# The Construction and Observation of the Database of the Chinese Organization Names

Chen Hui

Department of Chinese Language and Literature, Beijing Foreign Studies University

***Abstract*—Based on Dynamic Circulating Corpus(DCC), we extracted the Chinese organization names with their context information, text information and so on, and built a dynamic-updated database of the Chinese organization names. On the database, we researched on all kinds of the distribution features, structural rules and the using regularities of the characters, words and strings in the Chinese organization names. At last, we developed a program to monitor them dynamically.**

***Keyword*—Chinese organization names, database, distribution, structural rule, dynamic monitoring**

# 中文组织名数据库建设与初步考察

陈 慧

北京外国语大学中文学院

**摘 要** 我们基于 DCC 动态流通语料库，抽取中文组织名实例与上下文信息、文本外信息，建设了一个动态更新的中文组织名数据库，并对数据进行了初步的分布特征研究、结构规则研究、字词符号使用研究和动态监测实验研究。

**关键词** 中文组织名，数据库，分布，结构规则，动态监测

## 1. 研究背景

中文组织名一直以来并未得到语言学研究和应用语言学研究的重视。而实际上，对它的了解对于中文信息处理、专名研究、社会语用心理研究、语言规范化、社会管理、传媒研究等都有很大的价值和意义。据《中国语言生活状况报告》2005-2013 年历年的统计结果，中文组织名在词语种数中的比例稳定在 36%左右，词语在不同年度中使用差异最大的是组织名，分别占到各年词种数的 40%－43%。

就拿中文信息处理来说，近二十年来中文组织名识别成为了各种统计技术的“沙场”。然而从中文分词评测结果来看，中文组织名识别仍是中文分词标注工作的瓶颈。组织名识别除了要应用成熟的技术，还要应用相关的语言知识。和其他词语成分的语言研究相比，中文组织名的很多基本问题没有得到很好的研究和解决。2001 年起我国建造了 DCC 动态流通语料库（Dynamic Circulating Corpus）。目前 DCC 动态流通语料库以几十家各类媒体十二亿字符次的年增长速度不断扩充，成为了我国规模最大、媒体覆盖面最广的语料库。DCC 依据流通度对主流报纸进行抽样，采录真实、语言规范的新闻文本。经过了文本预处理、分词标注和领域分类。经历了从 2001 年至今的动态更新，能实现历时稳态和实时动态研究的要求。我们基于 DCC（Dynamic Circulating Corpus）动态流通语料库，建构了一个组织名资源库，并在此基础上从多个维度进行统计、分析、考察、思考，获得了众多有价值的信息。现在我们就简要汇报我们的资源库建设及初步的一些考察结论。

## 2. 中文组织名资源库建设

我们所基于的语料如表 1、表 2 所示：

该语料库我们运用中科院自动化所分词标注系统进行分词标注，该系统的基本特点是：训练语料来自北京大学计算语言学研究所建设的《人民日报》六个月的语料；系统的命名实体识别模块完全通过统计技术训练获得；整个分词系统没有词典和规则等资源的支持。该分词系统在我国 2004 年 863 分词评测中取得了优秀的成绩，且在组织名识别方面的准确率达到了国内领先水平。我们的组织名考察研究工作就建立在此分词标注系统分词的基础上是更有意义和价值的。

我们从 DCC 主流报纸语料库中提取得到的组织名规模如下：

我们的组织名资源库结构如图 1 所示：

表 1：中文组织名研究语料库中的组织名规模

| | 总数（token） | 种数（type） | 平均频次 |
|---|---|---|---|
| 词语 | 247,257,749 | 8,750,105 | 28.258 |
| 组织名 | 3,954,716 | 615,681 | 6.423 |
| 比例 | 1.60％ | 7.04％ | 22.73％ |

表 2：中文组织名研究语料库语料量统计表

| 年度 | 媒体 | 语料量(字节) | 文本散布数 | 词语总数 | 词语种数 |
|---|---|---|---|---|---|
| 2002 \| 2006 | 北青报 | 514,664,332 | 341,770 | 83,941,419 | 2,263,096 |
| | 北京晚报 | 309,507,162 | 212,581 | 33,668,764 | 1,459,554 |
| | 法制日报 | 160,664,324 | 74,120 | 32,443,659 | 707,298 |
| | 环球时报 | 71,661,318 | 33,033 | 12,876,099 | 829,325 |
| | 人民日报 | 283,673,378 | 141,520 | 39,635,334 | 1,702,745 |
| | 羊城晚报 | 251,773,337 | 184,604 | 44,692,474 | 1,788,087 |
| | 总计 | 1,591,943,851 | 987,628 | 247,257,749 | 8,750,105 |

本研究专门设计实现了一个中文组织名辅助校对系统，界面如图 2 所示。

## 3. 中文组织名宏观分布特征考察

本研究基于 DCC 动态流通语料库，首次对我国现存的组织名的频率分布、领域分布、历时分布、报纸分布、字长分布进行了全面细致的考察。得到了以下有价值的分布特征：

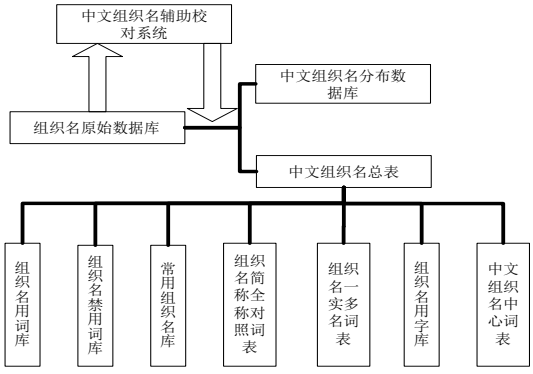


图 1：组织名资源库结构

1、频率分布特征：平均每个中文组织名出现 6.4 次。约 63％的组织名仅出现一次；约 17％的组织名出现两次；约 4.3％的组织名频次在 10 以上。排名第一的教育科研类组织为——北大。排名第一的国际、外国体育类组织为——韩国队。

2、领域分布特征：政治类语料出现的中文组织名最多，环境类语料的中文组织名种类最少。法制类、经济类的组织名种数在各领域类词语种数中的比例是最高的。领域独用组织名的比例较高，平均达到 56.63%的独用率。高频组织名在每个领域中都有可能出现。因此其领域特征确实不明显。

我们发现，衡量组织名是否进入通用领域，领域分布特征比频次更可靠。我们新创了“领域表征值”概念并给出操作标准，如，中国人民大学的政治领域表征值和法制领域表征值十分相近，说明中国人民大学至少在这两个领域上有很强的表征能力。再如，“清华大学”和“北京大学”都在政治、教育领域方面的领域特征很强，但是清华大学的政治领域特征强于教育领域特征，北京大学的教育领域特征强于政治领域特征。

3、历时分布特征：每年独用的组织名种类约占当年全部组织名种数的 2/3，独用组织名总数约占全年组织名总数的 1/5。年度独用组织名一般为频次为 1－2 的组织名。如频次较高，则为当年较热门的组织名。相邻两年会重复用到的组织名只有大约 1/5。两年共用的词语一般也就是多年共用的高频词语和历时关注度较高的非高频词语。对于组织名而言，也是如此。

4、字长分布特征：组织名的长度很不确定，在 2－17 的范围内均有分布。

组织名字长越大，其频次越低。从种数来看，字长为 6 的组织名最丰富。从词总数来看，三字组织名总数最多。频次和字长呈正相关。

下面将对 2－4 字长的组织名的形式特征进行进一步

图 2：中文组织名辅助校对系统

考察分析。字长为 2 的组织名都是组织名简称。字长为 3 的组织名结构最多的是“2＋1”式， 字长为 4 的组织名结构最多的是“2＋2”式， 更多字长的组织名结构实际上是在此基础上发展而来的。

5、报纸分布特征：报纸独用的组织名约占该报纸组织名总种数的 2/3。

《北京青年报》词语总数、词语种数均为六份报纸之冠。组织名分布、独用组织名比例是所有报纸中最平均的；《环球时报》的规模最小，所关注的组织名范围更集中，报道范围也更集中；《人民日报》上的组织名高度集中。更关注一些重要的、官方的组织机构的社会活动；《羊城晚报》组织名独用比率大，对于其关注的组织，其报道量也并不多；《法制日报》的组织名最丰富，关注的组织范围更广泛，报道范围更广泛。其报纸上载的组织名与其他报纸的差异性很大。

## 4. 中文组织名结构规则研究成果

汉语是没有形态变化的语言，如果单纯用西方语言的命名实体识别方法——主要依据机器学习和统计进行组织名识别——效果并不理想。因为单纯的统计模型无法解决数据稀疏和用词随机性带来的问题。因此中文组织名的识别必须引入规则。在这方面，我们全面查阅了从《马氏文通》至今的语言学文献与中文信息处理文献，发现这方面的研究主要集中在某一类别组织名的结构规则描写（如高校名、企业名）、中心词统计分析、用词分析、组织名形式化分类。主要的问题是，规则稍嫌琐碎，难以操作。随机获取的数据不能反映真实语料中的组织名分布。我们将在前人研究基础上，基于识别结果，以提高识别精度为目标，深入剖析组织名的中心词、上下文、结构等，建立了一套组织名规则研究的初步成果。其体系详见图 3。以《中心词词表》为例，其中包括：52 个单义中心词、26 个兼类中心词、25 个简称中心词、8 个小概率中心词、19 个非组织名中心词。我们依据中心词对全部组织名进行了形式化分类，并对每一类组织名进行了规则描述，以企业类组织名为例：

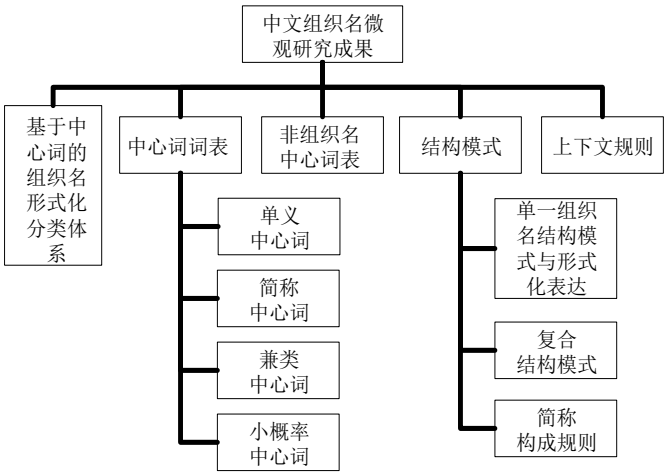


图 3：中文组织名结构规则研究成果

〈组织名〉::=&&{〈地名〉}〈字号〉〈内容说明词〉〈中心词〉

〈地名〉::=&&〈国名〉|〈名词：表地名〉|〈地名〉〈方位词〉|〈处所词〉

〈企业类中心词〉::=&&（中心词限定成分）〈中心词〉

在上下文规则研究方面，我们初步选取的是“英国广播公司”、“中国证监会”和“清华大学”这三个代表性词语进行了前接续成分关系、后接续成分关系的考察。厘清了直接搭配关系之外的伪搭配和间接搭配关系。并对每一种情况进行了规则描述。

## 5. 中文组织名识别结果字词符号考察：

我们基于识别结果，将组织名包含的通用汉字、其他字符、词性分布、词语使用、地名、字号、内容说明成分等行了初步的全面考察，考察内容如图 4 所示。数据库中共出现了 4883 个通用汉字，其中构成组织名简称最多的 10 个字为：中、大、航、军、北、共、部、行、铁、盟。我们对组织名表中前 60 万种[1]组织名识别结果，进行了二次分词，并根据分词结果的性质不断过滤错误的识别结果。根据字词符号考察结果操作我们的过滤程序发现，如果在组织名识别结果中引入禁用词性这一资源，能自动过滤 85475 种中文组织名。引入禁用字符串这一资源，能继续自动过滤 43930 种中文组织名。随后我们又对三大实词中的禁用词语进行了一一排查。

在以上结构规则研究和字词符号研究之后，我们提出了一个组织名识别流程，以说明中文组织名规则知识在识别中的具体应用。

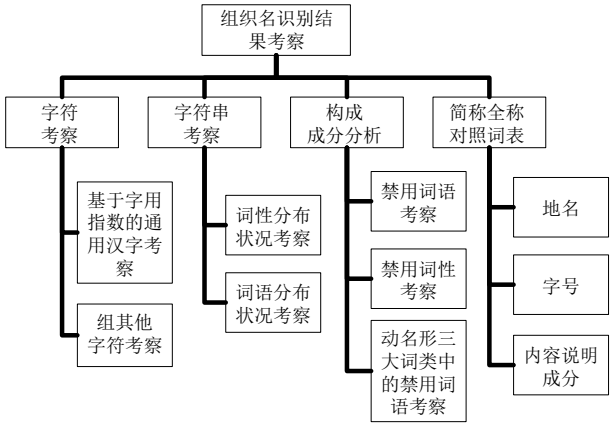


图 4：组织名识别结果考察内容

[1]之所以只取前 60 万，主要是因为后面的一万多条词语都是错误的识别结果。

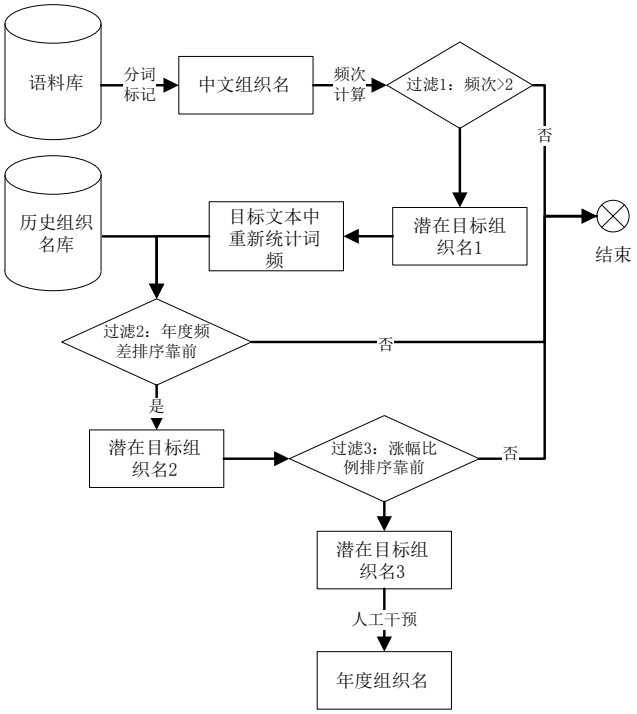


图 5：年度中文组织名获取流程

## 6. 中文组织名动态监测

我们研制了一套年度中文组织名获取流程如下：

然后我们选取 2005、2006 年六份报纸语料为语料进行实验，获得年度前十五位的政府组织名如下：民进党、国土资源部、铁道部、交通部、朝阳法院、中央综治委、红四方面军、国家药监局、丰台法院、国家安监总局、药监局、全总、红一方面军、市安监局、深圳市公安局。

除了整体监测，在这些数据的基础上我们也能很方便地实现对特定组织名的动态监测。如，可以通过年度频次等统计数据绘制其历时走势图，以了解某一组织名的历时分布状况。根据不同组织名的历时走势图，我们可以得到“持续高度关注型”中文组织名和“年度高度关注型”组织名。前者如“教育部”，后者如“中国女足”。动态监测的目标一般重点在“年度高度关注型”中文组织名上，但“持续高度关注型”组织名则反映了媒体历时稳定高度的关注情况。

## 7. 结语

中文组织名研究是一个值得深入和拓展的课题，而本文的研究只是一些初步的工作。下一步，我们将进一步研究名词、动词、形容词中的“禁用词”，结合组织名结构和语义词典，研究名词、动词、形容词在中文组织名结构中的条件限制和搭配规则，完善禁用词表，对组织名结构进行再分类，并将规则形式化，供中文信息处理使用。在组织名动态监测方面进一步深入研究。


图 6：“中国女足”、“教育部”历时走势图

## 参考文献(References)

[1] Anthony Giddens, Sociology (4thEdition), Peking University press, 2003.

[2] Language Situation in China: 2006, the Commercial Press, 2007.

[3] Huang Chang-ning1, Zhao Hai2, Chinese Word Segmentation: A Decade View, Journal of Chinese Information Processing, Vol.21, No.3, 2007.

[4] Yu Liming, Reseaches on Chinese Shortening, doctoral dissertation, Sichuan University, 2002.

[5] Zhan Weidong, A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing, Qinghua University Press, Guangxi Science and Technology Press, 1999.

[6] Zhangpu, Overall Thoughts on Dynamic Renewal of Language Information, Applied Linguistics, 2001.4.

[7] Zhang Xiaoheng, Wang Lingling, Identification and Analysis of Chinese Organization and Institution Names, Journal of Chinese Information Processing, vol. 11, No. 4, 1997.