法律语料库建设设想

◇杨海燕(上海电力学院)

关键词:语料库;法律语料库;建设;标注

摘 要:本文主要从法律语料库的定义、设计、语料的采集筛选和标注、检索软件的选择等方面探讨了法律语料库建设设想,指出法律语料库的建设是必要的,也是可行的。法律语料库的建设和应用为法律语言学研究的新领域。

Compilation of Forensic Corpus

YANG Haiyan

Key words: corpus; forensic corpus; compilation; markup

Abstract: This paper discusses forensic corpus compilation in terms of forensic corpus definition, design, text capture, text markup and software. It points out that it is necessary and feasible to compile a forensic corpus. The compilation and using of forensic corpus is becoming a new research field in forensic linguistics.

一、前 言

“从科学方法的角度,语料研究方法是一种更为强有力的方法,因为其结果是可以验证的”(Leech,1993)“。”基于语料库的研究不再是计算机专家的独有领域,它正在对语言研究的许多领域产生越来越大的影响”(Tomas&Short,1996)。语料库建立使人们对语言的研究从以直觉为基础转变为以语料为基础,带来了语言研究领域的一场革命。20 世纪 90 年代以来语料库更是迅猛发展,其研究趋势主要表现在以下三个方面:(1)大规模、多品种语料库的建设;(2)对语料库的深加工研究;(3)语料库广泛应用于与语言相关的各个领域。建立大型语料库已经不再是很困难的事,接下来的是如何建立各种有地域、语体、乃至语篇特色的大型语料库以及对现有的语料库进行深层次的开发与研究(何安平,2001)。国际法律语言学家协会(IAFL)搜集并划分的 20 个法律语言学研究领域中的第 15 个科技项目就是关于计算机软件和资料库方面的研究。因此,有必要也有可能建设法律语料库。法律语料库的建设和应用将给法律语言学研究提供强有力的研究工具,同时也给法律语言学研究注入新的活力,成为法律语言学研究的一个新领域。

二、语料库与法律语料库

1. 法律语料库的定义

语料库是“按照明确的语言学标准选择并排序的语言运用材料汇集,旨在用作语言的样本”(Sinclair,1986)。Renouf 认为,语料库是‘由大量收集的书面语或口头语构成,并通过计算机储存和处理,用于语言学研究的文本库’(Renouf,1987)。在现代语言学当中,语料库指为了某一研究目的收集的,用计算机可读形式保存的语言文本集合;这一语言文本集合由大量的自然出现的口语或书面语材料汇集而成,代表特定的语言或语言变体。语料库通常是指科学规范地标注了不同语言信息的语料库。Walker(1990)区分了四种不同的语料库,即异质的、同质的、系统的和专用的。这种分类方法的标准不清晰,是有争议的。现在已建立的和正在建立的语料库可作如下分类:(1)按照语料所代表的媒体形式可分为书面文字材料的电脑文本语料库、经过转写的口语语料库、视频语料库以及上述几种形式的混合语料库。(2)按照语料库设计结构可分为均衡结构语料库、无结构的随机开放式语料库、以及有若干子库叠加而成语料库网。(3)按照语料的来源有单语种语料库和多语种语料库之分,原文语料库和翻译语料库之分,母语语

料库与外语学习者语料库之分。(4) 按照语料的实效可分为共时语料库和历时语料库。(5) 按照语料的处理方式可分为未经附码的语料库与经过附码的文本语料库(潘永樑, 2000) 。

杜金榜(2004) 指出法律语料库是用于法律活动和法律语言学研究的语言信息库。法律语料库是为了法律语言学研究和法律实践收集的, 由书面语和口头语构成, 经过计算机处理和存储的, 大量真实的法律文本库集合。法律语料库的建设有着非常明确的目的, 用于描述法律语言和法律实践活动。法律语料库应该定位为专门用途语料库, 主要收集立法、司法、公安检察和普法等相关法律环节的文本, 形成一个由若干个子库构成的法律语料库网。法律语料库是既包括书面语又包括口语。法律语料库是一个动态的、开放的语料库。

2. 法律语料库的建库目的和用途

法律语料库的建设就是为了能够通过大量真实的法律文本揭示法律语言的本质特点, 为语言学的深入研究提供进一步的证据; 终极目的是为了能够把法律语料库提供的信息材料应用到具体的法律实践当中, 为我国立法、司法和执法实践提供有价值的意见和建议。

从法律语料库的定义可以看出法律语料库有两个用途: 一是用于法律语言研究, 查询法律语言各层面的语言特征, 验证有关法律语言的某些假设等; 二是用于法律实践活动, 为立法、司法和执法等法律环节提供证据支持, 解决具体的法律问题。法律语料库要实现法律语料的资源共享, 可进行法律语料库的在线索引、查询等功能。

更详细地说, 首先, 可以实现用语料库研究法律语言。以立法语篇特点分析为例,“法律语篇一般性特点的确定, 往往基于合理的语篇分类和足够的语篇样本”(杜金榜, 2004) , 而法律语料库恰好可以提供大量真实的语篇样本。基于法律语料库的立法语篇特点的确定是建立在大量的真实立法语篇的基础上, 具有足够的代表性。法律语料库可以运用到法律词典编纂、法律术语研究、法律语篇分析、法律文本资料翻译、法律语言教材编写和法律语言教学等法律语言学研究领域。其次, 法律语料库中大量的法律语料资源也可以运用到具体的法律实践活动当中, 如文本鉴定。法律语料库建设全部完成后, 其功能是强大的, 可以为法律语言研究者、法律工作者乃至法律语言学习者提供强有力的工具和资源。

3. 法律语料库的设计原则及结构

在建设语料库时需要考虑以下几个主要问题:(1) 语料库是静态的还是动态的;(2) 语料库的代表性和平衡性;(3) 语料库的规模(Kennedy, 1998) 等。结合法律语言的特点, 法律语料库的建设要体现以下原则:(1) 开放性原则。法律语料库本身是一个开放的系统, 而不是封闭的系统。法律信息库的资源经过语料库建设者和管理者的筛选、标注可以进入到法律语料库, 同时和其它的信息库和语料库也可以实现对接。(2) 动态性原则。法律活动的丰富多彩和千变万化决定了法律语料库是一个动态的语料库, 语料库中的语料可以随时添加或删除, 使法律语料库可以更及时准确的代表法律语言。因此, 我们设计时不规定法律语料库的具体容量规模。(3) 真实性原则。法律语料库收集的文本必须是自然真实的文本。如法庭话语, 警察讯问等文本应该是现场录音或录像材料经过规范的转写获得的。只有这样的语言才是最真实, 最自然的体现了法庭话语和警察讯问等语言的特征。(3) 语言从法原则。法律语料库中法律汉语、法律英语语料库分别按照法律活动的过程划分为立法、司法、公安检查和普法四个二级子库。

法律语料库的结构框架如图 1 所示。

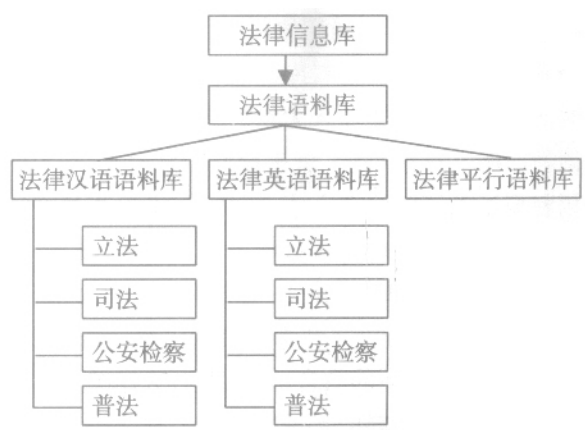

图 1　法律语料库结构框架

从图 1 可以看出: 法律语料库有一个强大的支撑: 法律信息库。法律信息库包含书面语、口语(转写) 、音频和录像四个子信息库, 可以为法律语料库提供充足的语料资源, 而且信息库里采集来的语料资源是经过初步筛选的, 经过进一步的筛选和标注可以非常方便的进入法律语料库。法律

语料库包含法律汉语语料库、法律英语语料库和法律平行语料库三个一级子库。法律汉语和法律英语语料库又分别包含立法、司法、公安检察和普法四个二级子库。法律平行语料库由于其自身的特点需要另行进一步考虑。法律语料库中的语料都是经过深加工的,是标注过了的语料。

当然,同时开发法律汉语语料库、法律英语语料库和法律平行语料的任务是艰巨的,因此在具体的建设过程中考虑分步进行。先进行法律英语语料库的建设,首先可以开发一个小型的实验库,建立一个工作模型,经过充分的论证和实验后,可进行大规模的法律英语语料库的建设,然后根据法律英语语料库建设的经验再着手进行法律汉语语料库和法律平行语料库的开发。

三、法律语料库的建设

1. 语料的采集和筛选

法律语料库中语料的采集并不是语言材料的简单堆砌,而应该在采样的时候考虑到法律语料库的代表性和平衡性的问题。一个有实际应用价值语料库决不是任意文本任意集合,而是应该"尽可能的代表一个语言的全貌或一个特殊的语域(Biber,1993)"。从理论上讲,语料库越大,语料库的代表性也就越强,但是这是一种理想状态。从某种意义上讲,法律语料库对法律语言的代表性要好于像 BNC 之类的通用语料库。法律语料库是一个开放的语料库,语料库建设和管理人员可以随时向法律语料库中加入大量的语料样本,以弥补语料库的不平衡问题。就平衡性问题而言,采集的语料要尽可能的涵盖立法、司法、公安检查和普法各个领域,同时这些领域的语料要占有一定的比例。举例来说,如果采集来的语料绝大部分都是法律条文、规定、司法文书等之类的语料,而缺乏法庭话语、警察讯问的录音语料,那么这样的法律语料库由于采集语料的不平衡,也就不能代表法律语言的全貌。当然,由于某些客观的原因,采集法庭话语、警察讯问录音是一项比较艰辛的工作,需要付出长期不懈的努力。

法律语料的采集途径有以下几种:可靠的网站,电子光盘、扫描仪扫描、手工输入。可以直接通过互连网搜索和现有的电子光盘法律数据库中获取一部分有价值的语料,这是一种比较方便快捷的方法;对于那些非常有用但又无法找到现成的电子文本的语料,可以选择扫描的方法;人工输入虽然费时费力,但对于一些比较珍贵的笔录如手工写的警察讯问笔录等只有采取手工输入的方法。对于音频和录像只能采取人工转写然后手工输入电脑的方法。

最后对于采集来的语料要按照立法、司法、公安检察和普法分类存储于法律信息库。由于立法的语料比较多可以按照部门法再进行进一步分类。对于录音和录像的原始材料经过格式转换后另外存储于法律信息库当中,以备以后对语料的进一步深加工。

2. 语料的标注

未经任何处理的电子文本语料库被称为生语料库,其中没有包括词法及语法等信息,应用价值非常有限,要实现语料库的多种价值,必须对语料库库进行多层次的标注。Leech(1993)年提出了语料标注应该遵循的七个基本原则:(1)标注附码可以删除,可以恢复到原始语料。(2)所做出的标注可以单独抽出,另外储存。(3)语料的最终使用者应该清楚标注的原则和附码的意义。(4)在语料库的使用和说明文件中,应该说明标注者以及标注所使用的方法。(5)应向用户表明,语料的标注并非完美无缺,他只是一种可能有用的工具。(6)标注应该尽量采用被人们普遍接受的中立的模式。(7)任何标注模式都不能作为第一标准。即使有,也只能通过大量的实践和比较才能得到。在对法律语料库进行语料标注时也要遵循这七大原则,寻求一种既方便语料库使用者,又对标注者来说是切实可行的一种标注方案。

现代语料库大都采用了 XML 标注技术。XML 是一种开放性的"可扩展的标记语言",允许用户设计自己的标签和文档结构,比较适合语料库标注。但是 XML 的句法要求很严格包括标签大小写一致;标签不能乱套;属性一般使用双引号等。法律语料库的标注方案和规范应该在参照其他语料库标注方案的基础上,根据法律语料库自身的特点,制订出一套基于 XML 语言的标注方案。下面的例子就是用了 XML 标记语言。例如:

<p id="1">

<s> All legislative powers herein granted shall be vested in a Congress of the United States, which

shall consist of a Senate and House of Representatives. </s>

法律语料库的结构标注格式参照 COLSEC 的结构标注格式有三部分:句子以<s>开始,以</s>结束;段落以 开始,以 结束;文档在一行 XML 声明下以<html>开始,以</html>结束。中间内嵌两对标记,表示文档公共信息的<head>和</head>,表示文档正文信息的<body>和</body>头部信息包括了一些检索时需要的信息以及其他一些公共信息;正文就是语料库的主要部分(李文中,2006)。

对文献信息标注,应该标注出语料类型<type>和</type>、语料类别<category>和</category>、语料名称<title>和</title>、语料作者<author>和</author>、语料长度<length>和</length>、语料发布时间<publishing time>和</ publishing time >、语料采集人<collector>和</ collector >、语料采集地点<collecting place>和</ collecting place >、语料采集时间<collecting time>和</ collecting time >和语料采集途径<collecting means>和</ collecting means >等 10 类背景信息,放在文档公共信息中。

对于语法标注,应该针对法律英语、法律汉语和法律平行三个语料库分层次进行。词类附码可采用计算机加人工的方法,可用现有的软件如 CLAWS、ICTCLAS_Win 等词性附码器对语料库进行 POS 附码,然后辅以人工校对。法律语料库再进一步的深加工就是对法律语料库进行语言信息码的标注,这一步的标注可采用人工加计算机辅助的方法。对语料库进行语言信息码的标注有助于揭示英汉语篇的信息特征。

3. 检索软件的选择与开发

目前,语料库检索软件很多,较成功的共享软件如 Wordsmith Tool v4, Concordance v3, Monoconcord 以及 WordCrucher 等,自由软件如 Microconcord, Tact 2.1(基于 DOS 平台),Wconcord, Concap, Mlct - concordancer, Wordpilot, Antconc, Paraconc 等(李文中,2002)。在法律语料库建设的初级阶段我们不考虑开发自己的语料库检索软件,而是根据需要使用现有的语料库检索软件。但是这些现有的检索软件并不是万能的,并不一定完全适用于法律语言研究。当法律语料库建设到高级阶段,我们就应该考虑开发专门适用于法律语言研究的检索软件。自己开发的检索软件有以下优势:(1)进行现成的软件无法完成的分析;(2)更准确地进行多项分析;(3)输出形式更适合于自己的专题研究需要;(4)对分析的语料库规模没有限制。(Biber et al, 2000)

四、结　论

法律语料库的建设是一个很浩大的工程,将会耗费人们大量的精力。但是在目前的技术条件和环境下,按照科学的设计方案,法律语料库一定是可以建成的。法律语料库的建成和投入使用将会给法律语言学研究开辟一片新的天地,将会对法律语言本体研究、法律语言教学和培训以及法律实践活动产生积极作用和影响。

参 考 文 献

[1] 杜金榜.法律语言学[M].上海:上海外语教育出版社,2004.

[2] 何安平. 导读 [A]. 用语料库研究语言[M].Thomas, J& Short. 北京外语教学与研究出版社,2001:23-24.

[3] 李文中.语料库索引工具[A].语料库语言学导论[C] 杨惠中.上海:上海外语教育出版社,2002.

[4] 李文中.COLSEC 的设计思想和建库方案 [A].中国学习者英语口语语料库建设与研究[C]. 杨惠中,卫乃兴. 上海:上海外语教育出版社,2006.

[5] 潘永楑.导读[A]. 语料库语言学[M]. Biber D, et al.北京:北京外语教学与研究出版社,2000:12-13.

[6] 王建新.计算机语料库的建设与应用[M].北京:北京大学出版社,2005.

[7] 文秋芳,王立非,梁茂成[M].中国学生口笔语语料库[M].北京:外语教学与研究出版社,2005.

[8] 杨惠中. 语料库语言学导论[M]. 上海:上海外语教育出版社,2002.

[9] Biber D, et al. Corpus Linguistics [M]. 北京:北京外语教学与研究出版社,2000.

[10 Kennedy, G.. An Introduction to Corpus Linguistics[M]. Beijing: Foreign Language Teaching and Research Press, 2000.

[11] Leech, G. Corpus annotation schemes [J]. Literary and Linguistic Computing, 1993(8).

[12] Sinclair, J. Corpus, Concordance, Collocation [M]. Oxford: Oxford University Press, 1991.

[13] Thomas, J. and Short , M. Using Corpora f or L language Research.[C].London: Longman, 1996.