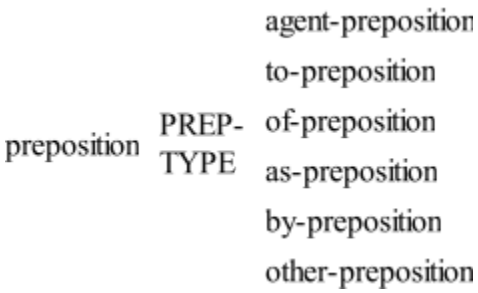
6 Number



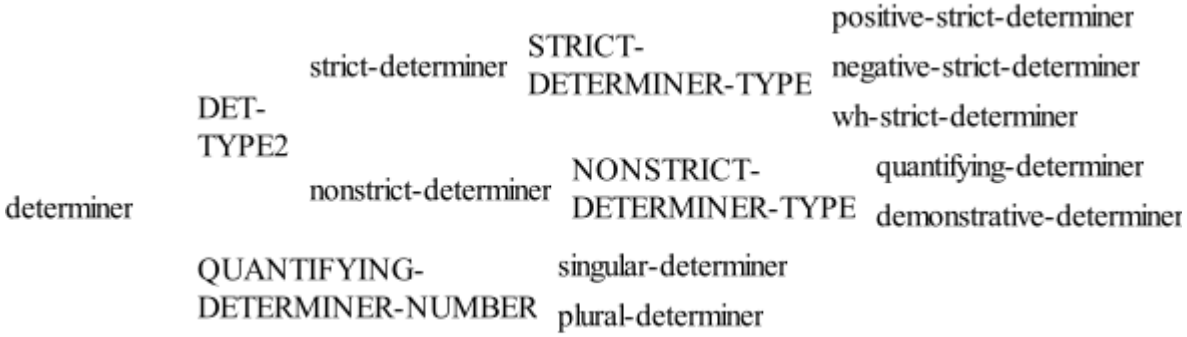
7 Conjunction



8 Prepositions



9 Determiners



10 Punctuation