中文地名结构的定性与定量分析[1]

唐旭日 陈小荷
南京师范大学文学院
{tangxuriyz@hotmail.com; chenxiaohe@njnu.edu.cn}

摘要：本文依据语言递归性本质特征，提出“地名成分”的概念，并以其为基本单位，对中文地名内部结构进行了定性和定量分析。“地名成分”为确定性的地名结构分析单位，由区别性词素、方位词素、描写性词素、类型词素和部位词素按一定顺序构成，各类词素又可分为开放性词素和封闭性词素两种类型。对地名结构数据库的统计分析显示，地名成分作为分析单位能够很好地揭示地名结构的规律性，并能为地名自动识别中各种机器学习模型提供基本知识结构框架。

关键词：地名成分；词素；地名；结构分析；地名自动识别

Structural Analysis of Chinese Toponyms

Xuri TANG Xiaohe CHEN
School of Chinese Language and Literature, Nanjing Normal University
{tangxuriyz@hotmail.com;chenxiaohe@njnu.edu.cn}

Abstract: Based on the recursive nature of language, this paper proposes a new analytical unit — Toponym Constituent for the analysis of toponym structure and conducts a quality and quantity analysis of the structure of Chinese toponyms on the bases of Toponym Constituent. Toponym Constituent is a determined structure which itself is constructed of distinguishing morphemes, orientation morphemes, descriptive morphemes, category morphemes and part morphemes in a fixed order. These different morphemes fall into two classes: open class and closed class. Investigation based on Chinese Toponym Structure Database indicates that the adoption of Toponym Constituent as basic analytical unit enables deeper insight on principles of toponym construction and provides a convenient knowledge framework for machine learning models in automatic toponym recognition.

Key words: Toponym Constituent; Morpheme; Toponym; Structure Analysis; Toponym Recognition

1. 前言

在近几年举行的 SIGHAN 中文信息处理 Backoff 中，包括地名自动识别在内的命名实体识别都是竞赛的主要项目之一。信息处理的快速发展迫切需要对地名，包括其内部结构和外部使用环境进行深入和全面的研究和描写。地名的研究在不同的语言学子领域中被归属为不同的类别。在语法上，地名被认为是表示地理名称的处所词，其语法分布与名词类一致，如被"不"修饰，而能被"不是修饰"，因而是名词的一个小类(文炼 1957)。在语义上，地名是专名的一种类型，是自然语言的一种专有名称，用以特指客观世界中的某一特定区域，因而归属于专有名词一类。而在语言信息处理中，地名自动识别是命名实体自动识别的三大任务之一(Hirschman and Chinchor

[1] 本文获国家 863 课题（编号：2007AA12Z221），国家社科基金项目（编号：07BYY050）和国家自然科学基金项目（编号：60773173）的支助，谨表谢忱。

1997)。

已有与地名相关的研究主要集中街道名称、方位词以及地名命名与民族历史、民族文化三个方面。(马庆株 1991)对 55 个街道的名称和构成方式进行了详细考察，使用词类如数字、动词、形容词、通名对不同形式的街道构成模式进行了细致分析。(张清常 1985, 张清常 1996a, 张清常 1996b)对北京街巷地名的变化及其历史、人文渊源、地名“数码化”（即数字在街巷命名中的应用）做了详尽考察。

在方位词方面，(文炼 1957)将其划分为单纯方位词和合成方位词，指出方位词具有虚词的特征，讨论了方位词的语法功能。(张清常 1996a)对北京街巷地名中 14 个方位词（东南西北，前后左右，上中下，内里外）的使用情况进行了细致分析，指出方位词在地名命名中存在系统性的特点。(邢福义，李向农，褚泽祥 1999)采用方位标的概念描述方位词，并将方位词分为准方位标和典型方位标两种类型，其中准方位标包括“旁”、“边”、“头”、“部”、“心”等，从而扩大了方位词的范围。

地名命名与民族历史、民族文化之间的紧密联系在以上文献中都有论及，而(邓慧蓉 2003)在这一方面有系统的研究，认为是一种特殊的文化现象，“是人类的认识成果，积淀了人类的思维方式和心理特征”。地名的用字选择、用词选择以及结构形式，都反映了民族文化，民族认知的特点，承载着民族历史的沉淀(吴志荣 2006)。

然而地名结构的研究还不能满足信息处理的需要。一方面，已有文献对地名的考察范围主要集中在城市、街道名称，而对于其他地理名称，如县、乡、镇、村的名称，以及各种地理要素，如河流、山脉、小山丘、溪流等名称的研究相对缺乏。而这些地名在在真实文本中大量出现。另一方面，计算机信息处理要求对地名结构的分析具有可操作性，要求用形式化的方法来表示地名的内部结构，这方面的研究还比较缺乏。目前能够提供给各种自动地名识别模型的语言学知识主要是地名用字和地名中的通名词表，如(刘开瑛 2000, 李丽双 et al. 2007)，而对于地名内部的组合结构，通名在地名结构中的地位及其与其他部分之间的组合关系，还有待深入研究。描写机制的缺乏使得在使用机器学习模型时直接使用四位标注[2]或六位标注来简单描述地名的内部位置关系。地名结构研究的滞后是地名自动识别的准确率提高所面临的一大问题。

本文在前人地名结构研究的基础上，提出地名成分的概念，并使用这一概念对北京大学 1998 年上半年语料库中的出现的所有地名（共计 18587 个）进行分析，建立起具有一定规模的地名结构数据库，并以此为基础，对地名成分中区别性词素、方位词素、类型词素、部位词素以及描写词素分别进行统计分析，探讨了这些要素参与地名命名过程中的一些文化、认知特点。基于地名成分和地名结构数据库的分析同时也给出了各种地名结构模式在汉语中的实际分布情况。这些分析一方面使得我们对地名的内部结构有了一个较为全面和深入的认识，同时也为地名自动识别提供了地名结构知识描写框架，为提高地名自动识别的精确度提供了语言知识资源。

2. 地名成分

一般而言(朱德熙 1982)，按照内部结构的复杂程度可将词分为单纯词和合成词两种类别。对地名而言，亦可分为三种类型[3]：

A） 简单地名（或单纯地名），如北京、高界、卢湾等

B） 复合地名，如北京市、玄武区、南大街、中华人民共和国等

C） 简称，如京、沪、辽、湘等

[2] 即以 B，I，E，S 代表汉字在地名中出现的先后顺序及其在结构中相对于开头、结尾的位置。B 代表词的开头汉字，I 代表词中汉字，E 代表结束汉字，S 代表成词汉字。

[3] 在语言信息处理中还有一种结构值得注意，即所谓的“多层复合地名”，如“北京市海淀区”、“南京市鼓楼区”等。这些多层复合地名具有词的一些特点，在意义上仅指向一个处所，而不是两个处所，如“北京市海淀区”仅指向海淀区，而北京市仅起到限制的作用。

其中简单地名由一个或多个语素组成。复合地名由简单地名与其他附属结构，或其他成分复合而成。语素、单纯词、复合词构成了地名分析的基本单位。

作为普遍使用的词汇结构分析单位，语素、单纯词、复合词具有高度的概括能力，适用于一般的词汇结构分析。然而由于地名是一类特殊的词汇单位，上述三个单位自然不能满足细致分析的需要，不能提供完整的形式化描写机制。例如，单词地名与复合地名可能存在界线不够清晰的问题。例如，对于“南大街”、“广州东”这样的地名，其中包含处所名和方位词，应归属为复合地名。但从功能分布角度讲，与“北京”、“广州”没有实际的差别，这样简单地名与复合地名之间的区分就失去了实际的意义。此外，语素、单纯词、复合词的高度概括能力使得构成成分之间的关系描述需要采用其他的机制，且存在信息量不充分的问题。例如“北京 市”与“北 三环”中“市”与“环”都是通名，但“北京”与“市”之间的构成关系与“北”与“三环”之间的构成关系存在明显的不同，而联合结构或偏正结构的高度概括不足以确定它们之间的构成关系。

为此，我们依据“递归性是语言的根本性质之一”[4]，引入一个用以描述中文地名内部结构的特殊结构单位，称为地名成分。地名成分的定义及其与地名之间的关系如下：

*地名成分：地名成分（**L-Struct**）是构成地名的基本结构，由区别性词素(**Dst**)、方位词素(**Ori**)、描写性词素(**Mod**)、类型词素(**Cat**)、部位词素(**Part**)和方位词素（**Ori**）按照先后顺序构成，各类词素可不同时出现；*

地名：地名是由一个或多个地名成分构成的语言结构形式。

地名成分这一概念的提出所遵循的一个重要原则是语言的递归性，即语言中复杂结构单位是通过简单结构的重复出现而构成的(钱冠连 2002)。历史上地名的不断变化要求其结构机制具有内在的创新性，而递归性是地名不断变化，不断创新的源泉。“地名成分”是地名递归构造地名的基本单位。采用地名成分，简单地名、复合地名都可获得统一的分析方法（表一）

表一：地名成分分析示例

北京	L-Struct [Ori<北> + Cat<京>]
广州东	L-Struct[Mod<广> + Cat<州> + Ori<东>]
南大街	L-Struct[Ori<南> + Mod<大> + Cat<街>]
昆仑山	L-Struct[Mod<昆仑> Cat<山>]
西山	L-Struct[Ori<西> Cat<山>]
北京市	L-Struct[Ori<北> + Cat<京>] L-Struct[Cat<市>]
广西壮族自治区	L-Struct [Mod<广> Ori<西>] L-Struct[Mod<壮族> Cat<自治区>]
南横街东口	L-Struct[Ori<南> Mod<横> Cat<街>] L-Struct[Ori<东> Part<口>]

地名成分中所使用的“词素”与(刘叔新 2005)中定义的词素相同，即可由一个或多个语素组成的词的构成单位。地名成分与单纯词、合成词的概念并不一样。地名成分是构成地名的最小单位。从范围上讲，地名成分可以是单纯词，也可以是复合词，以上的“北京”，“北三环”都是可以看作地名成分。地名成分的内部形式可以给出确定性描述。如定义规定的，地名成分由给出的一种或多种词素按先后顺序组合而成，结构中各类特征成分不要求同时出现，允许其中的一个或者多个词素缺失。对于某一特定的地名成分，其组成成分是确定性的，成分之间的先后顺序相对固定，成分之间的相互关系也是确定性的。

采用地名成分作为地名结构分析单位避免了简单地名和复合地名之间各成分层次纠缠不清的问题。地名成分之间与地名成分之间的界限是明确和清晰的。在表一中，采用地名成分分析，“广州东”与“南大街”被分析为一个地名成分，这个地名成分能够构成独立使用的地名。而“北京”与“北京市”则具有不同的结构类型，其

[4] 钱冠连. 2002. *语言全息论*. 北京：商务出版社.，第 155 页

中“北京”由一个地名成分构成，“北京市”由两个地名结构成分构成，“北京市”中包含有“北京”的地名成分分析。地名成分为简单地名和复合地名提供了统一的描写模式。

采用地名成分作为分析单位同时也使地名内部结构成分之间的相互关系得到有效表示。不同地名中具有相同功能的地名成分很容易得到确认。例如，“北京市”中的“市”与“广西壮族自治区”中的“壮族自治区”具有同样的分布特征和结构功能。同一词素在地名中的不同分布也得到了明确表式。“北京市”中的“北”和“广西壮族自治区”中的“西”都是方位词，而与描写词素的位置关系不同，结构方式也不一样，使用地名成分可以对这两种结构给出不同的分析方法。

采用地名成分分析的优势还在于其为地名在文本中的动态变化提供了描写机制。汉语中一个地名可能存在多种表达形式，如“西双版纳傣族自治州”，在应用中可能采用“西双版纳”、“西双版纳州”、“西双版纳自治州”等多种形式。这种现象使用简单地名和复合地名很难加以描述，而采用地名成分分析，则可以对地名使用中部分结构的脱落现象和规律给出概括性地描述。是观察表二：

表二：地名成分脱落

西双版纳傣族自治州：	L-Struct [Mod<西双版纳>] L-Struct [Mod<傣族> Cat<自治州>]
西双版纳：	L-Struct [Mod<西双版纳>]
西双版纳自治州：	L-Struct [Mod<西双版纳>] L-Struct [Cat<自治州>]
西双版纳州：	L-Struct [Mod<西双版纳>] L-Struct [Cat<州>]

采用归纳法可以得出一系列的规律。例如，从上例可以看出，如果一个地名中包含多个地名成分，则第一个地名成分中不能脱落，第二个地名成分可以整体或部分脱落。第二个地名成分在部分脱落中，类型词素往往被保留，而描写成分可能脱落。

3. 地名成分的元素分析

(刘叔新 2005)认为，词素是构词的基本单位。地名作为一种特殊类型的语言单位，其外在句法功能、语义特征必然导致其内部构成词素具有区别于其他词类的特点。这些特点主要表现在三个方面，其一是出现在地名成分中的词素类型是有限的，地名成分中包括了区别性词素、方位词素、描写词素、类型词素和部位词素；其二是在所有词素类型中，除描写词素外，其他词素类型所包含的词素成员是相对封闭的，其中的词素可以例举出来；其三是在描写词素具有鲜明的民族文化特点，与此相适应，其用字也存在一定的分布规律。

3.1 封闭性词素

地名成分中区别词素、方位词素、类型词素、部位词素四种类型的成分都是相对封闭的，这些类型的词素在汉语中基本可以全部例举出来。

3.1.1 区别性词素

区别词素在地名成分中起到区别性作用，即当两个地名其他构成一致的情况下，区别词素将两个地名区别开来。例如：

*小*西山 *老*城区 *民主*德国 *白*尼罗河

*大*西山 *新*城区 *联邦*德国 *青*尼罗河

从理论上讲，所有具有区别性功能的形容词都能够用来充当区别词素，但是在地名成分数据库中统计结果中，只有 13 个区别性词素，其中包括“大、小、新、老、古、白、青”等。

3.1.2 方位词素与部位词素

方位词素与部位词素在某种程度上具有相似之处。事实上，这里的方位词素与(邢福义，李向农，褚泽祥 1999)中提及的典型方位标一致，主要包括“东南西北，前后左右，上下中里”等。而(邢福义，李向农，褚泽祥 1999)所提及的准方位标则与部位词素基本一致，如“口”、“头”、“尾”、“底”等。方位词素和部位词素空间方位表达上存在一定的区别，一般而言，方位词素可表达物体外部的空间方向，也可表达物体内部的空间方位，而部位词素则表达物体内部的空间方向或处所。试比较：

岔口 城关 桥头 杉树脚

岔南 城南 桥西 港北

将“方位词素”和“部位词素”区分开来的第二个原因是两者的分布不同。方位词素一般即可位于描写词素之前，如“北大仓”、“北大街”，也可位于描写词素之后，如“百下区”、“察南”。而部位词素一般只能位于描写词素之后，不能出现在描写词素之前。

从空间方位认知的角度上分析，位于描写词素之前的方位词素与位于描写词素之后的方位词素表达了不同的空间认知结果。前者常表示以言语发出者为参照系，地名投射为空间的一个“点”，而后者以描写词素代表的处所为参照系，地名投射为空间的“面”。对地名结构数据库中的统计显示，出现在描写词素之前的方位词共有37种，包括“东“、”南”、“西”、“北”等单音节词素，也包括“中心”、“东北”、“北部”等双音节词素，含有前方位词素的地名有1235个，占总数的7%。出现在描写词素之后的方位词有24个，含有后方位词素的地名有1080个，占总数的6%。值得注意的是，地名中同时含有前方位词素和后方位词素的地名也有38个。

地名成分中出现的部位词素共23种，如“口”、“头”、“边”、“滨”、“脚”、“尾”“嘴”等，含有部位词素的地名有336个，占总数的2%。

3.1.3 类型词素—通名

类型词素也可被称为“通名”，通常用来表示地名成分所表示的地名的类型。类型词素往往表示语言使用者对地名所指处所的突出地理特征的认识。不同的处所，其地理特征差别较大，从而导致在地名中所使用的类型词素的种类和数量相对较多。地名结构数据库中的统计显示，类型词素共有351个。类型词素又可分为三种自类型：行政区划类、自然要素类和人工物类，分别举例如下：

行政区划类：市、国、区、街、州、省、村、镇、疆、屯、都、县、乡、盟等

自然要素类：湾、山、圳、江、海、湖、地、峡、浦、川、林、岛、岭、陵等

人工物类：里、路、坊、门、内、坛、城、道、园、亭、庵、港、楼、津、铺等

由本文第4节地名结构分析可以看到，近80%的地名包含有一个或多个类型词素，因此，在地名判断中，类型要素具有十分重要的作用。

3.2 开放性词素

3.2.1 描写词素

描写词素是指地名成分中除方位词素、区别性词素、部位词素和类型词素之外的，对地名所指处所进行描写的那一部分结构，如下面的划线部分：

“安定门”，“临高”，“麟游”，“凌源”“马达加斯加”

描写结构按照符号与所指的历史联系过程可将地名分为以下四类：

A. 音译。即将原有语言中对地名的发音借鉴过来。汉语中的地名中采用音译的不仅仅是国外的地名，还包

括一些少数民族地区的地名。而且，许多是在音译规范之前就已经广泛流传并被接受的译名。

B. 古字。如泗、淮等源于远古时代的单字在长期使用中形成的地名。这类地名往往单独运用即能够确定其所指。

C. 特征描写。摹状描写是地名形成的一种重要方式。就其历史来看，摹状描写中所描写的内容并不一样。有的地名所描写的内容是地形，有的是其中生物状况，有的是物产，有的则与重要历史事件、重要历史人物相关。在结构上，为满足描写需要，可以使用主谓、偏正、述宾、述补等多种形式。

D. 意愿表达。从地名的历史考察同时也可看到，地名本身不表示一种描述方式，而表现为一种语言崇拜，即对"言语行为"的一种夸大性的运用，表现出语言使用者的一种良好愿望。地名中许多褒义词的使用，都与这一良好愿望分不开。在结构上，偏正、主谓、述宾等多种结构都能够表达良好愿望。

4. 地名结构模式

地名成分为地名结构化分析提供了基本单位，在此基础上对大量地名进行处理、分析，然后进行统计分析，可获得对地名结构较为全面和深入的认识。我们在地名结构库上的分析发现，地名的内部构成具有一下几个特征：

A. 绝大多数地名由一个地名或两个地名成分组成。在地名结构数据库中，仅由一个地名成分构成的地名有12939个，占总数69%；

B. 描写词素是仅有的一个能够独立构成地名的词素，如"天目"，"偃师"，"福建"。类别是由两个地名成分构成的地名有5177个，占总数的28%，其他词素不能够单独构成地名。

C. 有的地名的内部结构相对复杂，最多由四个地名成分顺序组成，如下例：

井冈山市： L-Struct[Cat<井>] L-Struct[Cat<冈>] L-Struct[Cat<山>] L-Struct[Cat<市>]

凤阳府城镇：L-Struct[Mod<凤> Ori<阳>] L-Struct[Cat<府>] L-Struct[Cat<城>] L-Struct[Cat<镇>]

沟河庄乡：L-Struct[Cat<沟>] L-Struct[Cat<河>] L-Struct[Cat<庄>] L-Struct[Cat<乡>]

其中重复部分多为类型词素。从形成过程中分析，一方面是因为地名的命名与以地理特征相关联，而地名特征往往需要使用类型词素表示；另一方面是因为在历史变迁过程中一些行政区划名称附加在地名之后固化而成。

D. 地名的内部结构模式相对复杂。对地名结构数据库中的结构类型统计（见表一）显示，虽然类型1、2、3的出现频次占到了总数的77.8%，但是其他类型的分布相对分散，且所占比例也比较大。

表一：频次超过200的地名结构类型

类型序号	结构类型	频次	比例	例
1	L-Struct[Mod<> Cat<>]	7724	41.5%	狗叫屯，古巴共和国，古交市，三惠桥
2	L-Struct[Mod<>]	3510	18.8%	天目，偃师，福建
3	L-Struct[Mod<> Cat<>] L-Struct[Cat<>]	3258	17.5%	福井县，九台市，酒泉市，喀拉山山口
4	L-Struct[Mod<> Ori<>] L-Struct[Cat<>]	392	2.1%	阿北乡，巴东县，巴南县，白下县
5	L-Struct[Mod<> Ori<>]	348	1.8%	安西，白面下，保北
6	L-Struct[Ori<> Cat<>]	342	1.8%	北碚区，北道区，北甸子

	L-Struct[Cat<>]			乡
7	L-Struct[Ori<> Mod<> Cat<>]	327	1.8%	北冰洋，北大仓，北大街
8	L-Struct[Ori<> Cat<>]	286	1.5%	北部湾，东庄，南海
9	L-Struct[Cat<>] L-Struct[Cat<>]	218	1.1%	厂甸，池河，关镇，界岭
10	L-Struct[Mod<>] L-Struct[Mod<> Cat<>]	202	1.1%	楚雄彝族自治州，椿树二期，广安大街

5. 结语

作为自然语言的一部分，地名的内部结构存在递归性特点，这是语言无限生命力的一种表现。上述分析显示，地名的递归构成所依赖的基本单位就是地名成分。因此，使用地名成分作为地名内部结构基本分析单位，能够有效地揭示地名的构成规律。地名内部的各类构成词素具有其自身特点和分布规律。而基于地名分析的统计分析显示，虽然绝大多数地名在结构上具有很强的规律性，但是其内部构成模式相对复杂，存在多个地名成分构成一个地名的现象。这些分析说明在地名的自动识别中需要采用规则和统计相结合的方法。基于地名成分的结构分析，为统计分析模型提供了基本知识结构表示框架。如果将地名识别看作是一个隐藏状态标注问题，那么地名成分中的各种结构成分就给出了地名的内部状态标注集合。这一状态标注集合能够为基于最大熵、条件随机场等状态标注模型的地名自动识别系统提供语言学基础，为提高地名自动标注的精度提供条件。将基于地名成分的地名结构分析结果与机器学习模型相结合，构建高效的地名自动识别系统，也是我们下一步要研究的主要任务之一。

参考文献

Hirschman, L. & N. Chinchor. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Fairfax, Virginia.

文炼. 1957. *处所、时间和方位*. 上海: 上海教育出版社.

朱德熙. 1982. *语法讲义*. 北京: 商务印书馆.

李丽双, 黄德根, 陈春荣 & 杨元生 (2007) 基于支持向量机的中文文本中地名识别. *大连理工大学学报*, 47, 433-438.

邢福义，李向农，褚泽祥. 1999. 时间方所. *语法研究入门*, ed. 马庆株等, 472-483. 北京: 商务印书馆.

刘叔新. 2005. *汉语描写词汇学*. 北京: 商务印书馆.

刘开瑛. 2000. *中文文本自动分词与标注*. 北京: 商务印书馆.

吴志荣 (2006) 地名用字琐谈. *地图*, 2006, 42-43.

张清常. 1985. 明清以来北京城区街道名称变革所涉及的一些语言问题. *二十世纪现代汉语词汇论文精选*, ed. 周荐, 62-67. 北京: 商务出版社.

--- (1996a) 北京街道名称三题. *中国语文*, 1996, 428-432.

--- (1996b) 北京街道名称中的 14 个方位词. *中国语文*, 1996, 10-15.

邓慧蓉 (2003) 从中国地名透视汉族人的思维方式和社会心理. *学术交流*, 2003, 138-141.

钱冠连. 2002. *语言全息论*. 北京: 商务出版社.

马庆株. 1991. 街道名称及其构成方式. *《语言研究论丛》第六辑*, 153～177 天津: 天津教育出版社.