UAM CorpusTool

版本 2.0 用户手册

（2008 年 9 月）

Mick O'Donnell
michael.odonnell@uam.es
(Translated by Liu Xiaohan)

目录

第一节：关于 UAM CorpusTool

1 简介

UAM CorpusTool 是文本和图像语言标注的工具集，其核心内容包括：

- 用户制定一个“项目”，即一组文档，和应用于每一文档的分析规则。
- 每一类“分析规则”可视为标注的“层系”。CorpusTool 现在有两种类型的标注。
  1. **整体文档标注**：文档（文本或图像）的整体特征标注。例如，这些特征可以表明该文档的语域（语场、语旨、语式）或文本类型。
  2. **切分段标注**：用户选择文件中的节，并分别赋予特征。节可通过鼠标在文本/图像中拖动指定，同时提示用户指定该节的特征值。

后续版本将添加其他标注类型，可以是修辞结构理论（RST），体裁结构（GSP），参与者链条（participant chaining），句子结构（比如主语、谓语、情态、附加语 adjunct 等），口语数据标注等等。

UAM CorpusTool 取代作者之前的 Systemic Coder 软件仅能在单一层系上对单个文档标注。UAM CorpusTool 是为了克服 Coder 用户诸多限制的一个尝试。我希望感谢广大 Coder 用户近年来作出的评论和本新软件的评论。参见附录一讲 Systemic Coder 研究结果导入 CorpusTool。

CorpusTool 相关在：
http://www.wagsoft.com/CorpusTool/
访问该网站以指导 CorpusTool 在电脑中的安装。

第二节：项目的创建

1 创建一个新的项目

1.1 打开 CorpusTool

UAM CorpusTool 在电脑中安装后即可工作。首先要创建一个新的“项目”：

Windows：

安装 CorpusTool 时可选择放置图标到桌面。点击此图标启动 CorpusTool。

另外，在开始菜单中的程序菜单含有 UAM CorpusTool 图标。选择以启动 CorpusTool。

Macintosh：

CorpusTool 安装在应用文件夹中，双击以启动 CorpusTool。

也将该应用程序置于 Dock 方便使用。

（在已创建项目后，可双击项目文件夹中的.cptr 文件打开。该文件图标如下：

MacOSX： 

Windows： 

启动窗口

窗口应如图 2.1 所示。所使用软件版本号在窗口中显示（在交流缺陷时有用）。窗口提供的选项有“创建新项目”或“打开项目”继续已创建的项目。如果电脑之前打开过某一项目，此项目也会出现在按钮上。

图 2.1 启动窗口

1.2 点击“创建新项目”按钮
点击此按钮后，用户被询问标注“文本”还是“图像”。CorpusTool 可标注纯文本和图像（但不能在同一项目中同时标注两者）。
选择其中之一后，会出现一个“创建项目向导”，做必要的步骤引导。
1. 为新项目提供名称
2. 指定新项目存储的文件夹。比如电脑的桌面文件夹
3. 文本文件：用户将被问及是否向项目添加文本文件夹。文件夹中应是纯文本文件，扩展名都是.txt。此文件夹将被复制于项目文件夹中的语料库文件夹中。也可跳过此步以后添加。
点击“完成”按钮后 CorpusTool 就创建了项目，该文件夹中含有与项目相关的所有细节，包括语料库、标注文件和一个直接启动该项目的图标（.cptr 文件）。
一旦完成“创建项目向导”，CorpusTool 主窗口就会打开，出现“项目管理面板”。参见图 2.2。利用此面板可控制项目的细节，比如增添文件和分析类型。

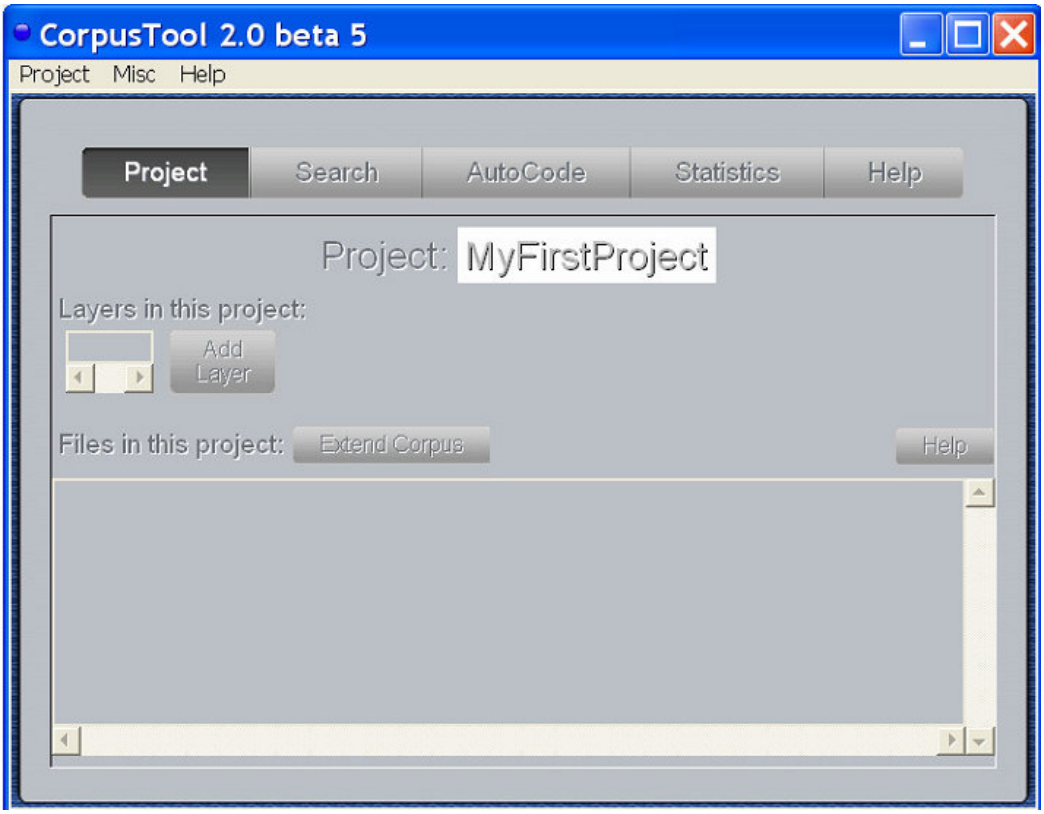


图 2.2 项目管理面板

面板上部的按钮可在 CorpusTool 不同面板间切换：搜索（第五节）、自动标注（第六节，只针对文本标注）、统计（第七节）和帮助。
现在我们选中的是项目面板。顶部的大字显示的是项目名称。
下面显示何种分析（层系）使用于项目。最初为空。
“层系”往下的区域显示项目中的所有文件（最初为空），每一文件旁都有可能用到的分析按钮。
我们首先来给项目添加一个层系。
2 添加层系（layer）

建立新项目后首先要确定所需何种分析。我们来开始添加一个层系。

1．点击“添加层系”按钮

“层系”是对文本文件的一种分析类型。添加的层系可以标注小句、集合、整个文本的语域、评价系统分析等。

点击“添加层系”按钮后即有窗口弹出，询问几个问题，用“下一步”按钮在问题间切换：

- 层系名称：层系的名称。输入“语域”。
- 标注对象：在这里指定是将文本作为整体赋值（如语域、文本类型）（文档标注），还是文本中的切分段赋值（如小句）。这里我们假定对前者感兴趣，选择“文档标注”。
- 标注体系：标注体系是对文本标注的特征描述。有如下两个选择：
    i. 创建新体系：用户多数情况下关注于制定自己标注体系，展示所感兴趣的特征，并进行组织。CorpusTool 提供简便界面来创建和修改这些体系（参见第三节）。
    ii. 复制已有体系：有时用户可以再次使用自己或别人以前开发的体系。CorpusTool 预装了几套体系供使用，其中有 Peter White 的评价网络和基于 Granger 的错误标注体系。

作为帮助文件，我们这里选择“创建新体系”。再点击“完成”按钮，新的层级即可添加到项目窗口。

图 2.3 展示添加一个层级后的项目窗口。层级区域提供层级的信息：名称（语域）、类型（文档标注）和与层级相关的体系名（语域.xml）。

层级控制面板上有两个按钮：

- 删除：删除层级和该层级在文本中的所有分析。一般在层级标注之前使用该按钮，否则将真正删除次层级。
- 编辑：点击打开窗口以编辑标注体系。下节中会详述。

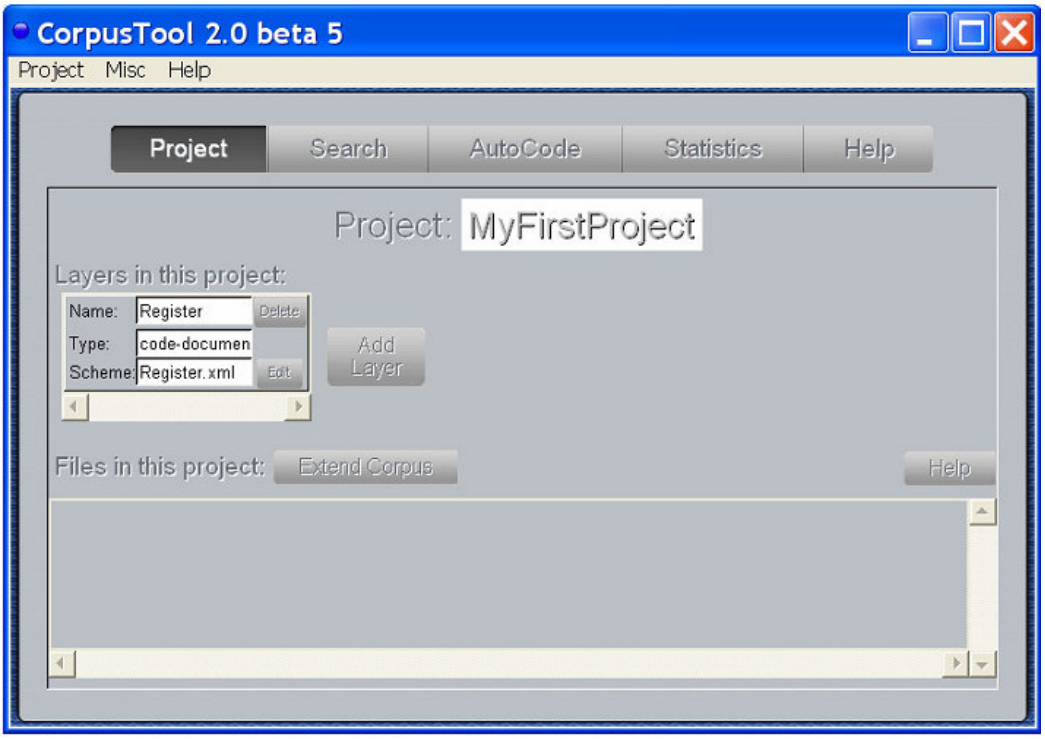


图 2.3 已添加层级的项目窗口

选择“导入层级”用以添加 Systemic Coder 中的层级（.cd3 文件）。详述于附录一。

3 添加文件以供分析

下一步是向项目中添加文件。如果在项目创建时已经指定文本文件到项目中，在项目窗口的文件面板中就能显示出来。现在假定此项工作尚未进行，这样文件面板如图 2.3 所示为空。

3.1 扩充语料库

添加文件到语料库：

1. 点击“扩充语料库”按钮：“向导”会一路引领添加文件。添加文件可以是单独文件，也可是一个文本文件夹。
2. 若选择添加单个文件，既可以将其添至已有的子语料库（项目语料库文件夹中一个文件夹），也可以添至新的子语料库（此时需提供新文件夹名称）。若选择添加文件夹，待磁盘特定文件夹指定后，被复制至项目下的语料库文件夹。
3. 文件或文件夹指定后点击下一步和完成按钮。

文件将显示于文件面板中（见图 2.4）。新加文件有标示“文件在语料库中但未加入项目”。CorpusTool 区分“**已加入**”文件，即带有所有可标注按钮，和“**未加入**”文件，即存于语料库但尚未供标注使用。

这种区分可以容易地跟踪已开始编辑的文件，和那些将来再编辑的文件区别开来。假如语料库中有 100 个文件，但只标注了 5 个，这 5 个标注的文件能清晰显示。语料库逐步扩充虽然需要较长时间，但结果在每一阶段都能获得。

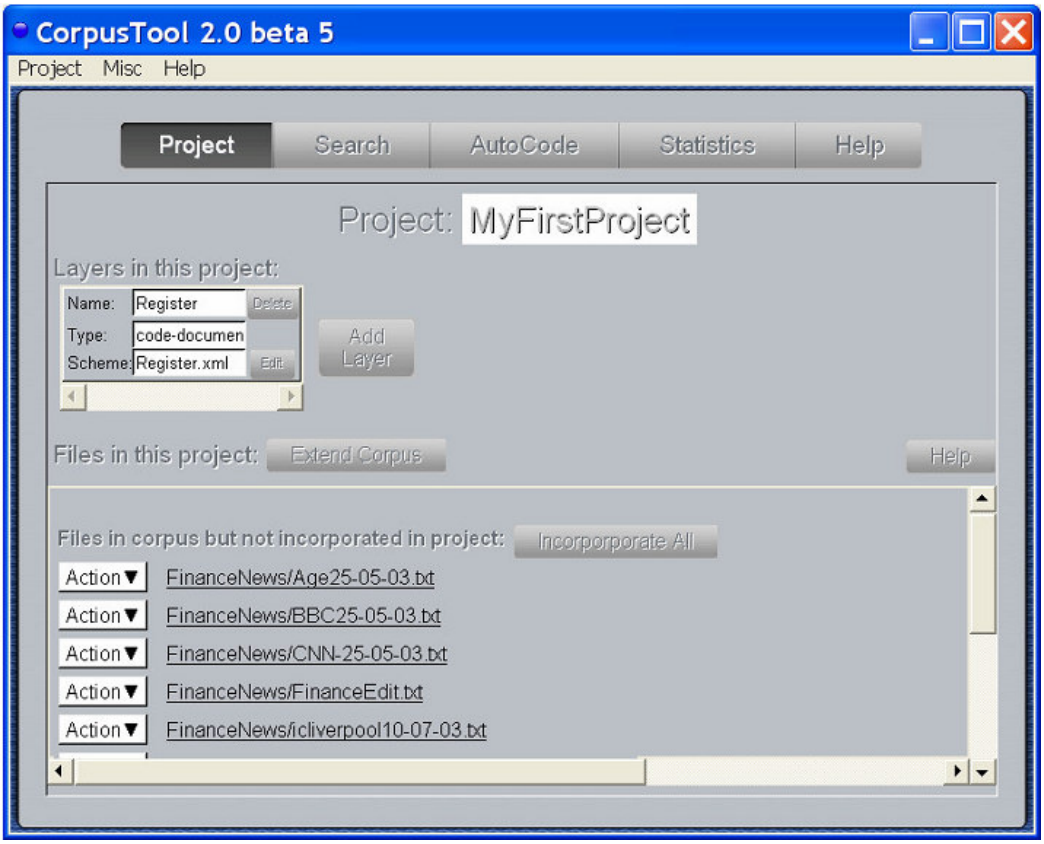


图 2.4 扩展语料后的项目窗口

3.2 加入文件

加入文件到项目中，使其能够标注，则点击文件旁边的“**加入**”按钮。

确定**语言、编码和显示字体**：在加入文件时窗口中会询问关于文件的某些元数据（参见图 2.5）。包括：

语言：文本的书写语言是什么？本项可确定文档的语言资源，包括词汇（供检索查询、词汇密度计算）、分析器（供自动切分）和赋码。英语是现在唯一支持的语言，其他语言的词汇资源将后续提供。

编码：文本文件由特定文本编码所储存。选择此项告知 CorpusTool 文件所用编码。CorpusTool 提供的缺省选项仅为猜测，若文本不能正确显示应当修改。为获得文档编码，可尝试右击文件，选择“打开方式”（或 MacOSX 相应选项），用 MS Word 打开，可帮助选择最佳编码。否则，在“打开方式”中选择 Firefox，在 “视图”子菜单下选则“字符编码”，看是采用的何种编码。

显示字体：从此项选择字库和字号，将文本显示于标注窗口。有些字体最好由非西方书写体系处理，如一些字体是为显示中文而设等。但许多现代字体应能显示任何的书写体系。

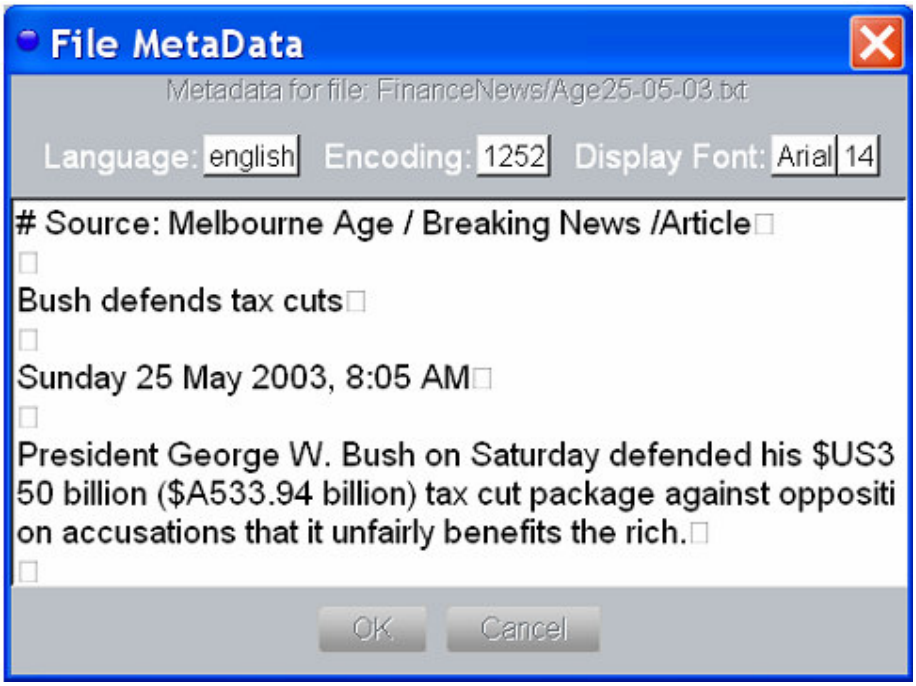


图 2.5 文件元信息窗口

加入两个文件后，项目窗口如图 2.6 所示。注意这两个文件出现在顶部。

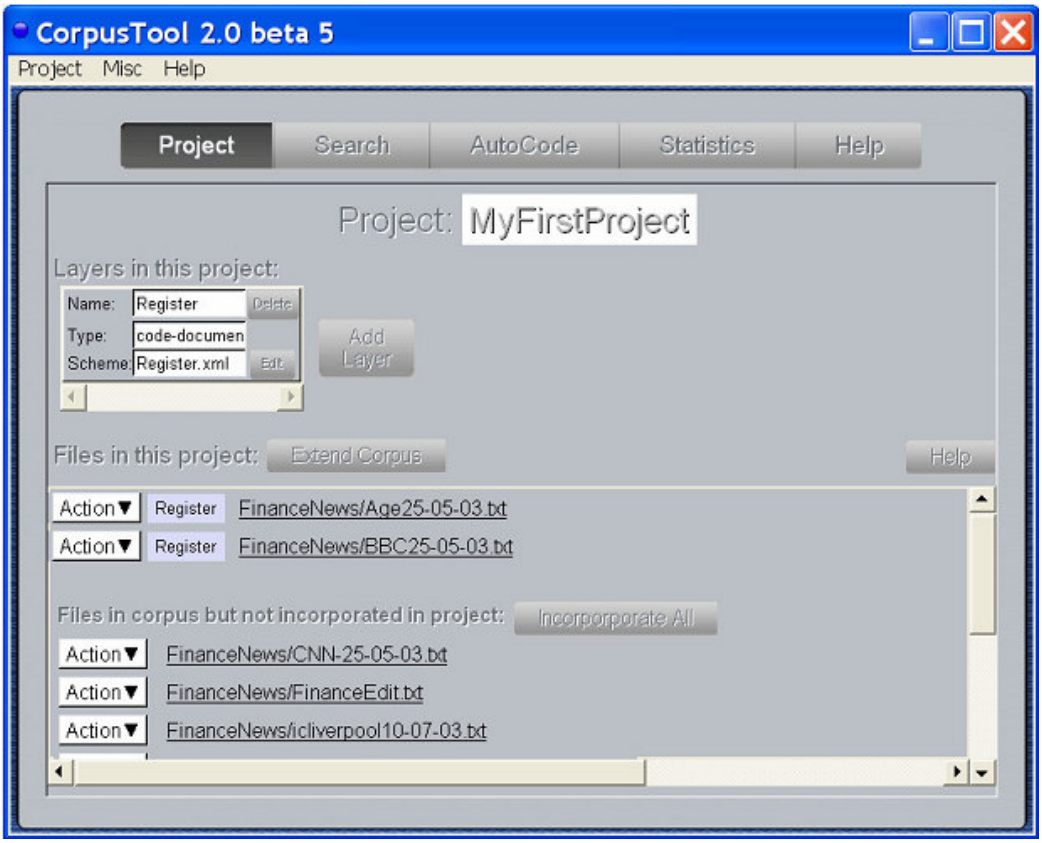


图 2.6 加入文件后的项目窗口

3.3 “撤销文件”的其他选项

“撤销文件”的其他选项有：

- **信息**：提供文本文件的一些统计信息，词数，句数，平均句长等。英文文件还有词汇密度测量，代词使用的一些描述（见下）。
- **删除**：从语料库中移去文件。同样从项目下的语料库文件夹中删除文件。
- **文件名**：点击文件名显示整个文件。

4 “加入文件”按钮（操作）

文件添加后，对每个层系都有相应按钮。在本样例项目中，我们现在为止只定义了“语域”，所以加入的文件只有一个按钮。随着其他层级添加，相应按钮也会出现。

4.1 更改文件元数据

（仅针对文本标注）我们上面看到当“加入”文件时，需随即确定其语言、编码和显示字体。这些选项的更改可随时通过选择文件相关联的“按钮”菜单中的“更改文件元数据”来实现。

4.2 浏览文件一般统计信息

（仅针对文本标注）每个文本文件的一般统计信息可通过选择每行按钮菜单下的“浏览基本文本统计”观看。所提供的基本文本信息**不**依赖于文件的任何标注（参见图 2.7），包括有：

- 文本的词数
- 平均词长
- 文本句数（在欧洲语言下）
- 平均句长的词数（同样在欧洲语言下）

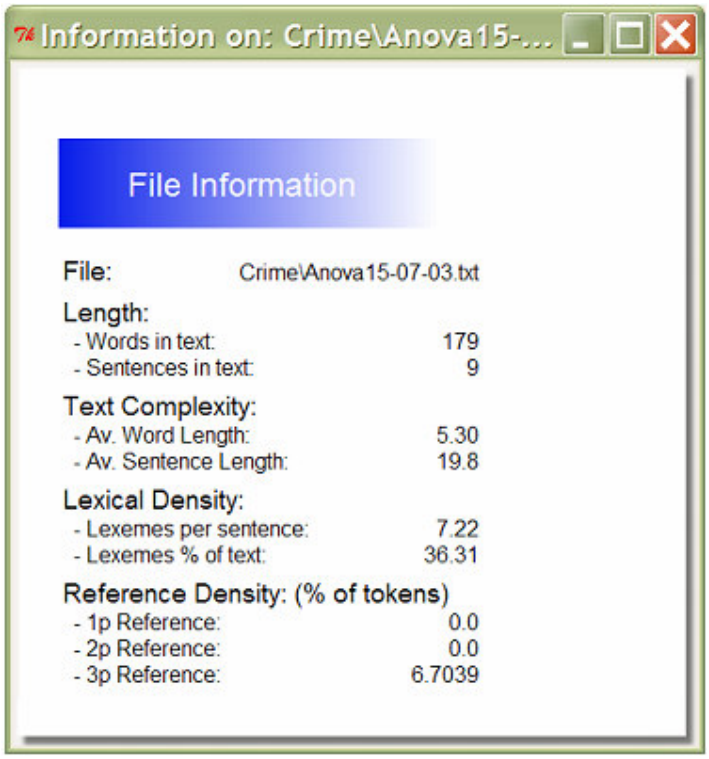


图 2.7 文件的信息窗口

对于**英文**文本，信息还有：

- **词汇密度**：平均每句中的开放词类数，或整个文本的%开放词类项。
- **代词指称密度**：第一、二、三人称代词使用的详情，以整个文本为参照的百分比。

注意：随着其他语言词汇的加入，其语言的相关统计也将会提供。

4.3 从语料库撤销文件

“撤销”按钮将文件从研究中移除。

警告：该文件以往的标注会被删除。文本文件会放入“未加入”列表，可以后再加入（但完全没有标注）。

4.4 打开标注窗口

每行其他按钮各对应项目中一个标注层级。点击打开文本在此指定层级的标注窗口。

按钮颜色：文档每层级的按钮有颜色标示，显示完成程度。

- 白色：全部标注
- 浅蓝：部分标注
- 深蓝：较高程度标注

注意这些颜色仅起提示作用。

5 退出 CorpusTool

注意所有项目改动都是自动保存的。如果退出项目管理窗口（点击右上角 X），即可退出 CorpusTool，改动全部保存。

6 继续某项目

项目创建后，打开 CorpusTool 处理项目最简捷方式是：

1. 打开桌面上的工程文件夹
2. 双击.cptr 文件（为蓝色球形图标）

CorpusTool 即直接打开项目窗口。

撤消操作：现在不支持撤消操作。后续版本将会支持。

第三节：制定标注体系

1 打开“体系编辑器”

根据某一指定层系标注文件之前，需要为该层级制定标注体系。第一步是打开体系编辑器。在层系工具栏中点击“编辑”（见图 3.1）。

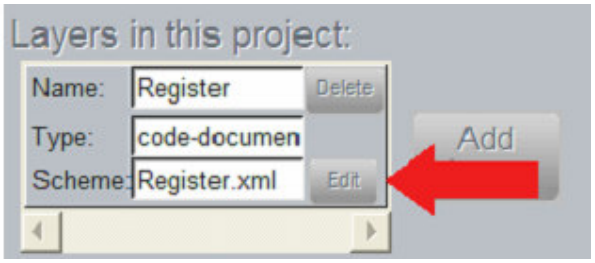


图 3.1 体系编辑按钮

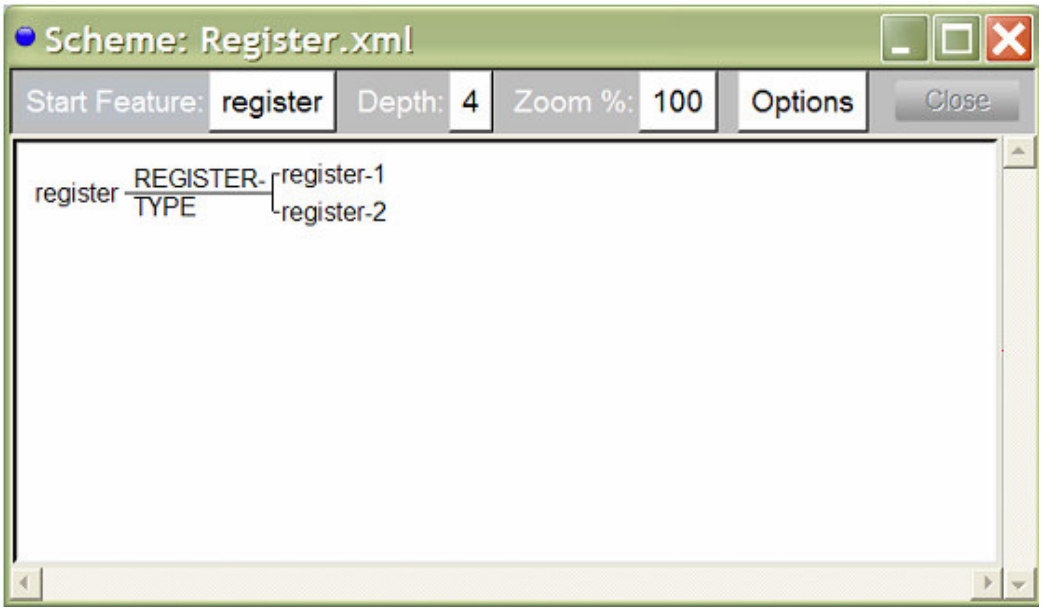


图 3.2 编辑前的语域体系

弹出如图 3.2 的窗口。它展示一个小的“系统网络”（特征层次），“语域”是最起始的概念，有语域-1 和语域-2 的选择项。

2 编辑体系

这些特征是自动生成的，我们将改动特征以提供更多信息。

点击“语域-1”会有待选择的菜单出现，如图 3.3。

图 3.3 特征选项

这些选项以后会详细解释。现在，我们要更合理地更改“语域-1”。假设所有文本都是新闻，属于头版新闻或评论。因此我们要将“语域-1”更改为“fpn”，“语域-2”更改为“editorial”。

相关选项是“特征重命名”。点击该选项。弹出窗口询问该特征值的新名称。键入：fpn 然后点击“返回”。

对“语域-2”同样操作，更名为“editorial”。

另需注意，在 fpn 和 editorial 的选择，系统自动提供“语域-类型”。将其重命名为“文章-类型”。点击“语域-类型”，选择菜单中“系统重命名”。

标注体系可以相当复杂。图 3.4 的体系相对更为复杂，能容扩几百个选择项。但是现在体系越小，标注就越快。

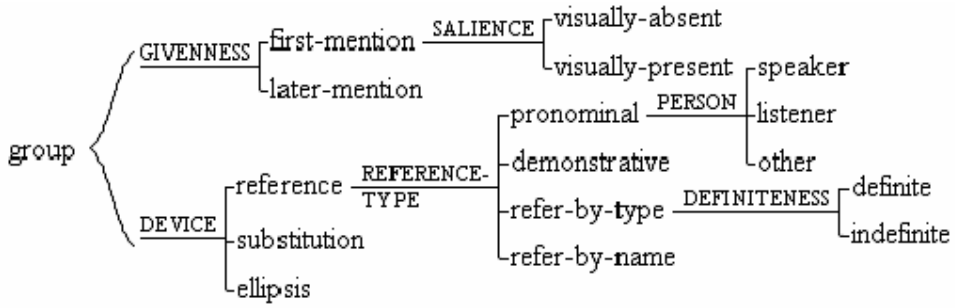


图 3.4 更复杂的体系

2.1 系统网络

CorpusTool 使用的层次体现来自系统功能语言学。这种层次成为系统网络。它包括数个相互依存的选择节点，称为系统。图 3.2 所示系统就是一个系统。图 3.4 显示进入网络的由 6 个系统组成。

一个系统由三部分组成：

- **系统名称**：即选择项的名称。语法中典型的名称可以是情态、极性、限定性等。系统名称由字母串组成，还可有数字和连字符，但不允许有空格。系统名称须是唯一的——CorpusTool 不允许两个系统使用同一名称。软件自动将系统名大写。

- **特征**：即各个选项。上例中，editorial 和 fpn 是系统的特征。无论如何键入，均以小写显示。同一名称不能在两个系统中使用。
- **入列条件**：每一系统都有入列条件，此特征（或复杂特征组）构建语境，使得选择相关联。上例中系统入列条件是“语境”，恰好是系统网络的根特征。

几个系统可拥有同一入列条件，这时的系统称为同时系统。由入列条件形成交叉分类。比如引入另一系统，也将语域作为入列条件，其特征可以是金融、军事、体育等。

制定的系统集合形成一个系统网络，其中一个体系的特征成为更精细系统的入列条件。如何创建这些网络在下文中描述。

2.2 创建、修改系统

点击网络其中的一个特征值（小写）或系统（大写），启动一个弹出操作菜单。这些操作可对网络扩展或修改。

2.2.1 对系统的操作

- 增添特征：增添新特征到系统。
- 重命名系统：更改系统名称。
- 删除系统：从网络删除系统。注意：同时属于该系统的特征和依存与该系统的系统也被删除。并且现在无法撤销。已被特征赋值的任何标注也将被删除该特征。
- 更改入列条件：更改系统的入列条件为另一特征。
- 上移：移动系统到图像的较上一层，重组显示。
- 下移：移动系统到图像的较下一层，重组显示。

2.2.2 对特征的操作

- 增添系统：创建该特征下一个虚拟系统。
- 重命名特征：更改特征名称。
- 删除特征：从系统删除特征。注意：依存于该特征的任何系统也将被删除。并且现在无法撤销。任何文本之前若被该删除特征所标注，其特征将从标注中删除。
- 上移：移动特征到系统较上一层。注意系统中的标注中的缺省值是第一个特征。
- 下移：移动特征到系统较下一层。
- 编辑实现：可对特征附加实现。程序不对其操作，但借此可以标注特征，如对选择项的释义。
- 显示实例：文本标注后选择此项将打开语料库搜索窗口，显示库中所有赋予该特征的实例。

2.2.3 更改入列条件

更改系统的入列条件，点击系统选择“更改入列条件”。出现图 3.5 对话框。

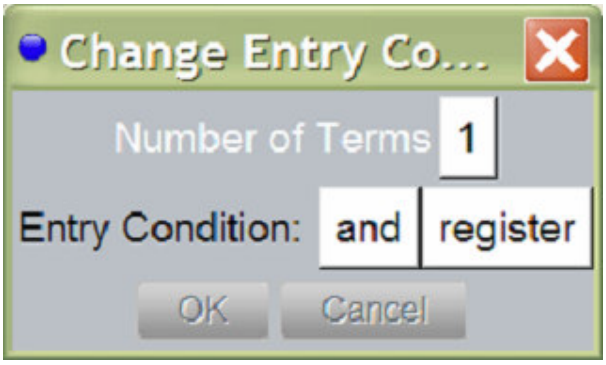


图 3.5 更改入列条件对话框

简单入列条件：若需要一个简单入列条件（系统从单一特征扩展），将“条件数目”设为 1，再选择作为入列条件的特征。点击“完成”后图像相应重绘。CorpusTool 自动更新词更改所涉及的标注。

复杂入列条件：也可以引入网络复杂入列条件。一个复杂入列条件由特征间的合取（“和”）或析取（“或”）组成。设置条件数目为 2 或更多。再选择输入特征。注意现在还不能在同一入列条件中混合使用 AND 和 OR。但可以首先伪装为一个“门”（仅有一个特征的系统）。比如要构造“A 和（B 或 C）”的入列条件，首先创建特个特征的系统（称为 b-或-c 特征）。再将此特征和 A 特征构成 AND 作为初始系统的入列条件。此时系统的入列条件就是“A 和（B 或 C）”。

2.2.4 移动特征到另一系统

若将一个特征从一个系统移动到另一系统，点击需增添特征的系统，选择“增添特征”。键入要移动的特征名称。该特征就被移至此系统。所有标注随更改作出调整。

3 增添“释义”到特征

可以对每一个标注特征做出描述。点击特征选择“增添释义”。弹出窗口可以输入。输入特征描述，即选择标准。此释义在标注文档时可显现（见下）。

4 选项菜单

每一个体系菜单都有“选项”菜单，可以：

- 另存为…：保存体系到单独文件夹。
- 显示/隐藏系统名称：选择此项可以隐藏或显示特征名称。若隐藏系统名称，系统编辑就更为困难（通常点击系统名称进入“增添特征”等功能）。
- 保存图像为 PDF：选择保存当前显示的网络为 PDF 文件。
- 保存图像为 SVG：保存当前网络为 SVG（Scalable Vector Graphics）格式。下文详见 SVG。
- 复制到剪切板：（仅对 Windows）复制显示图像到剪切板。可以粘贴到 MS Word 或其他程序。注意，依微软标准，复制前需打开 MS Word，否则无法粘贴。

5 生成图像嵌入文档或网页

SVG 不为广泛支持，可转换为其他格式，因为它是以几何形式而不是位图形式存储的。

生成 SVG 文件的其他格式，下载安装 InkScape。该软件免费，运行于 Windows，Macintosh 和 Linux 系统。下载地址：http://www.inkscape.org/

打开 InkScape，选择文件菜单的“打开”。选择.svg 文件。

文件在此也可编辑。

有两种方法保存为另一种格式：

1. 选择“另存为”保存为 PDF，EPS，EMF，或其他基于矢量的文件格式。
2. 选择“输出位图”保存为 PNG 格式，作为位图格式可嵌入网页或 Word 文档。图 3.6 是通过 InkScape 得到的 PNG 文件：

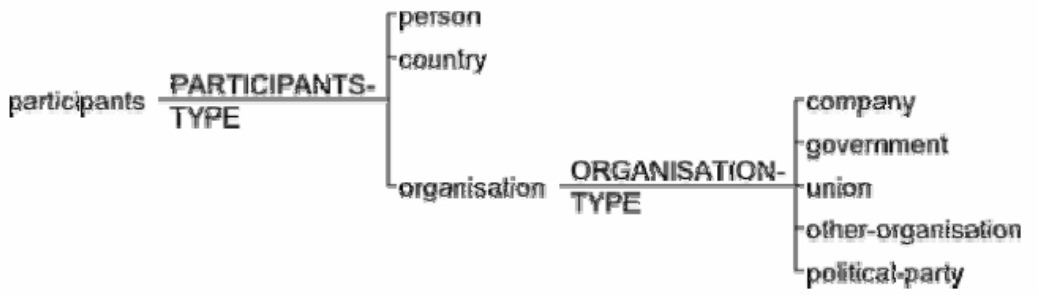


图 3.6 PNG 文件输出

第四节：标注文件

1 标注类型

CorpusTool 现在支持两种标注类型：

1. 文档标注：文档作为整体被赋特征。对于确定文档语言、文本类型、语域等有用处。也可标注作者特征（如语言能力）。
2. 切分段标注：用户确定文档中的切分段，并对其赋予特征。例如小句、NP，词，话轮等。

以下介绍对两种方式的标注。

2 标注文档整体文件

加入项目的每一个文本文件都有按钮进行层系分析。点击标明“文档标注”的层系按钮，就出现图 4.1 窗口。

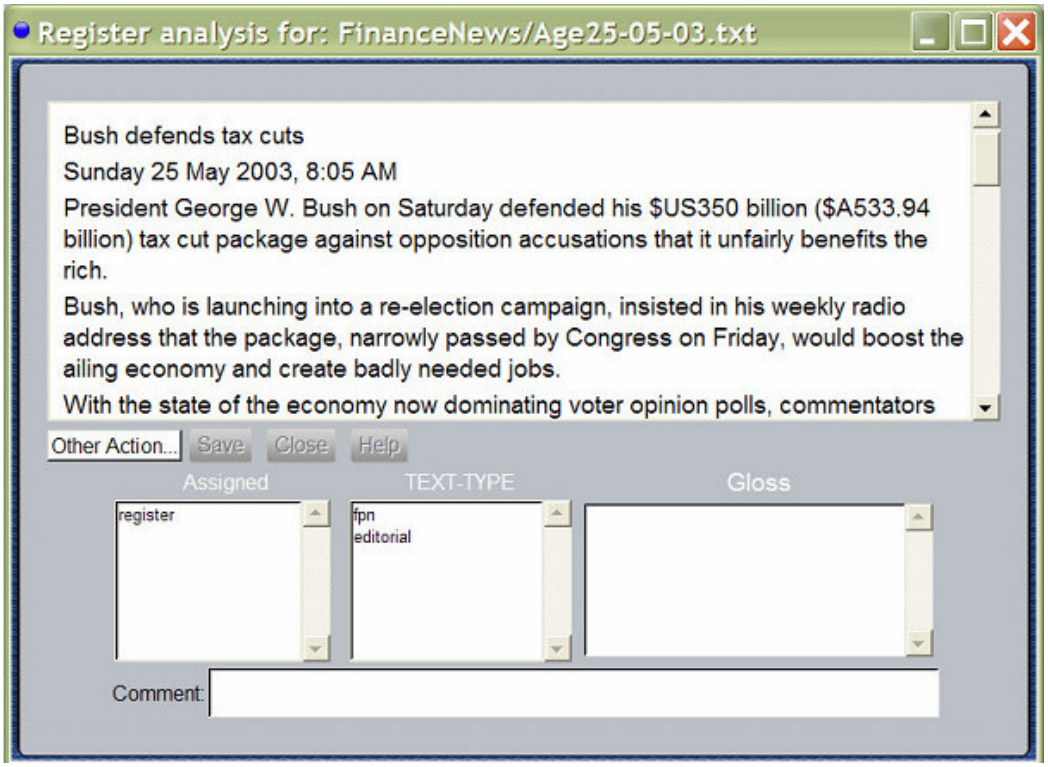


图 4.1 文档标注窗口

文档标注窗口有四部分：

1. 文本框显示文本文件。滚动鼠标可显示整个文本。
2. 工具栏：提供各类操作，如保存、关闭和帮助（见下）。
3. 标注框有三栏：
   a. 已选特征（标签名“已赋值”）：文本已赋值的特征。最初只含有一个特征，即层系标注体系的最左侧（“根”）特征。双击已选特征可删除选择。根特征不能被删除，因为它用作所有文档缺省值。
   b. 当前选择：中间栏需要对文档做选择。双击其中一个选项，就将其移至已选特征栏。如果标注体系有更多选项，下一个选择又会显示。
   c. 释义栏：若已在体系中引入特征释义（见上第 3.3 节），单击当前选择的该特征，其释义在此栏出现。当用户不能准确明白该特征标注标准时有帮助。

4. 注释框：在此可键入对于当前切分段的注释，以提醒自己某个问题或与同一项目下他人的交流。例如，可写“这是一个物质小句还是行为小句？查一下 SFG。”

总结，要标注整个文档：

1. 选择当前选择里的选项，直至穷尽。
2. 出现错误可双击已选特征栏的特征，撤销选择。
3. 关闭窗口，保存标注。

3 标注切分段文件

标注文档的层级为“切分段标注”时，过程较稍复杂。

首先，为方便用户，先增添新层级做研究。

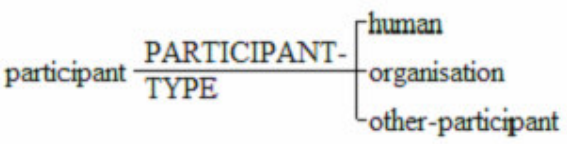
1. 调出工程窗口
2. 点击屏幕右侧“增添层级”按钮
3. 起名为“参与者”
4. 选择“标注切分段”
5. 选择“不自动切分”
6. 选择“创建新体系”
7. 点击“完成”按钮

注意这样在层级区增添了一个新层级，也在每个加入的文件旁增添了新按钮。

现在确定此层级的体系：

1. 点击“参与者”层级区的“编辑”按钮。
2. 当体系窗口出现后更改“参与者-1”为“人”，“参与者-2”为“组织”。
3. 点击“参与者类型”，选择“增添特征”选项。键入“其他参与者”。

网络应如下所示：



现在关闭此窗口，返回工程窗口。

点击其中一个文本文件旁的“参与者”按钮。

打开此文档在此层级的标注窗口，见图 4.2。

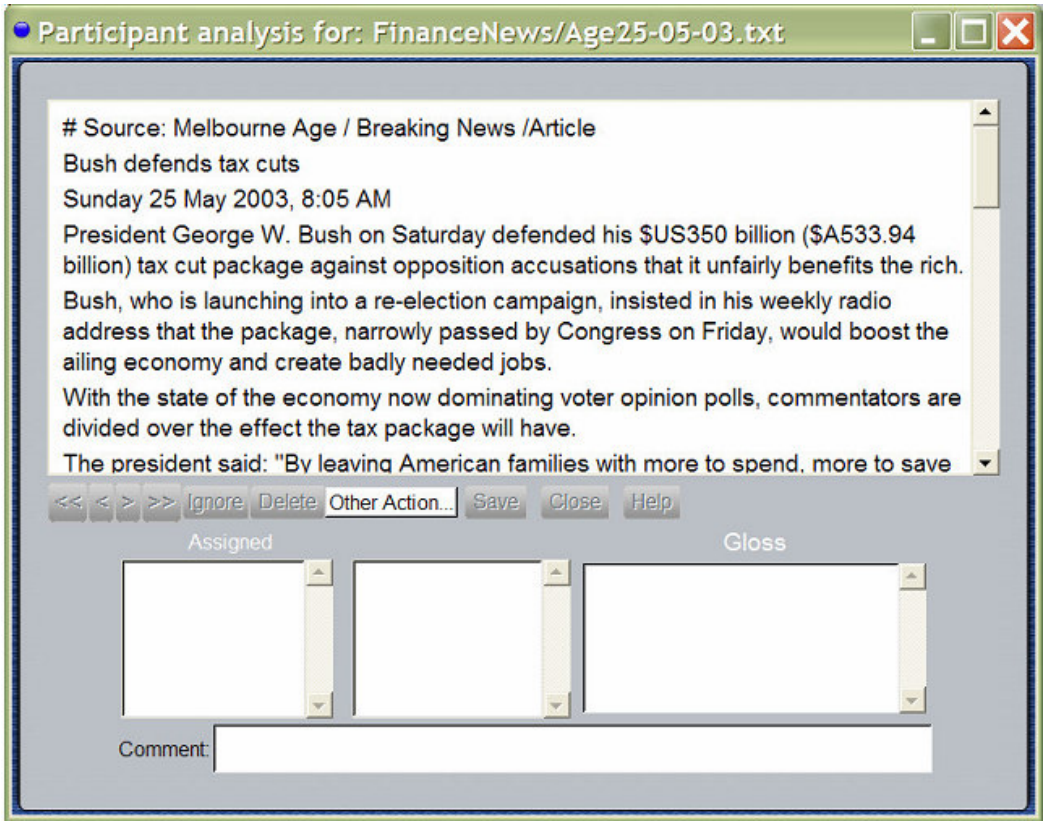


图 4.2 切分段标注窗口

与标注整个文档不同之处在中间的工具栏有更多的按钮。这些按钮主要用于在切分段间移动。

3.1 划定、移动和选择切分段

- 划定切分段（通过“横扫”文本）：在文本一处点住鼠标并拖动至切分段的终点后松开鼠标。
- 选择切分段：通过选择切分段下的线条来选择。选择切分段后该线条高亮显示。
- 选择后/前未完全标注的切分段：使用工具栏<<和>>按钮在未完全标注的切分段间前后选择。
- 调整切分段：将鼠标移至切分段一端边界点（垂直线），候其变红，按下鼠标拖动至位置。
- 删除切分段：错误创建切分段，可选定后点击工具栏上删除按钮，或键盘删除键来删除。

3.2 忽略切分段

选择切分段后点击“忽略”按钮，该切分段则不被统计分析。被忽略的切分段在文本窗口中呈灰色。该按钮也可恢复切分段。

4 标注图像文件

上述大部分讨论都是假定标注的是文本文件。换做图像标注，需另有讨论。

图像标注的窗口功能一般上和文本标注窗口一致。

创建图像的切分段，点击图像一点拖动至另一点。封闭的矩形是切分段。对其可在标注面板中赋值。

选择切分段，应点击切分段的外部线段（或利用<和>按钮）。

调整切分段，应点中切分段一角后拖至位置。

工具栏多出三个按钮：

- 放大：使图像变大
- 缩小：使图像变小
- 适合页面：调整图像精确符合当前图像空间范围

5 “其他按钮”菜单

此菜单显示附加选项，依据标注的文档（文本、图像）和类型（整体、切分段）而定。

编辑体系：打开标注层系的体系窗口，可编辑体系或增改特征的释义。

增添新特征：提示键入新特征名称，加入到当前显示的选择项中，并赋值给此切分段。

复制特征：复制已赋值给切分段的特征至内存。

文档重切分：抹去文档该层系所有切分。注意：这将删除文档该层系所有标注。

显示 XML：显示当前打开文件在磁盘存储的 XML 格式。

显示结构：切换至另一形式显示切分界面，更接近功能语言学的标准结构表达。见图 4.3。

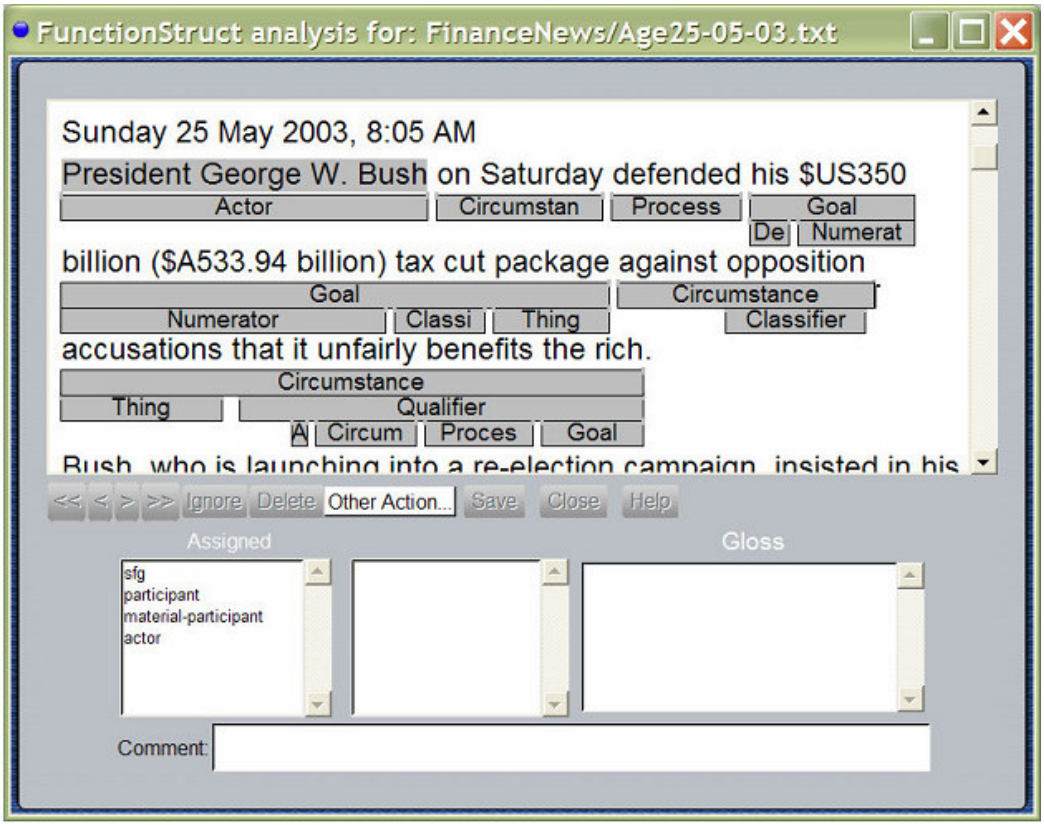


图 4.3 功能结构显示结构

显示文本流：打开新窗口，可浏览文本随选择的变化。用“系统到图像”菜单来选择浏览的系统。使用“平滑”菜单改变平滑度。平滑度为 0 时，每个标注选择都在序列中显示。较高的平滑度更好地浏览选择项在文本各阶段的分布。比如在图 4.4 中，文本流显示被动句在文首更多出现，之后减少。

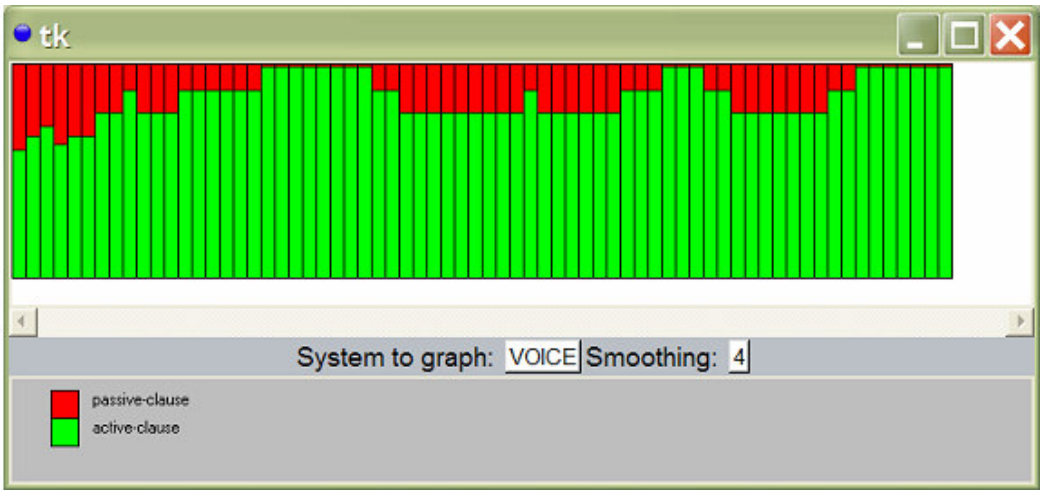

tk
System to graph: VOICE Smoothing: 4
passive-clause
active-clause

图 4.4 文本流窗口

第五节：语料库搜索

1 简介

点击项目窗口中的语料库搜索按钮打开搜索界面。窗口如图 5.1。

注意：要打开搜索窗口也可通过：

体系窗口：点击特征，选择“显示例子”。CorpusTool 将打开搜索窗口显示所有具有该特征的切分段。

描述或对比特征统计：点击任一集合的总数栏，将显示其示例。

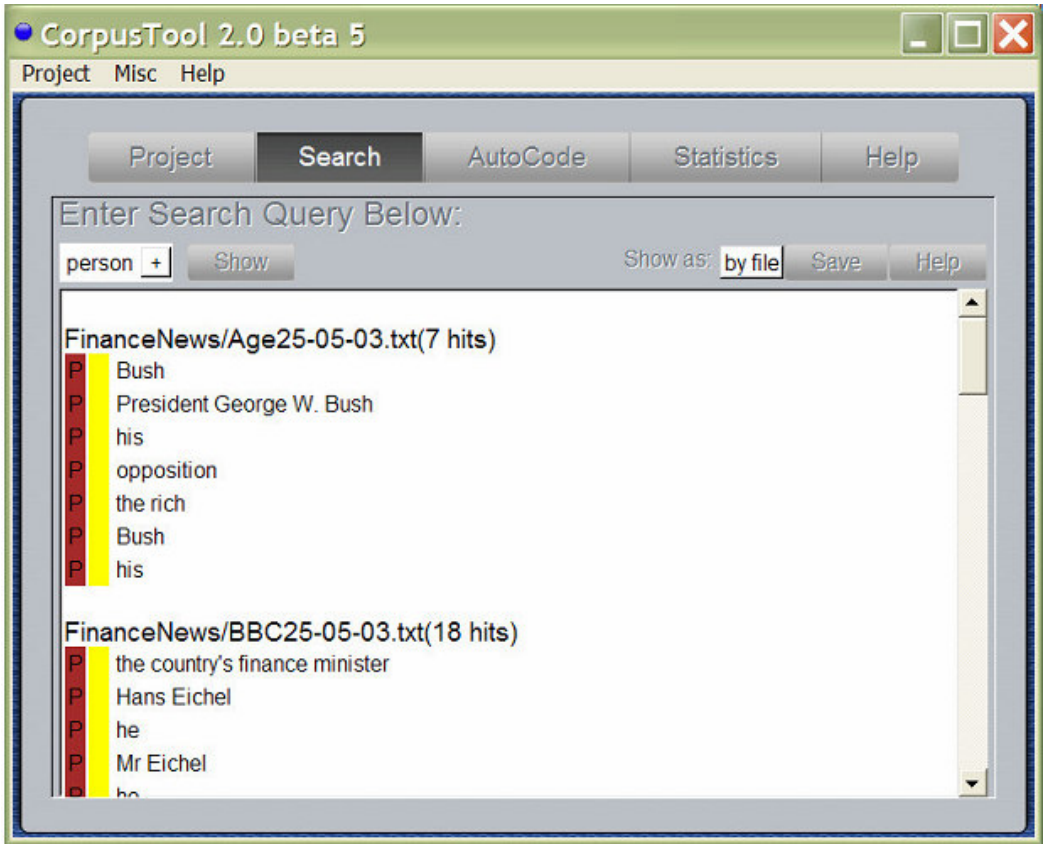


图 5.1 语料库搜索窗口

2 制定搜索式

窗口顶部有菜单式构件来制定搜索查询。

本帮助中使用一个小型项目“金融”，可从 CorpusTool 网站下载页面下载。

1. 简单特征搜索：搜索包含某特征的所有切分段，可点击构件左上角（图 5.1 中的“人称”），选择层级中的特征。然后按下“显示”得到所有示例。点击“保存”将搜索结果保存至文件。
2. 更复杂的搜索：点击特征选择器旁的小“+”号来扩展查询。
   - “与”：可增加另一特征，搜索返回同时包含指定的两个特征的所有切分段。
   - “或”：可增加另一特征，搜索返回包含至少一个指定的两个特征的所有切分段。
   - 注意：*与*和*或*不能混合使用！
   - “与非”： 可增加另一排除特征，搜索返回包含第一个特征但不包含第二个特征的所有切分段。
   - “包含切分段”：可跨层系搜索。返回标注有第一个特征的所有单元，同时又包含

在另一层级里的第二个特征。比如，搜索“限定性小句 包含 人称和主语”可找到所有限定性小句，并在其切分段内包含有人称和主语特征的切分段，此特征是属于另一个参与者层级。

- “包含字符串”：查询指定特征所有切分段，并包含某字符串。匹配不分大小写。
- 注意：此功能也可用于检索（基于词汇特征、通配符匹配等的搜索，详见下文）。
- “切分段内”：跨层级搜索，指定匹配项只包含于第二个层级的切分段内。比如，搜索“人称 在 评论”，可得到评论类别中标注为人称的切分段。

立即包含：注意：对于含有“包含切分段”、“包含字符串”或“切分段内”的搜索式，有“立即”和“任意”两种选择，区别为：

- 任意：若指定的切分段或字符串包含在所含切分段中，则会匹配。
- 立即：有时用户允许单元可嵌在同一层级的其他单元中，比如，小句可嵌在其他小句中。若指定“立即”，当搜索的切分段或字符串没落在嵌入单元时，便不会匹配所嵌入的单元。比如，在“[They left because [she was tired]]”中，搜索：clause containing immediately ‘was’只能匹配到内部的小句。

3. 复杂搜索组合：用户可以组合复杂搜索式，如，person containing immediately “bush” in finite-clause in editorial&english。

3 检索

CorpusTool 可搜索词汇类型（目前大多功能仅对英语）。

3.1 制定搜索式

如果制定“包含字符串”（见上），可不用简单的字符串，而用词汇类型。例如，为查询被动小句，“be% @participle”匹配含有“be”任一形式后跟分词（-en 动词）的所有切分段。

注意的是语料库并没有在词性（POS）上赋码。实际上 CorpusTool 含有一部大型英语字典，通过字典查询单词。因此单词会匹配到所有 POS 属性。比如“be%”匹配“being”的所有出现的形式，即使不是动词，如“the being”。

匹配规则如下：

不区分大小写：所有的搜索都不区分大小写。因此“Birch”匹配“Birch”、“birch”和“BIRCH”。

搜索字符串包括一系列的搜索字符，并被空格隔开。每个搜索字符可以是以下格式：

1) 纯字符：不含有*、#、@或%的字符仅匹配字符自身。
2) 通配符字符：如果查询字符含有*，它将匹配任意字符，因此
   - ca* 匹配 ‘cat’，‘carburettor’等
   - *ed 匹配 ‘weed’，‘lived’等
   - bro*en 匹配 ‘broken’，‘Brollerglen’等
3) 匹配任一：‘#’自身匹配任一个单独字符。

（凡是单词被空格或标点隔开的语言，上三例应都适用）

4) 词类限制：通配符后接@和一个单词特征，依据系统词库，匹配字符指定词类。如：
   - ca*@noun 匹配 以‘ca’开头的名词
   - *ing@mental-projecting 匹配 以‘ing’结尾的心理投射动词

   星号不能单独出现，只在文本前或后。

   附录二有完整的词汇特征列表，在 CorpusTool 的 Misc 菜单中选择“显示词类网络”也可见。

5) 普遍词类匹配：如果@前没有字符串，查询式匹配所有指定词类字符。如：
   - @noun 匹配任一名词形式
   - @verb 匹配任一动词形式

- @adverb 匹配任一形容词形式
- @mental-projecting 匹配任一心理动词形式
- @human-noun 匹配任一人类名词形式

6) 屈折匹配：‘%’在字符尾表示该字符的所有屈折形式，此字符应是根形式。因此：

- break% 匹配 ‘break’ ‘broken’ ‘broke’ ‘breaking’ ‘breaks’
- red% 匹配 ‘red’ ‘reds’（名词） ‘redder’（形容词） ‘reddest’
- be% 匹配 ‘be’ ‘is’ ‘are’ ‘was’ ‘were’ ‘been’ ‘being’
- is% 匹配 无（只能用根形式）

为限制屈折匹配数量可增加‘noun’、‘verb’、‘adjective’或‘pronoun’到%后面。如：

- red%noun 匹配 ‘red’ ‘reds’
- red%adjective 匹配 ‘red’ ‘redder’ ‘reddest’

注意，通配符不可以在%式中使用，%前也不可为空。

4 运行查询式

键入查询式后，点击“显示”按钮。如果光标在文字区（含有字符）可点击回车键。

5 修改查询式

更改特征选择，只需点击特征以更改。

删除搜索扩展，点击关键词（“&”, “/”, “containing”, “in”）后点击“移除”。

6 结果区域

查询区下面的白色空间显示结果。点击一个结果右侧出现含有该切分段的标注文件。

P/- 切分段是否完全标注（P=部分）

*/- 切分段是否有相关注释。点击切分段查看注释。

第六节：自动标注

1. 简介

自动标注窗口可以通过搜索式对现有的切分段赋予特征。比如，在英语中利用此式确认被动小句： ‘clause’ containing ‘be% @participle’

使用规则编辑器，制定如下规则：

（注意：同搜索，基于词汇的搜索式现仅限于英语）

之后点击“显示”按钮，即出现匹配查询式的所有示例，并各附有检查框。可以剔除错误匹配（非真正被动句）的任何一项。点击“选择标注”将“被动”特征赋予每个所选切分段。这种方法可以迅速地标注许多常见的语法形式。下面是一个用自动标注规则的例子，增添项目一个新层级，使用“clauses.xml”中的体系。包括有过程类型（心理、言语等）、语态（主动、被动）、情态、非限定性小句等。

1. 打开自动标注器：点击 CorpusTool 主窗口的自动标注按钮。
2. 增添新规则：若增添新规则，点击自动标注窗口顶部按钮中的“增添”按钮。出现如下窗口：

选择要自动标注的特征。制定使用的搜索式（见“语料库搜索”章节关于制定方法）。然后点击“保存”将规则存入内存。

3. 编辑规则：点击编辑按钮来编辑当前显示的规则。
4. 删除规则：点击删除按钮来删除当前显示的规则。
5. 规则标注：选定规则后点击“显示”按钮获得匹配搜索式成分的所有切分段。出现新的工具栏，带有三个构件：
6. 显示 所有/一致/冲突/不冲突：此选择列表允许过滤某些匹配：
   所有：显示所有匹配
   一致：显示已被该指定特征标注的所有切分段
   冲突：显示那些已被某特征标注，又与自动标注特征相冲突的切分段。比如，自动标注的是“被动”，这将显示已被标注为“主动”的所有切分段
   非冲突：显示既非一致也非冲突的所有切分段
7. 选择 所有/取消：选择所有/取消检查栏每个切分段。
8. 标注所选：点击此按钮自动标注已选择的所有显示的切分段。

提示：

对于某些语法现象，可有如下一些规则式：

Select passive if contains 'be% @participle'

Select active if clauses and not passive

使用第一个规则来标注被动句，然后用第二个规则将所有其他句子标为主动句。

如上被动句规则式提供一个规则，加以标注，编辑此规则，在搜索项间插入#，如：

Select passive if contains 'be% # @participle'

这将找到在动词中间有 not 或形容词的示例。

第七节：语料库统计

1 简介

语料库统计面板允许来自标注语料库的各种统计。点击主窗口工具栏的“统计”标签激活统计面板（如图 7.1）。

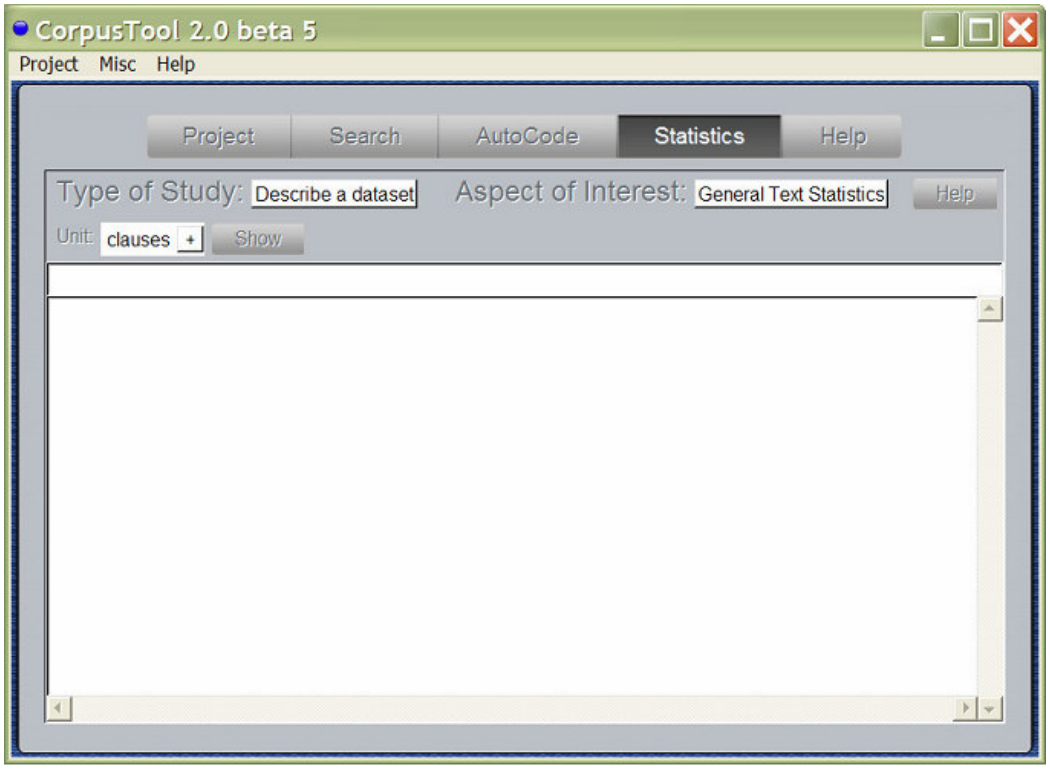


图 7.1 统计面板

此界面提供语料库的两类研究：

1. 一般文本统计：提供语料库的一般统计信息，如切分段总数，切分段平均词数，语料库的词汇密度，代名词使用等。
2. 特征使用：指定一个层级的特征（最典型的是层级的根特征），然后程序会描述语料库中该特征在层级中的使用（总数、平均值和标准差）。

这些研究可以是针对单一数据集（描述性统计），两个数据集（对比性统计），或者逐个显示每一个文档的结果。

1. 数据集描述：提供语料库或指定的子语料库的描述。
2. 两个数据集比较：提供语料库的两个子集（如英语 vs 西班牙语）的比较。特征选定后，两个集合在指定层级上所出现的标注特征示例进行比较。两集合间差异显著性水平得到显示，包括 Student T-test 和 Chi-Squared（见下）。
3. 多文件比较：提供语料库每个文件的细致信息，每个文件一栏。

2 一个对比特征研究

图 7.2 显示了利用“财经”项目的对比研究样例。注意现已标注的文本并不多，所以结果只是对应于小数目的。其结果若要获得相信需要标注一千或更多评论性或 fpn（头版新闻）文章。此项前期研究显示了两组有显著性的结果（人名比组织名更有显著性，达 98%；组织名类型的显著性差异已讨论），但数字仍太低，不能信赖。

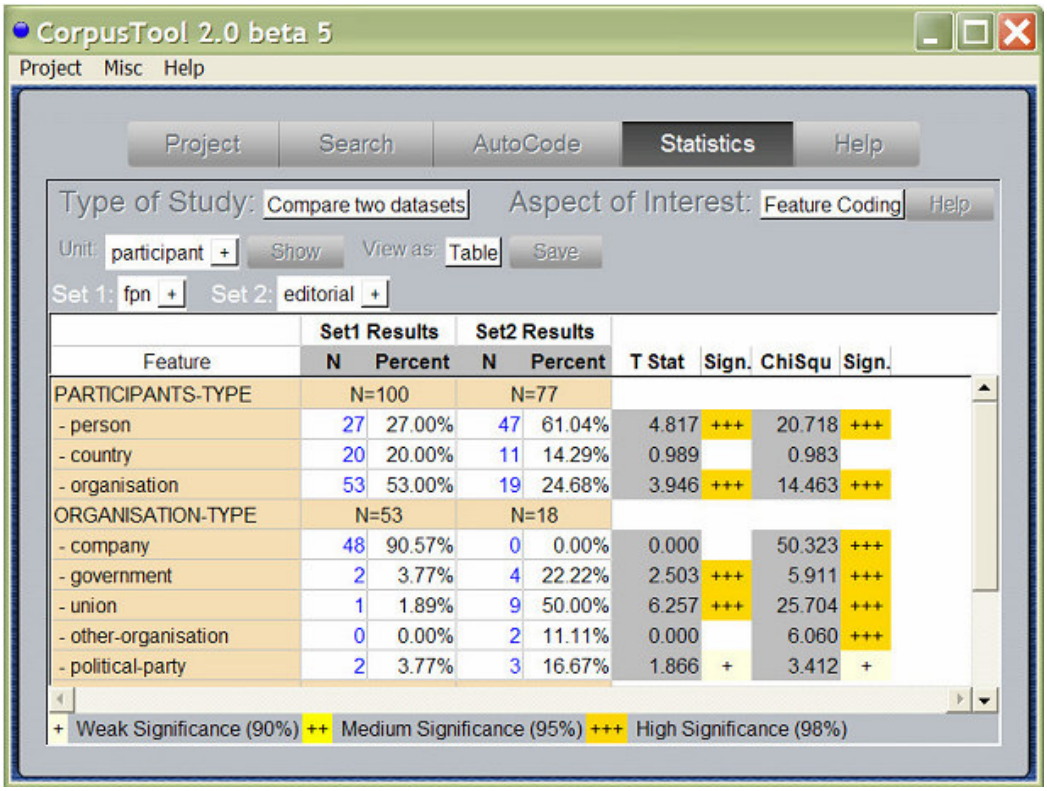


图 7.2 对比统计研究

3 进行研究

进行上面描述的研究需：

1. 从“研究类型”菜单选择选项：“数据集描述”，“两数据集比较”或“多文件比较”。
2. 从“兴趣角度”菜单选择：选择“特征标注”或“一般文本统计”。
3. 指定感兴趣的单元（见第五节第二部分：指定搜索查询式）：应为希望探寻差异的单元。可以是网络的根特征（如 7.2 的情形），或更具体的单元）。
4. 若已选择“两数据集比较”，键入第一个集中的特征和第二个集中的特征）。此单元应包含有研究兴趣的单元。在此例中，指定的是语域层级单元：fpn 和评论。因为这两个特征应用到整个文本，所以一定包含具有“参与者”的切分段。
5. 点击“显示”。

4 结果解释：基于特征的研究

显示的只有相关的系统。比如，指定的兴趣单元为上边的“人名”，研究则仅限于有“人名”特征的切分段。因此整个系统结果不会显示，因为“人名”占到 100%，系统其他特征为 0%。总数和百分比：每个特征的结果显示有绝对总数（在数据集中出现次数）和百分比数。百分比显示有该特征的切分段比例。注意在一个系统中的各百分比(给定的选择)相加总是 100%，所以实际测量的是在同一系统中选择此特征而不是其他特征的偏好。

统计显著性：对比研究完成后，可能需要测量两个数据集的差异是否在统计上具有显著性(是真实具有差异还是数据的随机性造成的)。

CorpusTool 使用两种统计显著性的方法，在结果中同时展现：

- T-Statistic：T-Stats 是来源于结果显著度的数字。数字越大显著度越高，但这也取决于数据量大小。在更严谨的论文中可能需要提供 T-stats，但在语言学中则很少。

- Chi Squared：近年来，特别是在语言学界，chi squared 统计在检测显著性时更受欢迎。CorpusTool 为每次比较都提供 Chi Squared 统计和相关统计度。

每个条目的最后都有 0 到 3 的“+”标记。显示的是本特征平均值和其他集的统计差异：

（none）没有显著性差异

+ 90%的显著性（10%的误差）

++ 95%的显著性（5%的误差）

+++ 98%的显著性（2%的误差）

显著度的重要性在于能够确认结果的重复性。没有显著性的结果可能是碰巧，如果拿其他文本重新研究，结果可能就不一样。如果结果有高度的显著性，分析应用到完全不同的文本中则结果很有可能重复。可以这样理解，一个+代表在十个结果中有一个是错误的（90%显著性，或 10%误差）。

5 结果展示为网络

进行基于特征的研究时，结果可以系统网络形式观看，不止是表格形式。见图 7.3。当结果以表格展现后，出现一个新菜单，标签为“View As”。选择“网络”转换到网络视图。

此种展示统计数据的方式来自 SysFan[1]的类似特点。感谢其作者 Wu Canzhong 允许我使用这个形式。

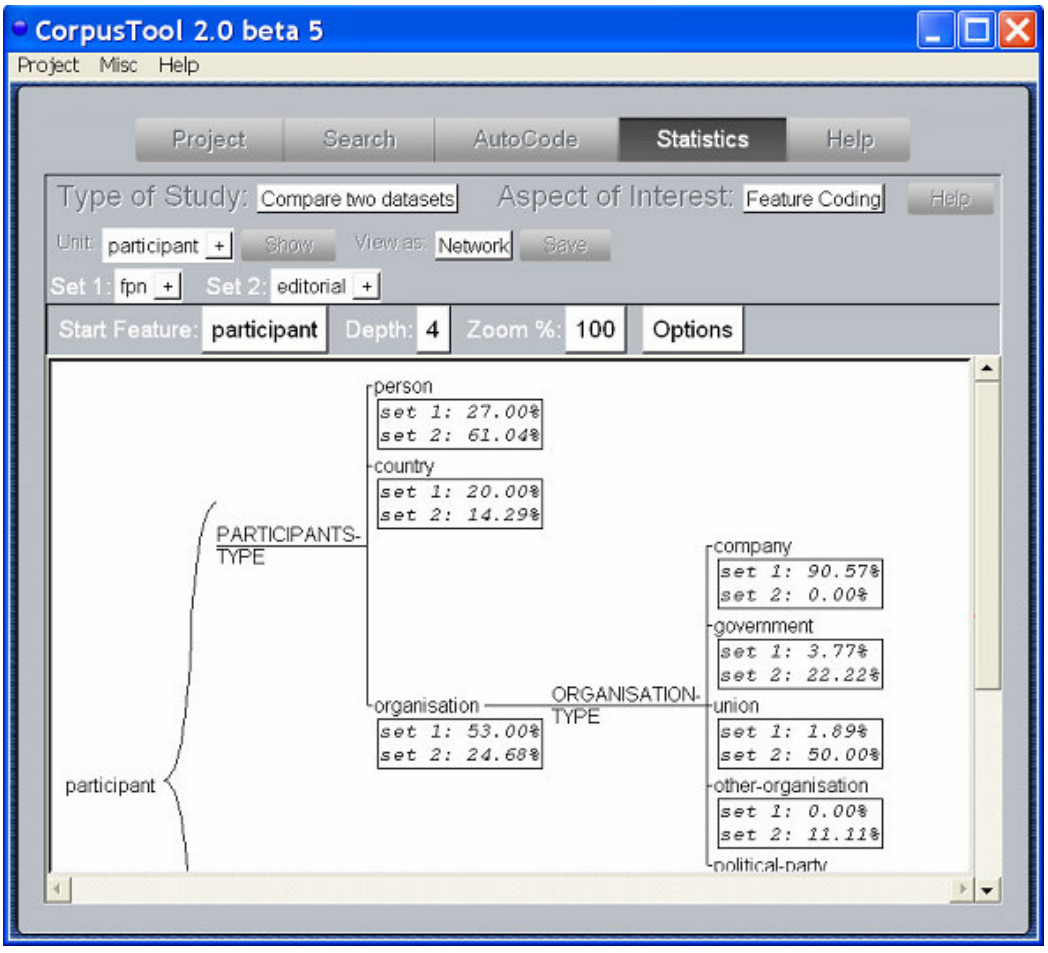


图 7.3 统计的系统网络浏览

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

6 保存统计结果

每个统计窗口提供“保存”按钮将结果保存到文件，可以是 HTML 格式、制表符边界格式或纯文本格式。

以 HTML 格式保存的结果可在 MS Word 中打开，再剪切/粘贴到文章中。

以制表符边界格式保存的结果可在 MS Excel（在 Windows 中，右击.txt 文件，选择打开方式…Excel）。文件在 SPSS 软件中同样有用。

第八节：文本样式

1 文本样式

有时视觉呈现文本标注很有帮助。CorpusTool 可以查看项目某文本文件时，指定（在任一层级中的）特定切分段显示为粗体、斜体、下划线、大字号或加色。见图 8.1 显示“财经”项目中的一个文件的文本样式。

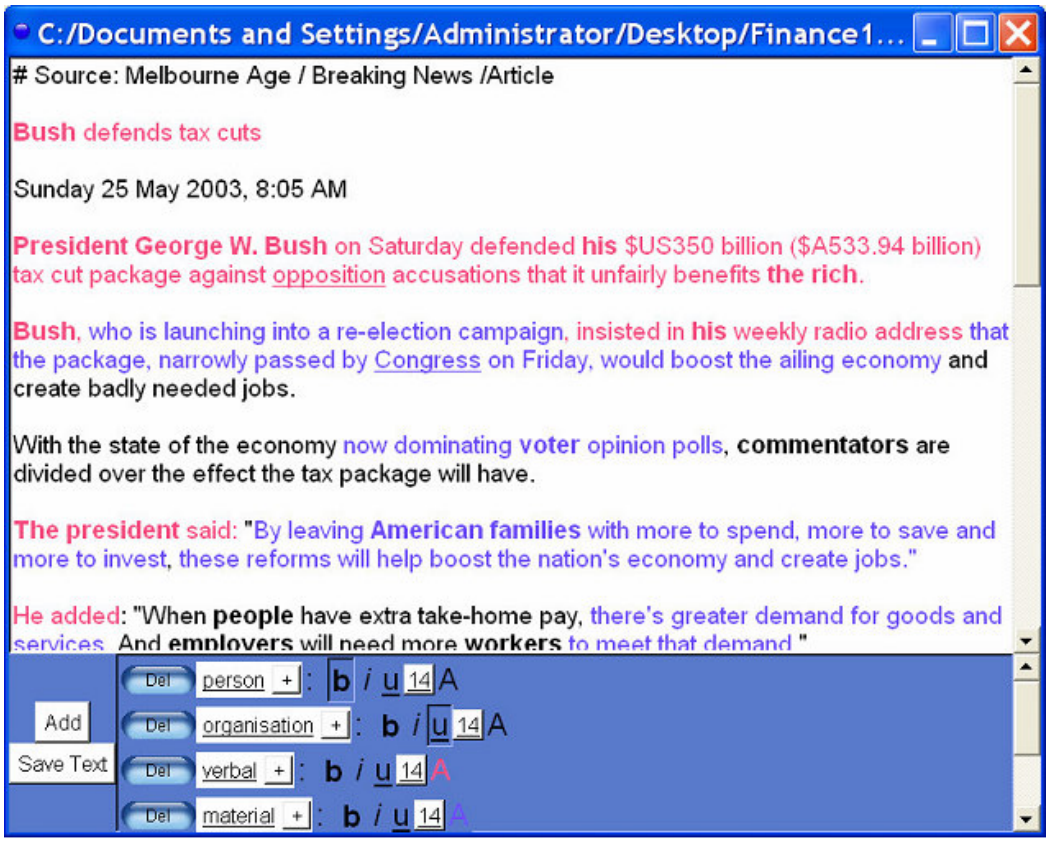


图 8.1 文本样式器窗口

2 打开文本样式器

从项目窗口（主窗口）点击一个文件名。注意，这仅对于已加入到项目的文件。同时项目也需至少一个已定义的层级。

3 编辑文本样式

文本中标注的给定特征或特征组合都可以加色，和/或增添字体效果（粗体、斜体、下划线）。这样选择的类型在文本中都可见。

例如，对评价范畴使用粗体/斜体/下划线，对小句类型加色，来观看评价在小句中的分布。

4 保存样式文本

可将样式文本存为 HTML 文件。要将其包含到 MS Word 文档中，需在 MS Word 中打开 HTML 文件，通过剪切/粘贴操作实现。

附录一：导入 Systemic Coder 的研究结果

1 如何导入 Coder 的研究结果

Systemic Coder 的分析文件可导入 CorpusTool。其步骤如下指示：

如果导入一个文件：

1. 确认标注体系是保存为一个单独的文件（主体系）。在 Coder 中打开文件，于选项菜单中选择“保存体系…”。选择“保存至主体系”指定保存路径。
2. 确认标注文件保存为.cd3 而不是.cd2 格式：如果文件扩展名是.cd2，需打开该文件，在文件菜单中选择“保存标注文件为…”。程序会提供.cd3 格式文件保存。
3. 现在建立一个新文件夹，将体系文件和标注文件放进去。
4. 打开 CorpusTool，建立新项目。
5. 从项目菜单选择“导入层级”。
6. 然后指定在上面步骤 3 中创建的文件夹。
7. .cd3 文件被分解为纯文本（放入语料库文件夹）和分析文件（放入分析文件夹）。下一窗口询问将文本文件放入哪个子语料库文件夹。
8. 分析体系导入为一个新的层级。下一窗口询问层级名。
9. 在 Coder 中不标注部分文本的方法只能是忽略它。在 CorpusTool 中用户可只选择需要标注的文本部分。如需要将 Coder 中忽略的切分段消失，下一窗口可以实现。
10. 点击完成，即增添新层级，cd3 文件也已导入。

如果有一系列的文件，都是同一体系标注：

1. 将所有 Coder 文件置于一个文件夹。
2. 确认所有文件都是.cd3 格式，而不是.cd2 格式。
3. 按照单独文件导入的步骤 1，至少操作一个文件（例如，确认文件夹中存折一个体系文件）。
4. 继续其步骤 4。

如果有一个或更多文件，同一文本有不同的网络标注（如同用 Coder 多层级标注）：

1. 对于同一层级标注的系列文件，建立一个文件夹，将 coder 文件和体系分析文件放入（确保文件是.cd3 格式）。
2. 确认对同一文件的分析文件都有同样的文件名，如 Text1-CLAUSE.cd3 分析小句，Text1-GROUP.cd3 分析小组，将两个文件重命名为 Text1.cd3。（CorpusTool 只能通过两个有同样文件名的文件来确认分析文件）
3. 打开一个新项目，按以上描述操作导入层级选项，完成文件夹中的一个文件。
4. 对其他文件夹，重复步骤 3。

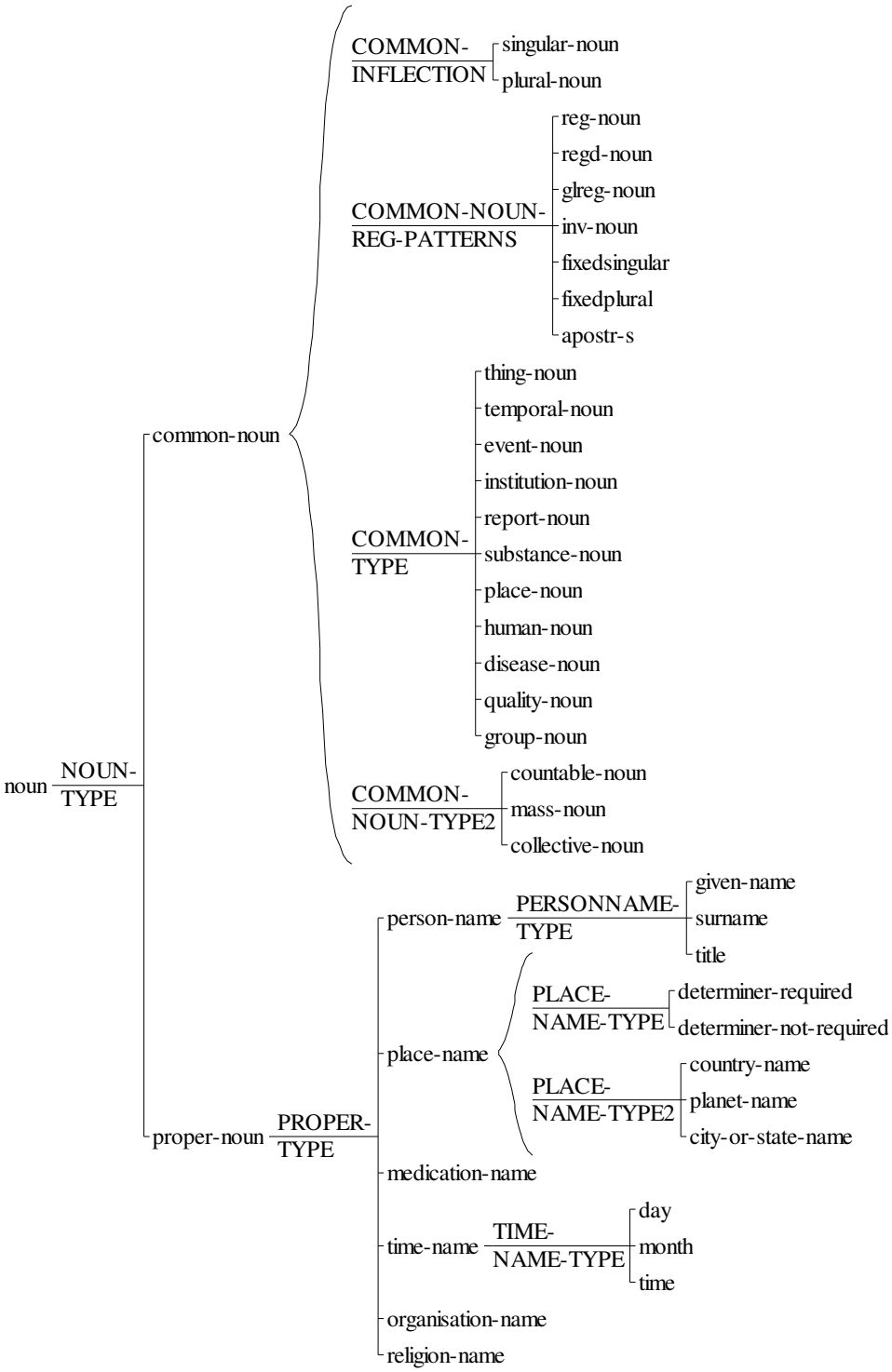
出现的可能部分问题有：

CorpusTool 提示不能读入一个或多个.cd3 文件：文件可能含有 ASCII 之外的字符。CorpusTool 应当能处理，但现在尚不能。发送文件给我替您转换。

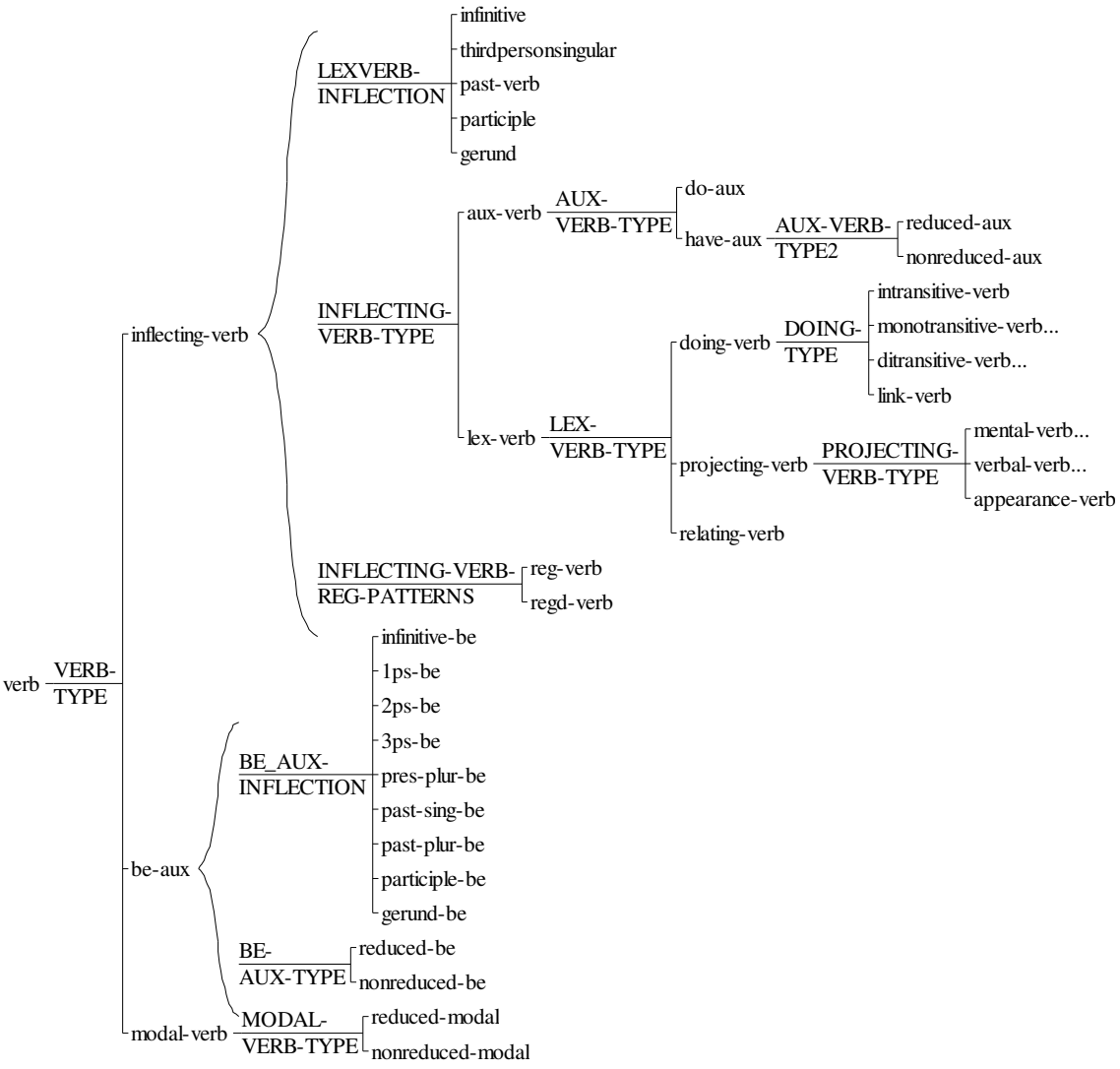
在导入.cd3 文件中的其他问题，将其（压缩文件夹为 zip）发送给我，我来诊断（这样本人可以处理各类问题，进行修理）。

Appendix II:
Lexical Features for Concordance Searching

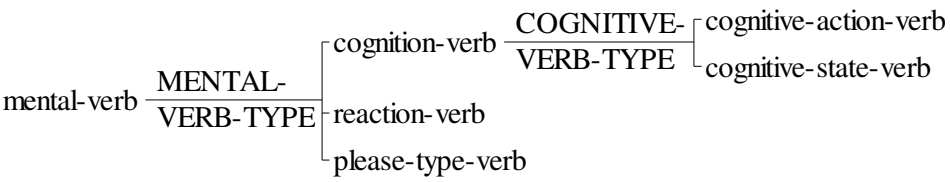
Nouns



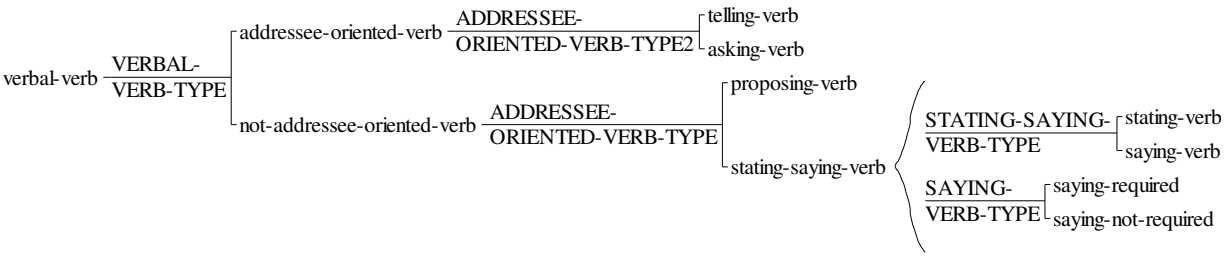
Verbs



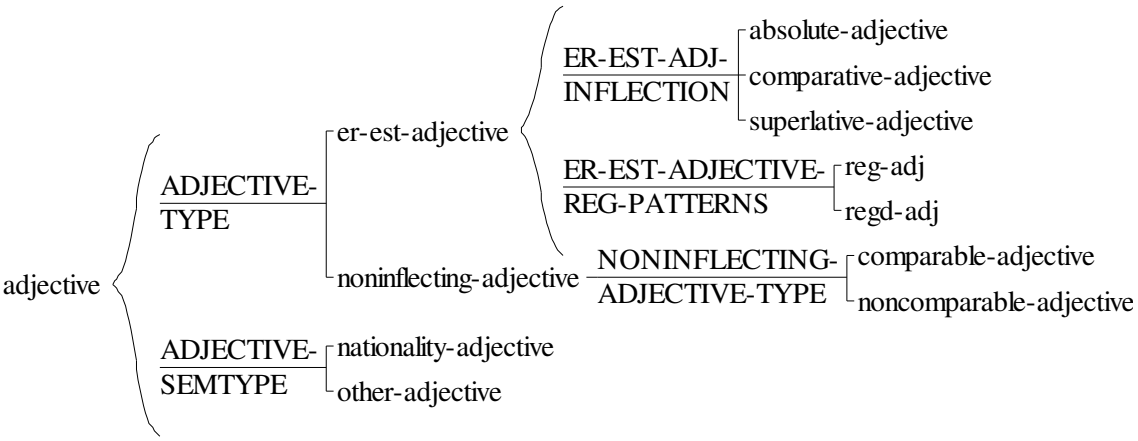
Subclasses of mental verb



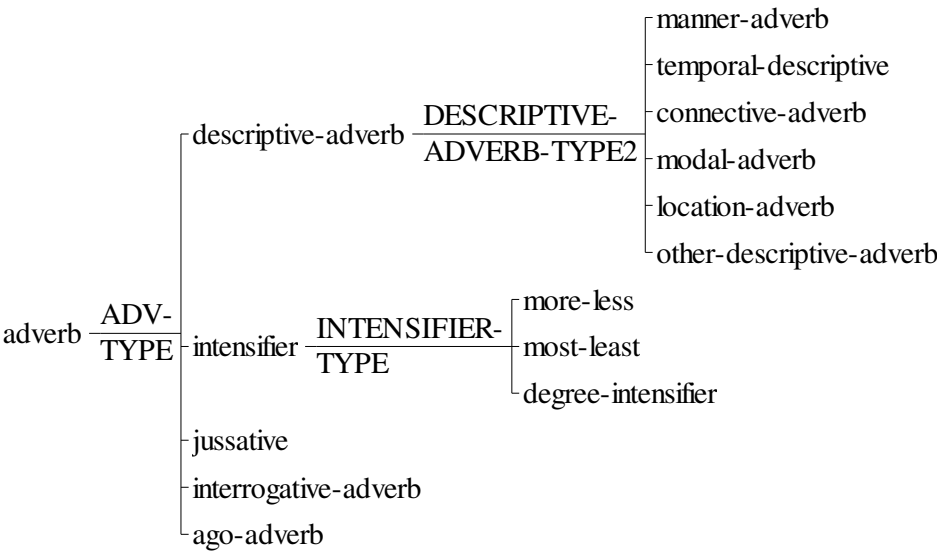
Subclasses of verbal verb



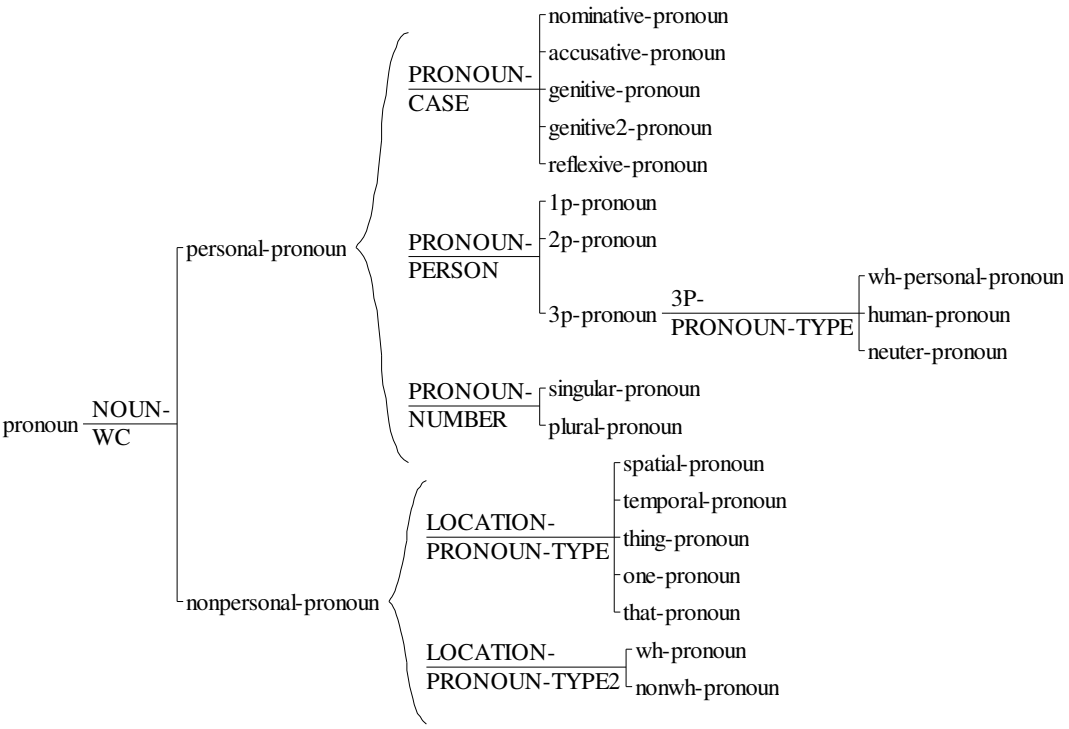
Adjectives



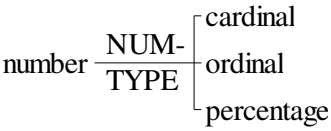
Adverbs



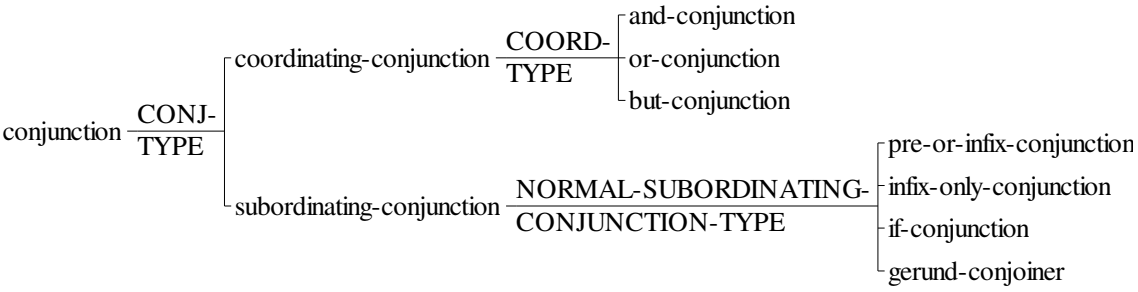
Pronouns



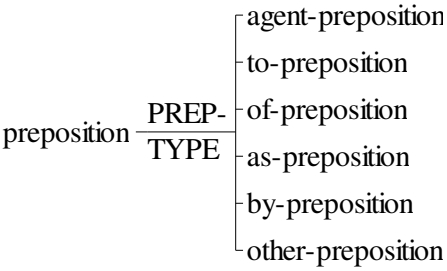
Number



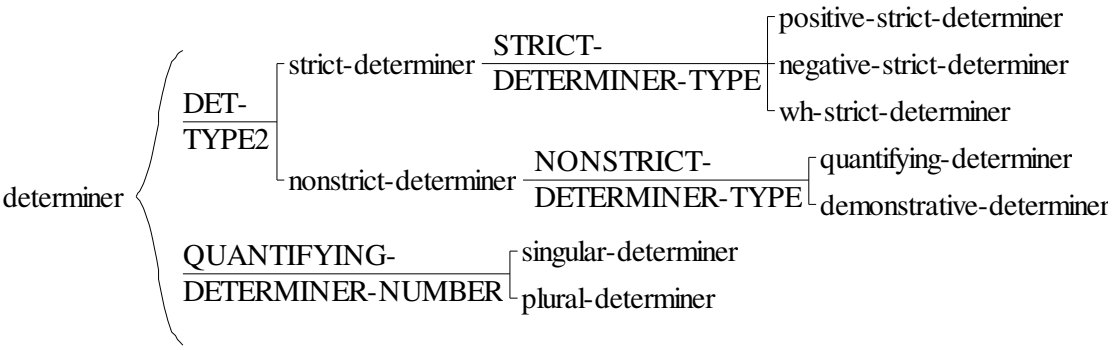
Conjunction



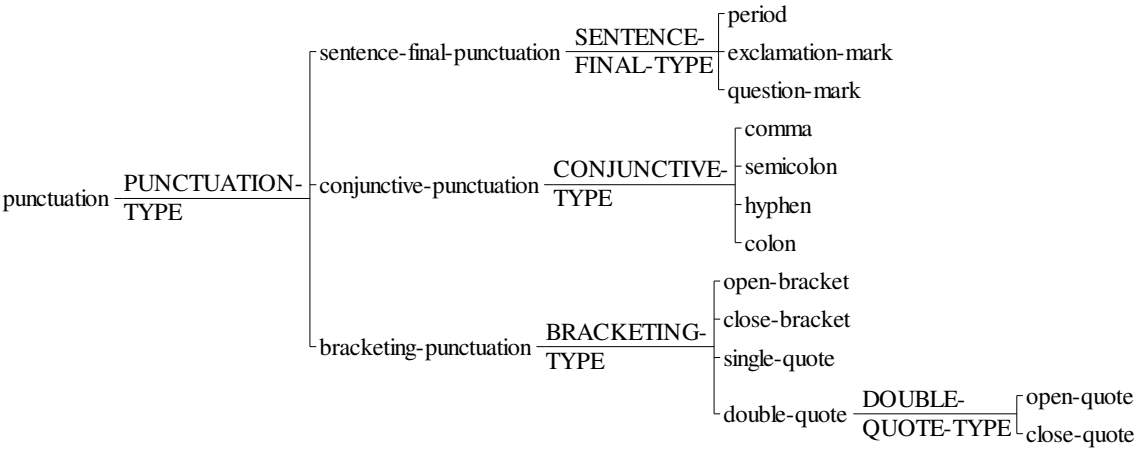
Prepositions



Determiners



Punctuation



genitive-s

附录二：检索使用的词汇特征

1 名词

2