基于语料统计的现代汉语量名搭配研究

王萌 俞士汶 段慧明 孙薇薇

北京大学计算语言学研究所，北京，100871

wm@pku.edu.cn yusw@pku.edu.cn duenhm@pku.edu.cn ws@pku.edu.cn

摘 要：本文对现代汉语量词与名词的搭配进行了定量研究，设计并实现了一个有效的数量名短语的识别方法，基于识别结果，统计了部分名词受量词修饰的情况。该统计结果不但可以为名词的概率语法属性研究提供数据，而且在对外汉语教学中也有借鉴意义。就数量名短语识别这项任务而言，它属于名词短语识别的范畴，可以为完全句法分析提供名词语块。

关键词：现代汉语；基本标注语料库；数量名短语；量词；搭配

Research on the Classifier-Noun Collocation Based on Corpus

WANG Meng, YU Shiwen, DUAN Huiming, SUN Weiwei

Institute of Computational Linguistics, Peking University, Beijing, 100871

Abstract: This paper studies the classifier-noun collocations quantitively, and presents an effective method of numeral-classifier-noun phrase identification. Based on the result of identification, we calculate the distribution of the classifiers which can collocate with a certain noun. The statistical results can be used not only in the study of probabilistic grammatical attributes of noun, but also in teaching Chinese as a foreign language. As for the task of numeral-classifier-noun phrase identification, it is ascribed to noun phrase (NP) identification, and it can provide noun phrases for full syntactic parsing.

Keywords: contemporary Chinese; POS tagged corpus; Numeral-Classifier-Noun phrase ; Classifier; Collocation

1 引言

量词是汉藏语系的独有特点。与印欧语系不同，汉语中的大部分名词都有固定的量词（特别是个体量词）与其搭配，例如“三本书”、“五棵树”、“一块玻璃”（称为数量名短语）等。上世纪三四十年代，量词开始受到汉语语言学者的重视，对量词的划类、定名进行了深入的研究。量词是汉语词类中最后定名的词类，它显示了汉语的一个重要的语言个性，完善了汉语语法系统的建设。

目前，关于量词的研究多限于对量词做定性的描述或区分，而基于语料的定量分析并不多见。例如，词典告诉我们名词“问题”可以受 “个、些、种、项、类、系列”等量词修饰，可是这些量词在语料中的分布情况我们并不清楚。本文试图对量词与名词搭配情况进行定量的研究，调查量名搭配的实际分布情况，即考察某个名词在真实语料中与哪些量词搭配，且每个量词所占比例如何。该统计结果不但可以为《现代汉语概率语法信息词典》中名词相关属性的统计提供数据，而且在对外汉语教学中也有借鉴意义。

为了得到量名搭配的统计数据，本文实现了一个数量名短语的识别算法，利用量名搭配词典和简单的句法规则，在《人民日报》语料上进行了数量名短语的自动识别实验，结果表明该方

法简单有效。基于识别结果，我们统计了量词和名词搭配的分布情况。

2 数量名短语

2.1 数量名短语定义

数量名短语是名词短语的子范畴，它具有以下特点：

- 每个数量名短语必须包含一个量词；
- 量词修饰的名词是该短语的中心词；
- 量词左边相邻的词语通常是数词或者代词。[1]

例（1a）和例（1b）都是形式上最简单的数量名短语，如果在量词和名词之间加入修饰成分，它们可以形成更复杂的数量名短语，如例 2 至例 6 所示。

(1a) 一/m 本/q 书/n
(1b) 这/r 本/q 书/n
(2) 一/m 本/q 新/a 书/n
(3) 一/m 本/q 刚/d 买/v 的/u 新/a 书/n
(4) 一/m 本/q 老师/n 刚/d 买/v 的/u 新/a 书/n
(5) 一/m 本/q 老师/n 送给/v 我/r 的/u 新/a 书/n
(6) 一/m 本/q 由/p 儿童/n 出版社/n 发行/v 的/u 新/a 书/n

由此我们给出数量名短语的形式化定义，由三部分组成：

数量名短语→(代词|数词)+ 量词成分 +(名词短语|名词)
量词成分→(修饰语 + 量词)| 量词
修饰语→大|小|厚|薄|长|满|整

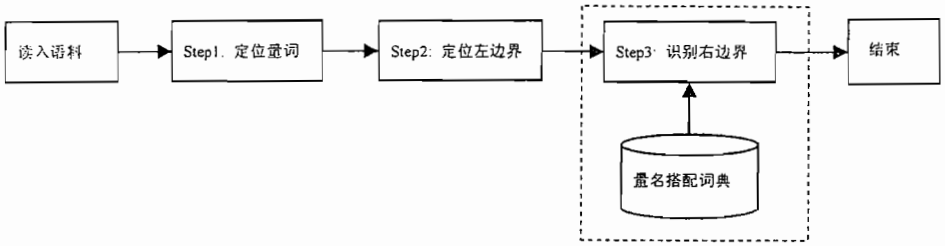
2.2 数量名短语识别流程

从上面的例子中可以看出，无论什么样的修饰成分被插入量词“本”和名词“书”中间，它们依然保持着选择性，这种选择性对界定数量名短语的边界很有帮助。因此，本文尝试利用量词和名词之间的依赖关系（即搭配），进行数量名短语的识别。那么，作为先验知识，我们需要一部量名搭配词典来体现这种依赖关系（词典的构造过程见第三节）。

通过对语料的观察并考虑到数量名短语的特点，确定其识别过程由 3 步构成，如图 1 所示。其中，输入是经过词语切分和词形标注的语料，输出是带有数量名短语标注的语料。该方法与现有的短语识别方法不同，第一步定位量词，因为量词的出现意味着一个潜在的数量名短语；第二步，定位左边界，根据数量名短语的定义，左边界词性范围有限，且出现位置相对固定，这使得左边界的判断非常容易；第三步，识别右边界，这是整个识别过程的难点所在，原因在于量名搭配的多样性。因此，本文提出了一种基于量名搭配词典和句法结构的识别方法，进行数量名短语右边界的识别，下文将对该方法做详细介绍。

[1]只有少数形容词可以修饰量词，如“大、小、厚、薄、长、满、整”等。

图 1 数量名短语识别流程



3 量名搭配词典的构造

如上所述，作为先验知识，需要构建一部量名搭配词典来体现量词对名词的选择性。构造一个量名搭配词典的数据基础是 ICL/PKU 的《现代汉语语法信息词典》(以下简称《语法词典》)和基本标注语料库。量名词典的结构如表 1 所示，每一量词后面列举它可以修饰的名词。下面将详细介绍词典的构建方法。

表 1 量名搭配词典样例

量词	搭配名词/后缀	
场/q	实例词典	座谈 梦魇 浩劫 大宴 剧目 流感 拍卖会 对攻战 事件 婚礼 传染病 好戏 竞技 关键球 报告会 小雨 和局 战事 卫冕战 争夺战 揭幕战 淘汰赛 泥雨……
	类词典	战 赛 雨 会 剧 病 害 灾
笔/q	实例词典	手术费 业务 资源 亮色 债券 细账 银款 汇票 安家费 安葬费 班费 搬运费 版税 办公费 包装费 保管费 保护金 保健费 保释金 保险费 保险金 保养费 保证金……
	类词典	费 金 税

3.1 基于实例的量名词典

首先，《语法词典》以词语的语法属性信息详实著称于学界，其中名词库中有若干属性描述了每一名词可与哪些量词搭配，例如，名词“土豆”，个体量词可为“个”，容器量词可为“筐”，度量词可为“斤，公斤，千克”，种类量词可为“种”，成形量词可为“堆”，不定量词可为“点，些”。所以，可以方便地从名词库中获取“种-土豆”等搭配实例。此外，量词库中设置了“后名”属性，列举某个量词可修饰的典型名词，例如，量词“杯”，可以修饰 “水、咖啡、酒”等名词。所以，从量词库中抽取“杯-水”等搭配实例。至此，我们从名词库和量词库分别得到了两部分量名搭配，然后对它们求并集，即合并每个量词能够修饰的名词，共得到 101302 条量名搭配实例。

其次，从语料中补充。使用的语料是 1998 年上半年《人民日报》语料，该语料经过了词与切分和词性标注。《语法词典》量词库中共有 525 个量词，将它们作为目标词（target word），在上述语料中寻找可与它们搭配的名词。为了保证抽取到更多的、可能的搭配，我们采用了 4 种假设检验的方法，分别是 Likelihood Interval Method、Likelihood Ratio Test、μ Test、χ2 Test 进行实验。每种方法各有偏置，例如，χ2 Test 通常对低频词语比较敏感，而 Likelihood Ratio test 对高频词的抽取效果较好。所以，用四种方法分别抽取，并取其交集，再通过人工检查，共得到 2390 条量名搭配实例。

最后，将从《语法词典》和语料中得到的量名搭配实例合并，共 102334 条，构成了基于实

例的量名搭配词典，简称实例词典。

3.2 基于类的量名词典

词典的收词是有限的，可利用的语料也是有限的，不可能穷尽某个量词所有可能搭配的名词。静态的实例词典的知识欠缺会造成某些数量名短语的漏标和错标，直接影响统计结果。为了解决这个问题，在实例词典的基础上进一步构建了基于类的量名搭配词典，简称类词典。

在中文里，指称同一类事物的名词往往是定中结构的复合词，且有相同的中心词，例如，“松树、桦树、柳树”的中心词都是“树”，而实际上，量词对中心词有很强的选择性。利用这个特点，类词典只收录了中心词。得到中心词的方法是，针对每一个量词，抽取它所修饰名词的中心词，对相同中心词计数，然后按照出现频次降序排列，依次选择中心词直至覆盖该量词所修饰名词总数的一半以上。设N是某个量词所修饰的名词集合，H是从N中得到的中心词集合。详细过程见算法1。

算法 1 抽取中心词

输入: $N = \{w_1, w_2, ...w_i ...w_n\}$

输出: $H = \{h_1, h_2, ...h_m\}$

变量: c=0, hash[ ]=NULL

1. For each w_i in N

1.1 抽取 w_i 的最后一个字，记为 h_i

1.2 if h_i exists in keys(hash)

1.2.1 hash [h_i]++

1.3 else

1.3.1 hash [h_i]=1

2. 对 keys(hash)的值按照降序排序

3. For each h_i in keys(hash)

3.1 add h_i to set H

3.2 c=c+hash [h_i]

3.3 if (c≥0.5*n)

3.3.1 break

例如，量词“场”，在实例词典中，与之搭配的名词共收录380个。在类词典中，只收录了中心词“战、赛、雨、会、剧、病、害、灾”（见表 1），拥有这些中心词的名词已经覆盖380个名词中的一半以上。在实际应用中，若“虫灾”没有被实例词典收录，而在语料中出现了“这是一场十分严重的草原虫灾”，利用类词典，匹配“X灾”，就能确定“场”修饰的名词是“虫灾”。

4 量名短语右边界识别算法

4.1 分析

如2.2所述，我们在语料中定位一个量词，使用量名搭配词典在当前上下文中选择可以被该量词修饰的名词，称为候选名词，这些候选名词就是可能的右边界。如果候选名词个数为1，则直接定位该名词为右边界。如果候选名词个数大于1，则需要一些句法结构来判断哪一个名词是正确的右边界。显然，第一种情况比第二种情况容易处理。所以，如果我们能通过某种方法，减少量词的候选名词，使第一种情况多多出现，就可以降低复杂数量名短语的识别难度。

例如，“从/p [百/m 余/m 种/q 参评/v 的/u 商品/n] 中/f 推荐/v 出/v [1 7/m 项/q 产品/n]”，该句中有两个数量名短语，其中“1 7/m 项/q 产品/n”的识别相对容易，因为量词“项”只有一个候选名词“产品”。显然，如果先将短语“1 7/m 项/q 产品/n”进行标注，就可以缩小量词“种”修饰的名词范围，即在“出/v”之前，所以很容易确定量词“种”修饰的名词是“商品/n”。

根据以上分析，利用量名搭配词典，设计了一种多遍扫描的标注方法（本实验中为5遍）。每遍处理一种情况，由易到难，在前一遍标注的基础上进行后续标注，每一遍标注都会缩小后续标注的搜索范围，从而降低难度。

4.2 算法

首先，对语料进行预处理，将“数量名”短语的识别范围限定在一个分句内，即以标点，；。！：？——结尾的句子。设经过预处理后的句子为 $S=w_1/t_1\ w_2/t_2\ \cdots w_i/t_i \cdots w_n/t_n$，$w_1^n$ 为词序列，t_1^n 为对应的词性序列。标注程序的大致流程如下（具体算法另文以详）：

1、第一遍，利用实例词典，处理情况（a）：候选名词的个数为 1。

2、第二遍，利用实例词典，处理两种情况，（a）：第一遍标注结果会使某些量词的候选名词的个数降为 1；（b）：量词右部相邻词语的词性是副词、动词、介词、连词，则确定同一分句内“的”字后面的名词为右边界。

3、第三遍，利用实例词典，处理两种情况，（a）：理由同上；（c）：从左往右顺序判断候选名词右部相邻词语的词性若是动词、副词、方位词、介词、助词，则确定该名词为右边界。

4、第四遍，利用类词典扩大名词搜索范围，依次处理情况（a）、（b）和（c）。

5、第五遍，所有词典失效，仅利用句法结构抽取，抽取形如“q_a_n” 或者 “q_b_n”的组合。

5 实验结果

5.1 数量名短语识别结果

测试语料是 1998 年《人民日报》1 月份后 10 天的语料，该语料经过人工标注了“数量名”短语，如“省局/n 还/d 专门/d 为/p 张家口/ns 增/v 开/v 了/u [一百二十/m 条/q 移动/vn 通信/vn 电路/n] 和/c [四/m 条/q 专线/n]”。标注程序的各项指标如表 2 所示，F-值达到了 94%。实验结果表明在该方法标注的语料上进行统计，得到的数据是可信的。此外，为了更细致地了解每一遍标注的情况，我们分别计算了它们的准确率，如表 3 所示。

表 2“数量名”短语标注程序的实验结果

数量名短语总数	7178
标记的数量名短语	7092
标记正确的	6724
准确率	94.81%
召回率	93.67%
F-值	94.24%

表 3 每一遍扫描的准确率

扫描	标记的数量名短语	标记正确的	准确率
1	4850	4727	97.46%
2	358	338	94.41%
3	1380	1222	88.55%
4	200	186	93%
5	303	251	82.57%
合计	7092	6724	94.81%

由表 2 可见，在第一遍中大部分数量名短语（4727/6724）都能够以很高的准确率被识别出来，这在很大程度上保证了整个方法的性能。但是，后续扫描的准确率有所下降，原因在于候选名词

个数的增加造成了识别的困难。在所有词典失效，仅使用句法模板时（第五遍），准确率最低。附录 1 列举了部分识别结果的样例。

5.2 量名搭配分布

我们用上述识别程序处理了《人民日报》1 月至 6 月的语料，然后统计了部分名词与量词的搭配情况，如表 4 所示。标注程序的错误会影响量词出现的绝对频次，而对各个量词出现的比例影响不大。

表 4 部分名词的搭配量词分布情况

名词	出现总次数	与量词搭配次数	量词分布情况
人/n	17788	2593	个 1177　些 725　代 126　名 85　种 78　口 65　位 64　家 53　批 48 岁 38　户 28　类 21　群 16　伙 16　点 11　号 11　辈 8　方 6　帮 5　拨 4　撮 3　行 1　队 1　圈 1　船 1　堆 1
农民/n	4206	369	户 99　位 92　个 69　些 46　名 32　岁 11　批 11　组 5　代 4
职工/n	6252	836	名 479　些 136　个 96　位 58　批 35　户 13　岁 6　群 1　成 1
专家/n	2320	283	位 97　名 65　些 64　批 37　个 17　代 3
同志/n	5890	363	位 164 些 116 名 35 批 20 个 19 级 7 岁 2
人员/n	5009	592	名 306　位 78　些 59　批 50　类 43　个 42　级 7　届 5　号 2
记者/n	14040	160	位 71　些 28　名 26　个 26　批 9
群众/n	6059	166	名 61　些 49　批 18　个 17　户 16　位 5
学生/n	2253	426	名 198　个 92　位 45　些 26　岁 20　级 15　批 12　届 9　群 7　帮 2
总理/n	4545	50	位 28　个 7　名 6　届 4　任 3　些 2
流氓/n	33	6	伙 2　个 2　些 2
罪犯/n	72	11	名 4 个 3　岁 2　批 1　些 1
大学/n	2264	127	所 87　些 18　个 13　批 8　种 1
政府/n	9143	762	级 589　届 122　个 33　些 16　任 2
公司/n	7309	810	家 544　个 133　些 111　批 16　种 6
城市/n	3094	505	个 338　座 91　些 64　批 9　类 3
图片/n	3646	62	幅 27　张 14　种 6　些 5　组 4　批 4　类 1　点 1
问题/n	12830	2200	个 1183　些 686　系列 129　种 128　类 35　点 11　批 10　项 5
方面/n	5664	810	个 699　些 101
措施/n	2712	795	项 274　系列 214　些 139　种 96　条 30　个 16　套 15　次 9　点 2
世纪/n	3154	448	个 448

表 4 的统计数据可以对一些语言现象提供例证。例如，某些量词也带有情感色彩，以“位、名、伙”为例，量词“位”表示敬重、褒奖的情感，表中数据显示，“位”多出现在“总理、专家、记者”等名词中。而量词“伙”表示厌恶、斥责的情感，它仅出现在“流氓”一词中。量词“名”的感情色彩不如前两者强烈，所以使用范围较广泛，出现在“学生、农民、总理、记者、专家、罪犯”等名词中。此外，表 4 也显示了量词“个”是使用最为广泛的个体量词。

6 结论

本文利用量名搭配信息和句法结构对名词短语的下位范畴—数量名短语的识别进行了探索，提出了一个有效的识别方法。基于识别结果，可以统计真实语料中量名搭配的分布情况，为现代汉语量词研究提供定量描述。此外，就数量名短语识别这项任务而言，它属于名词短语识别的范畴，是浅层句法分析，可为完全句法分析提供名词语块。

致谢 北京大学计算语言学研究所朱学锋老师为量名搭配词典提供了原始数据，博士生朱虹提供了搭配程序，南京师范大学李斌同学为本文的研究提供了测试语料，在此对他们表示衷心的感谢。

参考文献

[1] 俞士汶、朱学锋、王惠等. 现代汉语语法信息词典详解（第二版）. 北京：清华大学出版社，2003 年 2 月

[2] 何杰. 现代汉语量词研究.北京：民族出版, 2001

[3] 俞士汶　词的概率语法属性描述研究及其成果. 中文信息处理现代汉语词汇研究. 广州:广东教育出版社,2006 年 9 月第 1 版，227-283

[4] 俞士汶、段慧明、朱学锋、孙斌、常宝宝. 北大语料库加工规范：切分•词性标注•注音.汉语语言与计算学报, Vol. 13(No. 2)：122-158

[5] 朱德熙. 语法讲义. 北京：商务印书馆，1982 年 9 月

[6] Yu Jiangsheng, Jin Zhuihui, Wen Zhenshan. Automatic detection of collocation.Hong Kong: Proceedings of the 4th Chinese Lexica Semantics Workshop, 2003

[7] 俞士汶、段慧明、朱学锋、孙斌. 北京大学现代汉语语料库基本加工规范. 中文信息学报. 2002，16（5）：49-64，（6）：58-65

[8] 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. 中文信息学报，2004，18（5）：1-10

[9] 陆俭明. 现代汉语语法研究教程（第三版）. 北京：北京大学出版社.，2005 年 2 月

[10] 俞士汶、朱学锋、段慧明、张化瑞，以词义为主轴的综合型语言知识库. 第六届汉语词汇语义学研讨会论文集. 厦门：厦门大学. 2005 年 4 月，214-221

[11] 方芳、李斌. 基于语料库的数量名短语识别.第三届学生计算语言学研讨会论文集. 沈阳，2006 年 8 月

[12] 苏新春等. 汉语词汇计量研究. 厦门：厦门大学出版, 2001

附录 1 数量名短语识别结果样例

（rr1 说明第一遍标注，依次类推，符号“##”表示标注错误。）

该团/r　是/v　rr2[第一/m　个/q　全部/m　由/p　居住/v　在/p　美国/ns　的/u　华裔/n　青少年/n　组成/v　的/u　交响乐团/n]rr2　，/w　平均/a　年龄/n　17/m　岁/q　，/w　来自/v　美国/ns　rr1[22/m　个/q　州/n]rr1　。/w

当/p　她/r　在/p　##rr1[第一/m　个/q　被/p　评为/v　优秀/a　学员/n]rr1##　时/Ng　，/w　rr1[一/m　位/q　外国/n　教官/n]rr1　由衷/d　地/u　竖起/v　了/u　大拇指/n　出/v　了/u　汗/n　、/w　流/v　了/u　血/n　，/w　在/p　先人/n　们/k　的/u　欢呼声/n　中/f　rr4[--/m　座/q　飞桥/n]rr4　架/v　在/v　了/u　悬壁/n　之间/f　。/w