中医药古文献语料库设计与开发研究①

刘耀[1] 段慧明[2] 王惠临[1] 周扬[3] 王振国[3] 李宏展[2]
[1]（中国科学技术信息研究所 北京 100038）
[2]（北京大学 计算语言学研究所 北京 100871）
[3]（山东中医药大学 文献研究所 济南 250014）

摘要：专业领域语料库是对专业领域文献进行自然语言处理的重要的不可或缺的基础，是对专业文本内容与意图进行深层把握的必由之路。本文通过对研究背景的分析，进一步明析了专业文献进行自然语言处理的必要性，并在对专业文献语料库的研究特点进行分析的基础上，深入探讨了专业语料库的设计思想及原理，同时，对语料库词类的标注信息进行了深入研究。成功地开发了针对专业领域语料库的辅助加工系统，为专业领域语料库建设提供了理论指导和技术支撑。

关键词： 自然语言处理 语料库 中医药古文献 知识工程

Research on Corpus Creation and Development of Chinese Traditional Medicine

[1]Liu Yao [2]Duan Huiming [1]Wang Hui-lin [3]Zou Yang [3]Wang Zhen-guo [2]Li hong-zhan
[1] (Institute of Scientific and Technical Information of China，Beijing,100038，China)
[2](Institute of Computational Linguistics, Peking University, Beijing, 100871, China)
[3](Institute of Chinese Medical History and Literature，Shandong University of Traditional Chinese Medicine，Jinan, 250001，China)

Abstract: Domain corpus is the important base of natural language processing for domain documents. It is necessary for gripping the deep meaning and content of domain documents. Based on the research background analysis, this paper clarifies the importance of natural language processing for domain documents. After analyzing the specialty of domain corpus, this paper discusses the idea and principle of domain corpus creation in a deep degree. Meanwhile, it also further researches on part of speech tagging information of corpus. Finally we develop an assistant processing system of domain corpus for the purpose of providing theory instruction and technique support for domain corpus creation.

Keyword: natural language processing; corpus; Chinese traditional medicine document; knowledge engineering

自然语言处理(Natural Language Processing，NLP)是一种对自然语言信息进行处理的技术，从语言学角度来说，自然语言处理也叫计算语言学(Computational Linguistics)。自然语言处理包括自然语言理解(Natural Language Understanding，NLU)和自然语言生成(Natural Language Generation，NLG)两部分。自然语言理解是指对自然语言的内容和意图的深层把握。在人工智能领域中，自然语言理解特指计算机对自然语言的内容和意图的深层

① ［基金项目］本文得到国家科技支撑计划项目（2006BAH03B00）、国家 973 项目（2007CB512601）、教育部人文社科项目(06JC870001)、山东省中医药科技专项项目（2003-14）的支持。
［作者简介］刘耀，男，1972 年生，副研究员，北京大学信息管理系管理学博士，北京大学计算语言学研究所出站博士后，主要从事知识工程与中文信息处理方向研究；段慧明，女，1957 年生，北京大学计算语言学研究所高级工程师，主要从事计算语言学方向研究；王惠临，男，1948 年生，研究员，北京大学信息管理系博士生导师，主要从事自然语言处理方向研究。

把握。自然语言生成是指从非自然语言输入到自然语言输出的处理。自然语言理解与自然语言生成互为逆过程。如何将自然语言技术引入到中医药古文献的处理中来，是我们多年从事的研究课题之一。

1 研究背景

中医学理论体系带有浓厚的自然哲学色彩，表现为长期的、非常稳定的形态，形成了以《内经》、《伤寒杂病论》为主体的相对封闭的框架。现代中医基础学科的分化，基本上是从原著派生出来的，因此，难以超越原著所固有的架构体系。中医基础学科奠基于《内经》学术体系，临床课程则与当时的中医医疗分科相对应。作为学科建设的主要标志，是各科教材的编写。特别是一版教材，扎扎实实地从文献研究人手，在前人的理论建树和实践基础上梳理出已经分化明显的学科，正如二版教材“前言”所说，是“把祖国医学系统地画了一个前所未能画出的轮廓，对提高教学质量起到了积极的作用”，[1]使中医学理论向规范化迈进了一大步。在短时间内，从浩瀚的文献中由博返约，提纲挈领地构筑起了现代中医药学的基本框架，满足了当时高等中医药教育的需要。[2]但是，在上述规范化过程中，受到近代科学思想，特别是近代西方医学的影响，同时也受当时教育模式的制约，在学科学术体系的架构过程中，许多重要的、有价值的理论与方法被忽略了。[3]例如中医“证”的规范化是多年来的重点研究课题。但是，由于文献的覆盖面有限，大量证型被遗漏。在未能对全部古代文献进行梳理，并对“证”的文献做出系统分析和归纳的情况下，简单的或者人为的分型有可能掩盖疾病的复杂性、多变性，引导医者的思维趋向单一和片面，即病-证-方的线性模式，并妨碍中医临床疗效的提高，以至于中医药界在建国五十多年内无重大发展。究其原因，中医学固有的理论与思想体系由于近代科学与教育模式等原因而被忽视；当代中医工作者文献研究不足，未能进一步深入挖掘古代文献中的学说、思想与理论，对中医基础学科群的理论框架与学科体系进行充实、完善。

另一方面，中医古籍文献整理研究，是必不可缺的，并且人们企盼着能从古籍文献整理研究入手，起到保持中医学术特色的作用，认为这是按着中医学固有规律向前发展的最佳选择。因此，如何利用现代化手段，对中医药古文献进行深入加工，从而为智能检索和知识挖掘打开方便之门，也就成为当前中医药古文献的研究前沿问题，也是中医药信息化迫切需要解决的重要问题。经过多年的研究，作者认为建立针对中医药古文献的语言知识库，可以有效地解决这一难题。

2 中医药古文献语料库的构建意义

语言知识库（如：语料库、机器词典、句法规则库等）是自然语言处理系统不可或缺的组成部分，语言知识库的规模和质量在很大程度上决定了自然语言处理系统的成败。这已经是计算语言学研究者和自然语言处理系统开发者的共识[4]。特别是中文信息处理尤其需要重视知识库的建设。这其中更以语料库与词典的建设为重中之重。基于语料库的研究具有以下特点：

(1)基于语料库的研究是实证性的，能够用来分析自然环境下的实际模式；

(2)能以大量收集起来的自然文本作为语料库研究的基础；

(3)能大量使用计算机作为分析工具；

(4)能同时使用定性和定量分析手段。

我国古代医家善于从前人的文献出发研究医理，探索规律。他们往往通过博览群书，凭借笔记与大脑记忆来搜集资料和积累经验。这种实证的经验主义方法在计算机技术出现之后得到了强化，日益发展的计算机技术既增强了个人搜集医学资料的能力，又提供了处理资料的强大工具。建立中医古籍语料库的目的，就是要运用计算机技术通过语料库来研究古代医学文献。与传统的医家相比，用语料库来研究古代医学文献主要有两个特点：

一是突破了材料的限制，计算机强大的搜索能力使古代文献研究从过去的重在材料的搜集转变为重在对材料的处理和对医学规律的总结；

二是突破了个人的因素，穷尽式的搜索保证了医学资料的完整性，能够最大限度地避免由片面的材料得出片面的结论，增强了研究结论的普遍性和科学性。

语料库是贮存和处理语言材料的仓库，但它并不是语言材料的简单堆积；由于中医药古籍的特殊性，古籍语料库跟其他的语料库又有所不同。在对语料库进行规划时，必须根据中医药古籍语料自身的特点来确立建库原则。

3 语料库设计思想与原理

一般而言，一个计算机语料库的功能主要和下面三种因素密切相关，即语料库的规模、语料的分布和语料的加工深度。因为语料库容量的大小直接影响到统计结果的可靠性，语料分布的考虑则关系到统计结果的适用范围，而加工深度则决定了该语料库能为自然语言处理提供什么样的知识。

在建立语料库之前，首先必须要弄清楚建立该语料库的目的和组建原理。目前的语料库主要是针对语言学研究而建立的，包括：方言研究语料库、对比研究语料库、平行语料库、多语言语料库等。针对专业知识进行语料库的建设，目前鲜有人尝试，因此，中医药古文献语料库的建立的原理也就成了我们首要解决的问题。

中医药古文献语料库的建设和研究对中医药术语规范化研究，词的切分和属性研究，术语语义研究，字频、词频统计和词典编纂等方面具有重要的意义。在中医药语料自动标注生成的整个过程中，分析其过程就显得极为重要。从分析过程看，首先是词类分析，其次是语料的标注，语法信息分析及专业属性的层次越深，则语料标注就会越准确，其中语法信息包括词类信息、子类信息、语义信息、格助词添加等信息，专业属性又包括专业分类体系与知识结构，语料中每个词条的语法信息及专业属性需要同语法规则和相应的子类相结合，以实现由词项来自动标注，这是中医药语料库建设的核心技术之一。由于中医药古籍的数量有限，所以，我们希望穷尽中医药古文献，另外，由于采取是自动标注，必须进行机器学习，建库之初，应注意文献题材的多样性。

另外，词汇经过语义标记之后，需要建立符合医学知识结构及医学知识体系的知识架构，建立知识连结的轨迹，使全文检索从“索引式”提升为“思维联系式”的检索，进而实现对中医药文献所包涵的医理进行分析与研究的目的，因此，我们首先对词类的标注信息进行了深入研究。

4 词类信息的分类与标记

词语的分类既是任何一个自然语言处理系统的基础也是语法信息词典开发的基础。因为语法词典既要描述每类词都有的共同的语法属性，又要分别描述各类词特有的语法属性，只有这样，语法信息才会充分、完备，而又不致过于冗余。

4.1 通用词语的分类

在通用词汇方面，我们采用了北京大学计算语言学研究所俞士汶教授的《现代汉语语法信息词典》[5]的分类体系，该语法词典的词类体系是在朱德熙先生的语法理论指导下，依据词的语法功能建立的。该词性标注使用的是小标记集[6]。它除了《现代汉语语法信息词典》中的 26 个词类标记（名词 n 、时间词 t 、处所词 s 、方位词 f 、数词 m 、量词 q 、区别词 b 、代词 r 、动词 v 、形容词 a 、状态词 z 、副词 d 、介词 p 、连词 c 、助词 u 、语气词 y 、叹词 e 、拟声词 o 、成语 i 、习用语 l 、简称 j 、前接成分 h 、后接成分 k 、语素 g 、非语素字 x 、标点符号 w ）外，增加了以下 3 类标记：

①专有名词的分类标记，即人名 nr ，地名 ns ，机关团体单位名称 nt ，其他专有名词 nz ；

②语素的子类标记，即名语素 Ng ，动语素 Vg ，形容语素 Ag ，时间语素 Tg ，副语素 Dg 等；

③动词和形容词的名词用法标记 vn ，an 和副词用法标记 vd ，ad 。合计约 40 个左右。

同汉语信息处理学界的某些研究相比，这是一个小标记集。尽管使用的是小标记集，但由于规范及据此加工的语料库同《现代汉语语法信息词典》是紧密联系的，当这些基础研究成果同应用研究（中文信息检索、中文信息提取、汉外机器翻译等）相衔接时，以语料中的词语及词性为入口，可以快速、准确地检索到词典中词语的丰富的语法属性信息。[7]

4.2 专业词汇

根据中医药语料库建设的实际需要，依照“功能分类”思想，提出了信息处理用中医药的分类方案，同时，为了尽可能避免产生交叉，我们采用了尽量减少类的数量，从而加强属性的描述，进行进一步的区分的原则，因此，对于中医药专业术语，全部标记为名词的下位类。分为：中医基础理论、藏象学说、气血津液、经络腧穴、病因病机、诊断、中药、方剂、伤寒与温病、症状、病证、治疗方法、中医药器械设备、体质、著作等 33 类。

中医药术语的语义类型命名原则为：“n_＋语义类型的汉字简拼＋阿拉伯数字”的方式进行命名。如：“病名”这一语义类型，标注为：/n_bm;“症状”这一语义类型标注为：/n_zz 等。在同级内出现重复者可用阿拉伯数字来进一步区分，其命名规则与词的标记如下：

（1）首选命名规则：取名词术语中每一个字的拼音首字母，为该术语命名。如：整体观念（zheng ti guan nian）命名为 ztgn。

（2）若有重复，则取术语中最后一个字的拼音第二个字母。以此类推，取第三个、第四个。如：整体观念（zheng ti guan nian）命名为 ztgn，若有重复命名，则标为 ztgi、

ztga、ztgn。

（3）再有重复，则取术语中倒数第二个字的拼音首字母。以此类推，取第二个、第三个。如：整体观念（zheng ti guan nian）命名为 ztun、ztan、ztnn。

（4）仍有重复，依(2)、(3)法类推，取术语中倒数第三个字中的拼音。整体观念（zheng ti guan nian）命名为 zign。

例：辩证论治（bian zheng lun zhi）若已有术语标为 bzlz，则可以根据其重复出现的顺序依次标为：bzlh、bzli、bzuz、bznz、bhlz、belz、bnlz、bglz。

5 中医药古文献语料加工系统的开发与实现

从 1992 年起，北大计算语言所就开始了语料库多级自动加工的研究。从 1993 年开始开发基于《现代汉语语法信息词典》的“词语切分与词性标注”软件，经多年的改进与发展，现在已相当成熟,无论是切分还是标注，其精度都已经达到了国内领先水平。[8]但是，语言信息处理系统也需要专业知识的配合。特别是实现中医药古文献有效切分与标注，必须在构建大型中医药专业词典的同时，再根据中医药古文献的不同类型的行文风格进行提取归纳，对软件结构及部分规则加以修改。因此，这是一个反复叠加的过程。即：切分－提取－再切分－再提取。其示意图如图 1 所示。

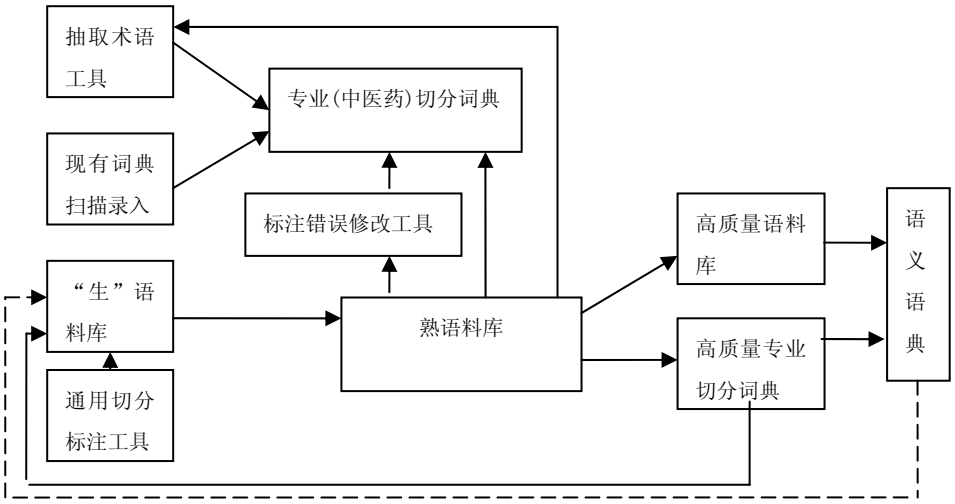

图 1：中医药语料库加工流程示意图

5.1 功能设计

该系统以北京大学计算语言学研究所自动切分与标注软件为基础,对语料加工所需的多种软件进行了开发与集成,形成了集加工、辅助修改及词典生成为一体的专业语料加工系统，主要有文件、编辑、检索、切分程序、词表替换、整理词典、抽词程序、环境设置、帮助等主要功能，如图 2 所示，现分别介绍如下：

（1）切分功能：自动切分标注，生成语料。是系统的主体程序之一。采用的是北大计算语言学研究所开发的自动切分与标注系统,该词语切分系统的抽取方法采用隐马尔可夫模型。

设文本 S 由单词串 W=w1,w2, …,wn 和标记集 T=t1,t2, …,tn 组成，汉语的词切分就是求使单词串和表记集的联合概率 P(W,T)为最大的词切分和词性标注的组合。P(W,T)可由如下隐马尔可夫模型近似求得。

在切分句子时，首先切出所有可能的切法，再用词典中单词出现的概率和语法规则中词性和词性的连接概率，计算所有切法的概率总值，取其概率值最大的为第一候选。

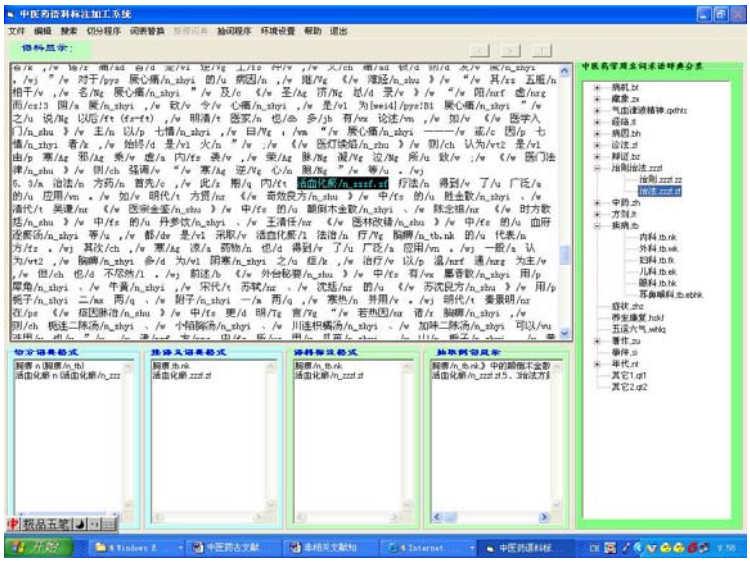

图 2：中医药语料标注加工系统界面 1

（2）词表替换功能

将切分标注软件中产生的错误改正，修改后的切分标注词语将自动复制到“新词文件”中，利用该功能便可实现词表的自动替换，将错误的切分全部自动替换成正确的切分。

实现本功能需要术语词典支持，“中医收词程序\文件目录\新词文件.txt”，待其中的术语以及“切分词典格式”输出的术语经验证确定无歧义后，方可导入“中医收词程序\可执行程序\SegTag.Lin\bin\lexicons\UsrLex6”（用户词典，收入用户定义的任意词或则词组）中。因为，这时提取出的只是术语的候选。在这些候选中，有些是术语，而有些不是术语，有些只是长术语的一个片段，还有一些在特定上下文中出现时才是术语，而在其它语言环境下出现时则不是术语。这些提取出的术语候选在脱离上下文后，即使人工校对也很难判断。因此，对于这些术语候选，必须进一步利用它们在特定文本中的上下文信息以及篇章结构信息进行确认。

（3）收词功能：生成三种形式的词典，并自动保存在指定文件夹中，如图 3 所示。

切分词典格式：升降浮沉 n｛升降浮沉/n_zh.sq｝

接语义词典格式：升降浮沉,zh.sq

语料标注格式：升降浮沉/n_zh.sq

并同时抽取例句显示，在词语前标有*号：升降浮沉/n_zh.sq,酒，则浮而上至巅顶。又一物之中，有根升稍降，生升熟降，是升降在物，亦在人也。经云：*升降浮沉顺之，寒热温凉则逆之。如春夏宜加轻宣升浮之药，秋冬宜加重涩降沉之药，以顺春升

①“切分词典格式”这个决定了词典的性质与格式，即，切分词典的结果针对于切分软件（现用的切分词典是北大计算语言所建立的通用词典），它与“词表替换”功能相结合，经过大量切分验证后，确定中医药切分词典，导入“中医收词程序\可执行程序\SegTag.Lin\bin\lexicons\UsrLex6”（用户词典，收入用户定义的任意词或则词组）中。

②“接语义词典格式”：建语义词典。

③“语料标注格式”这个决定了标注的深度——上位、层次，和语料的可用性。包涵一部分词法，比如：词性等，也可以用于运算统计处理。

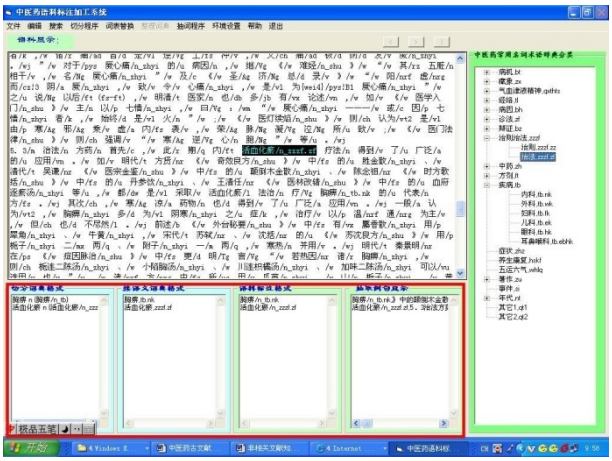

图 3：中医药古文献抽词与切分系统界面 2

（4）检索功能

该功能不仅能够进行多种形式的检索，同时，也可实现 KWIC(Key Word In Context 关键词)方式检索，将同一词汇的所有标注以高亮的形式显示在同一界面下，并支持同一界面下的修改与校对，不仅大大提高了人工辅助校对的速度，而且也确保了标注结果的一致性，如图 4 所示。

图 4：中医药古文献抽词与切分系统界面 3

（5）环境设置与词典整理功能

用户可以利用该功能对词典分类体系进行各种操作。详见图 5 所示。

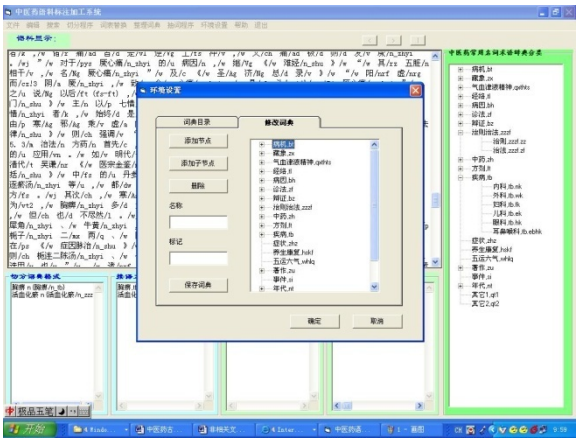

图 5：中医药古文献抽词与切分系统界面 4

另外，系统还包括文件、编辑、输出保存、无词典自动抽词功能等功能，在此不再赘述。

5.2 工作流程及结果

在以上工具的有效支持下，我们进行了大规模中医药古文献语料库的标注工作。其工作流程及结果示例如下：

a 生语料（原文）：

活血通窍法

用于瘀血阻络之头痛。症见头痛如锥如刺,痛有定处,头胀而痛剧,时轻时重,日久不愈,舌质暗,或有瘀斑、瘀点,脉涩。

b 汉语文本切分与词性标注工具标注后的语料：

活/v 血/n 通窍法/n

用于/v 瘀/Ag 血/n 阻/Vg 络/Ng 之/u 头/n 痛/a 。/w 症/Ng 见/v 头痛/a 如/v 锥/Ng 如/v 刺/n ，/w 痛/a 有/v 定/v 处/n ，/w 头/n 胀/a 而/c 痛/a 剧/Ag ,/w 时/Dg 轻/a 时/Dg 重/a ,/w 日/q 久/a 不/d 愈/Vg ，/w 舌/Ng 质/Ng 暗/a ，/m 或/c 有/v 瘀斑/n 、/w 瘀点/n ，/w 脉/Ng 涩/a 。

c 运用专业切分词典再次加工后的语料：

活血通窍法/n_zf

用于/v 瘀血阻络/n_bj 之/u 头痛/n_bm 。/w 症/Ng 见/v 头痛/n_bm 如/v 锥/Ng 如/v 刺/n ,/w 痛/a 有/v 定/v 处/n ,/w 头胀/n_zz 而/c 痛/a 剧/Ag ,/w 时/Dg 轻/a 时/Dg 重/a ,/w 日/q 久/a 不/d 愈/Vg ,/w 舌质暗/n_zf ，/m 或/c 有/v 瘀斑/n_zz 、/w 瘀点/n_zz ，/w 脉涩/n_zf 。/w

原始语料经过基本加工之后，文本由汉字串变成了“切分单位”的序列。“切分单位”基本是词，即语言学家所指的“句法词”。斜杠后的字母是根据该词语所表现的句法特性而加的标记。

经验表明，进行语料库标注，采取基于规则的方法与基于统计的方法相结合的策略是恰当的（特别是对于专业语料库），并且切分与标注同步进行是合理的。在进行这种标注时，语法词典可以发挥重要的作用。词典中的数以万计的词都已经划好了类，对标注的正确性与一致性可以起到基本的保证作用。标注程序只需集中力量解决兼类词的歧义消解及未登录词的确认与词性判定。

6 结论

中医药古今文献极为丰富，记载了大量方药疗疾防病的理论与经验，是巨大而宝贵的信息资源,中医药古文献语料库的建设和研究对中医药术语规范化研究,词的切分和属性研究,术语语义研究，字频、词频统计和词典编纂、信息检索、知识挖掘等都具有重要的意义，不仅是当前中医药古文献研究的前沿问题，同时也是中医药信息化迫切需要解决的问题。但专业语料库的建设并非一件简单的事情，经验表明，建库之初，应该注意以下几个问题：

1．对于专业语料库进行语料库标注，采取基于规则的方法与基于统计的方法相结合的策略比较恰当，这样可以充分利用专业词典，词典中的数以万计的词都已经划好了类，对标

注的正确性与一致性可以起到基本的保证作用。

2．语料标注时，应尽量减少大类的数量，进而加强属性的描述，这样可以有效避免类别的交叉，同时，也可以将专业术语与通用词汇进行有效的区分，有利于领域知识的发现与理解。

3．专业词汇要依据领域固有知识结构及体系进行描述，这样不仅可以有效地建立知识连结的轨迹，而且还可以建立该领域的知识架构，更加有效地进行专业领域的知识发现与挖掘。

参考文献：

[1]北京中医学院．方剂学[M]．上海科技出版社，1964

[2]王振国.当代中医基础学科群架构形成的历史局限性[J].山东中医药大学学报,2005(1):3-6

[3]张效霞，王振国．西医教育模式对中医基础学科体系形成的影响及反思[J]．中医教育，2004，23(6)：51

[4]俞士汶，段慧明，朱学锋，张化瑞．综合型语言知识库的建设与利用[J]．中文信息学报，2004(5):1-10

[5]俞士汶等．现代汉语语法信息词典详解（第二版）[M]．清华大学出版社，2003

[6]俞士汶，段慧明等．北京大学现代汉语语料库基本加工规范[J]．中文信息学报，2002(5):49-64

[7]段慧明，松井久仁於，徐国伟，胡国昕，俞士汶.大规模汉语标注语料库的制作与使用[J].《语言文字应用》，2000(2):72-77.

[8]俞士汶等.大规模现代汉语标注语料库的加工规范[J]．中文信息学报，2000(6):58-64