基于二元相关性的汉语三字格词语知识挖掘研究[1]

盛玉麒
山东大学 250100
yuqi-sheng@163.com

摘要：本文运用根词相关性理论和语料库语言学方法，在对 2003 年 1 月份《解放军报》200 多万字抽样文本分词标注基础上，建立二元相关性属性数据库，进一步进行词形相关性、词长（1+2）相关性、词长（2+1）相关性以及二元结构模式相关性统计分析，探讨三字格结构的准词语知识挖掘的可行性。结果显示，频次因素并不是知识挖掘的唯一可靠指标。将词长相关性和词类相关性相结合，并把未登录词语的发现和识别锁定在“名词相关性”上，具有结构合理性和现实可行性。

关键词：二元相关性 三字格 知识挖掘

A Study on Chinese Three-character Idioms Knowledge Mining Based on *Bivariate Correlation*

Sheng Yuqi
Shandong University 250100

Abstract: This paper establishes a bivariate correlation database by combining the root word correlation theory and corpus linguistics methods, based on more than 2 million characters sampling text from the Liberation Army Daily of January 2003. Meanwhile, this paper analyzes the correlation of the word form, word length (1+2), word length (2+1) and the bivariate structure model, and investigates the feasibility of quasi-word knowledge mining of three-character expressions. The results show that the frequency factor is not the only reliable indicator of knowledge mining. It is reasonable and feasible to combine the word lengthcorrelation with the part of speech collocation, and lock the unknown words into the "noun correlation ".

Keywords: bivariate correlation, three-character idioms, knowledge mining

一、词相关性

1．相关性理论

“相关性”通常指随机事件之间的关系。在数理语言学领域常指两个语言单位的关系程度。可以是字与字之间的相关性, 也可以是词与词、短语与短语、句子与句子之间的关系。

在计算语言学中，常用“共现”一词表示两个成分共同出现在一个语句中。相比之下，“共现”只是强调同时出现，而“相关性”则明确度表达了“关系的程度”。统计学有“相关分析”法，专门研究随机变量之间的相关性，包括偏相关、复相关、定序变量相关等不同类型相关性的统计分析。

本文所谈“二元相关性”是指任意两个词之间的关系程度。可以把这种关系想象成词的矩阵，纵横分别按照自然数列排列全部词，两两组合的节点数就是相关性组合的理论数值。

[1] 本文得到国家社科基金项目“基于语料库的汉语根词相关性句法模型研究”的资助，曾在澳门语言学会主办，中国社会语言学会、国际汉语教育协会协办的“澳门语言研究的回顾与展望暨庆祝程祥徽教授澳门从研从教 30 周年学术研讨会”（2011 年 10 月 23—26）上交流。

假设有 10000 个词，那么，相关性组合的理论数值就是 10000×10000=1 亿个。但是实际应用中绝不会有那么多。因为许多词之间由于句法功能的差异以及应用分布的关系，包括合理组合与非合理组合。例如下面的例子：

> 李嘉诚称，“事实上，中国人有好人，也都有差的；外国亦都有好有差，国籍没有特别的关系。”

没有分词的情况下，可根据标点符号分隔的两个相邻单位之间就存在相关性：“事实上，中国人有好人”，“ 中国人有好人，也都有差的”，等等。

分词标注词性后，得到如下文本：

> 李/nr 嘉诚/nr 称/v ，/w “/w 事实/n 上/f ，/w 中国/ns 人/n 有/v 好/a 人/n ，/w 也/d 都/d 有/v 差/a 的/u ；/w 外国/n 亦/d 都/d 有/v 好/a 有/v 差/a ，/w 国籍/n 没有/d 特别/a 的/u 关系/n 。/w ”/w

其中“李/nr 嘉诚/nr 称/v ，/w”算上标点一共 4 个单位，其中的二元相关性组合就有“李/nr 嘉诚/nr”、“嘉诚/nr 称/v”和“称/v ，/w”3 组。显然“李/nr 嘉诚/nr”和“嘉诚/nr 称/v”具有合理性，而“称/v ，/w”就不具有合理性。

根据相关性的程度可大致分为“高、中、低、无”四种。

相关性与结构关系不同，结构关系是具有内在逻辑规定性的关系，例如主谓、述宾、偏正、述补等基本句法结构关系。相关性是指“随机变量”之间的关系，随机变量是不确定性的量，并不能确定是否存在真正的逻辑关系，所以要通过统计分析来求证。

2．相关性的获得

获得相关性的途径有两条：

一是从理论预设为主，将所有的词建立二元搭配矩阵，然后逐一筛选排查；二是从抽样语料中进行统计分析，找出实际使用中存在的相关性组合。

前一种方法具有“穷尽性”，但是工作量大，可控性不足，因为人工筛选排查过程如何防止错漏是一个大问题。判断标准也不好掌握。因为词离开具体的语境后，往往很难判断其合理性。

后一种方法往往受限于抽样语料的规模。从理论上说，不管语料库多大，都具有不完备性，因此一定会有遗漏。二者相比，后者具有现实可行性。本研究采用基于语料库的统计分析方法进行词相关性的知识挖掘。

3．相关性的意义

用相关性来描述语言单位之间的关系具有统计学的意义。现代汉语关于词的定义中有“结构稳定性”、“历史继承性”、“社会通用性”等标准。其中的“结构稳定性”用相关性进行描述就属于“高相关性”。

新词语多属尚未达到“稳定性”的程度，可以用“中低稳定性”加以描述。因此，在新词语知识挖掘研究中，相关性研究就有了重要的意义。

从词长看，新词语一般多为 2—4 个音节，5 音节以上的很少。其中，双音节词与双音短语之间的区别难度最大。三字格中除了少数被收入辞典的惯用语等固定短语之外，多数为临时短语，真正的“三音节词”并不多。四字格中除成语外，情况与三字格类似。

固定短语一般都是从临时短语发展来的，都经过了“临时短语”——“准固定短语”的阶段，因此，从 3—4 音节的相关性组合中挖掘“准固定短语”应是新词语研究的一个重点。

二、抽样语料库

根据 2003 年 1 月份《解放军报》电子版文本语料，通过 word 字数统计结果为 2128619 个汉字，不计空格字符数为 3119202 个。

采用中科院计算所自动分词软件进程分词和标注词性作为预处理。

自动分词后得到 830797 词次，统计结果得到词种 37065 个。其中：单音词 4702 个，累计使用次 413186 次；双音词 24543 个，累计使用次 380400 次；三音词 4663 个，累计使用次 27167 次；四音词 2899 个，累计使用次 9398 次；五字以上词 258 个，累计使用次 646 次。列表如下：

表 1. 抽样语料库词频统计结果

词长	词种	所占比例	累计频次	所占比例
单音词	4702	12. 69%	413186	49. 73%
双音词	24543	66. 22%	380400	45. 79%
三音词	4663	12. 58%	27167	3. 27%
四音词	2899	7. 82%	9398	1. 13%
五字以上	258	0. 70%	646	0. 08%
合计	37065	100. 00%	830797	100. 00%

从表内可见，静态词种分布比例中双音词所占比例最高，达到 66. 22%；单音词和三音节词比例相当。但从动态使用频次看，单音词的所占比例最高，达到 49. 73%，其次是双音词，为 45. 79%，两项合计达到达到 95. 52%。三音节以上的词都比静态分布比例大幅下降。

这个结果一方面说明，在现代汉语中单、双音词的活跃程度，另一方面也反映出自动分词软件所依据的分词标准的局限性。因为自动分词软件基本上是以权威词典所收词语为分词依据，即所谓“语法词”或“词典词”。起码我们目前所使用的分词软件尚不具备动态“自动学习功能”，还无法识别标准词表之外的“未登录词”和“新词语”。这恰恰是新词语知识挖掘研究的重要任务和巨大的发展空间。

三、词形相关模式统计

“词形相关模式”指以词与词之间的相关性为特征的组合模式。按照 37065 个词种计算，二元相关性组合的理论数据应为 37065×37065=1373814225 组。实际统计得到 326490 组（含标点符号），去掉标点符号后共计 266924 组，仅占理论数据的万分之二。两个数据相差悬殊的主要原因在于理论数据为“任意两个词的组合”，而抽样语料库的统计结果是实际使用文本中的“相邻两个词语单位的组合”。

1.词形组合频级

266924 个二元相关性组合中，使用 100 次以上 233 组，累计使用 51740 次；使用 50–99 次的二元组 681 组；使用 40—49 次的二元组 946 组；使用 30—39 次的二元组 1433 组；使用 20—29 次的二元组 2476 组；使用 10—19 次的二元组 6553 组；使用 1—9 次的二元组 260371 组。

表 2. 二元相关性组合的频级分布表

频级	组数	所占比例	累计组数	频次	所占比例	累计频次
100↑	233	0. 09%	233	51740	8. 82%	51740
50–99	448	0. 17%	681	30575	5. 22%	82315
40–49	265	0. 10%	946	11669	1. 99%	93984
30–39	487	0. 18%	1433	16590	2. 83%	110574
20–29	1043	0. 39%	2476	24644	4. 20%	135218
10–19	4077	1. 53%	6553	53731	9. 16%	188949
1–9	260371	97. 54%	266924	397510	67. 78%	586459
合计	266924	100. 00%	266924	586459	100. 00%	586459

从上表可以发现，使用频级和对应的组数具有反比例关系，即频级越高，组数越少；频级越低，组数越多。频级在 10 次以下的低频区多达 260371 组，占总组数 266924 组的 97. 54%。

另一方面，从使用频次中发现，占总组合数 0. 26%的高频区（50 次以上）681 组相关性组合，累计使用频次高达 82315 次，占到总频次 586459 的 14. 04%。由此可见，使用频次对于知识挖掘的参考价值。

2．词长相关性

词长是词形的重要特征之一。从词长相关性看，三音结构有“1+1+1”、“2+1”和“1+2”三种模式。实际统计结果为，“1+1+1”模式有 23178 组、“1+2”模式有 54267 组、“2+1”模式有 50667 组，合计 128112 组。

表 3. 词长相关性分布统计结果

模式	组数	占总组数比例	累计频数	占总频次比例
1+1+1	23178	18. 09%	60464	19. 66%
1+2	54267	42. 36%	131694	42. 83%
2+1	50667	39. 55%	115343	37. 51%
合计	128112	100. 00%	307501	100. 00%

“1+1+1”属于“三元相关性”模式，按照降频选取使用频次 100 以上共得到 20 组。详见表 4。

表 4. “1+1+1”相关性组合高频样表

序号	词 1	词 2	词 3	频次	合理性
01	党/n	的/u	十/m	785	
02	新/a	的/u	一/m	547	
03	的/u	一/m	年/q	505	
04	了/u	一/m	台/q	429	
05	的/u	新/a	房/n	415	
06	**就/d**	**是/v**	**要/v**	293	√
07	时/ng	俱/dg	进/v	275	
08	与/p	时/ng	俱/dg	272	
09	**是/v**	**一/m**	**种/q**	242	√
10	在/p	新/a	的/u	153	

11	**有/v**	**一/m**	**批/q**	143	√
12	也/d	不/d	是/v	125	
13	多/m	年/q	的/u	123	
14	有/v	了/u	新/a	123	
15	要/v	有/v	新/a	117	
16	上/v	了/u	新/a	110	
17	多/m	万/m	元/q	108	
18	**几/m**	**年/q**	**来/f**	107	√
19	不/d	会/v	冻/v	107	
20	都/d	有/v	可/v	101	

从表 4 可见，20 组中只有“**就是要**”、“**是一种**”、“**有一批**”和“**几年来**”4 组具有句法结构的合理性，占 20%，其余几乎都不完整。由此可见，即使使用频次很高的组合，如“与时俱”和“时俱进”都是因为“与时俱进”的高频切分后形成的“伪高频”，没有句法合理性。

3．二元词长相关性统计

按降频选取“1+2”模式使用频次 100 次以上的共有 40 组，详见表 5。

表 5. “1+2”相关性组合高频样表

序号	词 1	词 2	频次	标记	21	团/n	党委/n	143	√
01	期/q	版条/q	1525		22	的/u	基本/a	139	
02	条/q	标题/n	1266		23	的/u	建设/vn	139	
03	十/m	六大/j	1083		24	的/u	同时/n	137	
04	个/q	代表/n	607		25	了/u	一个/m	124	
05	的/u	发展/vn	341		26	的/u	根本/a	121	
06	**新/a**	**战士/n**	316	√	27	**核/n**	**问题/n**	118	√
07	条/q	引题/n	259		28	**是/v**	**我们/r**	117	√
08	**江/nr**	**泽民/nr**	255	√	29	的/u	情况/n	117	
09	的/u	工作/vn	243		30	**新/a**	**世纪/n**	111	√
10	的/u	重要/a	232		31	的/u	政治/n	106	
11	的/u	问题/n	218		32	**一/m**	**系列/q**	106	√
12	的/u	思想/n	204		33	**江/nr**	**主席/n**	105	√
13	的/u	要求/n	182		34	的/u	生活/vn	104	
14	和/c	军队/n	175		35	的/u	时候/n	104	
15	的/u	精神/n	171		36	**胡/nr**	**锦涛/nr**	104	√
16	**是/v**	**一个/m**	170	√	37	和/c	人民/n	101	
17	**高/a**	**技术/n**	154	√	38	的/u	目标/n	100	
18	的/u	一个/m	151		39	**李/nr**	**岚清/nr**	100	√
19	的/u	基础/n	148		40	的/u	官兵/n	100	
20	的/u	历史/n	145						

上表所见，只有“**新战士、江泽民、是一个、高技术、团党委、核问题、是我们、新世纪、一系列、江主席、胡锦涛、李岚清**”等 12 组具有句法合理性，占 30%。其余都不具有合理性。

按降频选取“2+1”模式使用频次 100 次以上的共有 41 组，详见表 6。

表 6. “2+1”相关性组合高频样表

序号	词 1	词 2	频数	标记	21	他们/r	的/u	140	
01	**关键/n**	**词/n**	1242	√	22	精神/n	的/u	131	
02	建设/vn	的/u	415		23	**条件/n**	**下/f**	129	√
03	自己/r	的/u	339		24	取得/v	了/u	128	
04	工作/vn	的/u	335		25	国防/n	和/c	122	
05	军区/n	某/r	311		26	方面/n	的/u	122	
06	**本报/r**	**讯/ng**	296	√	27	建设/vn	和/c	121	
07	贯彻/v	十/m	263		28	关系/n	的/u	120	
08	进行/v	了/u	239		29	**基础/n**	**上/f**	120	√
09	发展/vn	的/u	229		30	发展/v	的/u	118	
10	官兵/n	的/u	228		31	**官兵/n**	**们/k**	117	
11	社会/n	的/u	192		32	军队/n	的/u	115	
12	部队/n	的/u	175		33	干部/n	的/u	114	
13	**特别/d**	**是/v**	168	√	34	**战士/n**	**们/k**	107	√
14	思想/n	的/u	166		35	提出/v	的/u	106	
15	我们/r	的/u	157		36	提高/v	了/u	106	
16	人民/n	的/u	155		37	事业/n	的/u	106	
17	群众/n	的/u	151		38	提出/v	了/u	105	
18	**我们/r**	**党/n**	148	√	39	国家/n	的/u	105	
19	重要/a	的/u	146		40	这样/r	的/u	104	
20	问题/n	的/u	145		41	学习/v	十/m	101	

从表 6 所见，只有“**关键词、本报讯、特别是、我们党、条件下、基础上、官兵们、战士们**”等八组具有典型的句法合理性，约占 20%。另外，还有许多“的”字结构如“自己的、建设的、官兵的、社会的、部队的、思想的、我们的、人民的、群众的”以及“动词+动态助词”如“进行了、取得了、提高了”等，具有句法“准合理性”，但不具新词语挖掘的价值，可以忽略。

消除或过滤“相关性”模式组合中存在的大量不合理组合的“冗余”信息，成为知识挖掘的关键。

分析这些不合理组合的词性关系发现，绝大多数都是因为在词性上不具有句法结构关系。例如，首位是助词、量词、数词、连词等的组合不具有句法合理性。此外，末位为连词、助词、副词等也不具合理性。为此，需要根据“词性相关性”对相关性模式做进一步筛选。

四、结构模式筛选

根据词性相关性进行统计，二元词词性相关性共计 1271 种模式（含标点符号），去掉标点后得到 1155 个组合类型。

其中，使用频次超过 100 的共有组，详见表 7。

表 7. 二元词性相关性组合高频样表

序号	**首位**	**次位**	**频数**	07	An	n,	473
01	A	n,	10746	08	A	v,	436
02	**A**	**u,**	**6159**	09	**A**	**y,**	**302**
03	Ad	v,	6156	10	An	vn,	297
04	A	vn,	1686	11	A	ng,	278
05	A	a,	653	12	**Ad**	**p,**	**271**
06	**A**	**m,**	**619**	13	**An**	**c,**	**261**
				14	Ag	n,	254

15	An	v,	233
16	**A**	**c,**	**202**
17	Ag	v,	194
18	**A**	**q,**	**191**
19	**A**	**d,**	**176**

20	**An**	**u,**	**152**
21	An	d,	121
22	A	r,	113
23	**A**	**p,**	**106**
24	A	an,	102

滤掉其中带有数词、副词、介词、助词、连词、语气词等 10 个相关性组合（表内标粗体行）后，还有 13 组具有句法结构合理性，重新列表如下：

表 8. 二元词性结构相关性组合高频样表

序号	首位	次位	频数	举例	
				模式 1+2	模式 2+1
01	A	n,	10746	高/a 技术/n；新/a 世纪/n	个别/a 人/n；文明/a 村/n
02	Ad	v,	6156	严/ad 要求/v；勤/ad 学习/v	深入/ad 到/v；认真/ad 做/v
03	A	vn,	1686	新/a 发展/vn；小/a 创造/vn	无
04	A	a,	653	小/a 聪明/a、高/a 标准/a	效/a 快/a；固定/a 好/a
05	An	n,	473	苦/an 日子/n；冤/an 干部/n	卫生/an 队/n；安全/an 网/n
06	A	v,	436	好/a 起来/v；小/a 混混/v	不好/a 找/v；
07	An	vn,	297	无	无
08	A	ng,	278	无	易爆/a 品/ng；忧郁/a 症/ng
09	Ag	n,	254	诸/ag 要素/n；主/ag 裁判/n	无
10	An	v,	233	无	迷茫/an 到/v；安全/an 度/v
11	Ag	v,	194	洋/ag 打工/v；速/ag 回家/v	无
12	A	r,	113	无	无
13	A	an,	102	新/a 矛盾/an；真/a 麻烦/an	无

上表所见，标为“无”的栏目，该相关性组合为“2+2”模式，没有“1+2”或“2+1”模式的用例。

可见，高频词性结构相关性组合模式用例，并没有得到我们所期望的结果。

原因有二：一是不应单纯考虑频次因素，而应扩大范围，从中低频区进行挖掘；二是应进一步从新词语性质上缩小目标作进一步的分析。

汉语中的虚词和代词、副词等都是“封闭性”词类，实词中动词和形容词属于“准封闭性”。“封闭性”和“准封闭性”词类中增加新词语的空间很小。名词是“开放性”词类。因此，新词语挖掘的重点主要是名词，包括人名、地名、组织机构、企业产品名称等专有名词、缩略简称等。

从“2+1”模式中“名+名”组合分析发现，783 组中使用频次为 1 的共有 563 组，其中有 359 组为合理组合，占 63. 77%。证明这个思路的可行性。统计结果详见表 9。

表 9. “名名 2+1”二元相关性组合低频样表

序号	词 1	词 2	频次	标记
01	爱心/n	车/n	1	√
02	爱心/n	卡/n	1	√
03	百灵/n	岛/n	1	√
04	板报/n	组/n	1	√
05	版权/n	局/n	1	√
06	帮工/n	队/n	1	√
07	包裹/n	库/n	1	√
08	保险/n	费/n	1	√
09	报告/n	表/n	1	√
10	暴风/n	雪/n	1	√

11	边城/n	村/n	1	√
12	边关/n	诗/n	1	√
13	边关/n	月/n	1	√
14	标杆/n	量/n	1	√
15	标志/n	旗/n	1	√
16	冰川/n	水/n	1	√
17	冰雪/n	路/n	1	√
18	冰雪/n	水/n	1	√
19	兵站/n	部/n	1	√
20	病房/n	门/n	1	√

“动兼名+名”2+1 模式二元相关性频次为 1 的 308 组低频组合中有 272 组为合理组合，占 88. 31%。例样详见表 10。

表 10. “动兼名 2+1”二元相关性组合低频样表

序号	词 1	词 2	频次	标记	11	烧结/vn	厂/n	1	√
01	斗牛/vn	舞/n	1	√	**12**	生活/vn	关/n	1	√
02	给水/vn	站/n	1	√	**13**	配电/vn	房/n	1	√
03	护理/vn	室/n	1	√	**14**	烹饪/vn	法/n	1	√
04	滑雪/vn	热/n	1	√	**15**	批评/vn	人/n	1	√
05	回收/vn	区/n	1	√	**16**	转移/vn	法/n	1	√
06	会晤/vn	站/n	1	√	**17**	着陆/vn	区/n	1	√
07	活动/vn	表/n	1	√	**18**	咨询/vn	室/n	1	√
08	摄影/vn	组/n	1	√	**19**	自律/vn	网/n	1	√
09	审计/vn	关/n	1	√	**20**	综合/vn	楼/n	1	√
10	审批/vn	权/n	1	√					

限于篇幅，谨将其他名词性为主的相关性统计结果列表如下作为参考。

表 11. 名词性二元相关性组合频数表

名词性相关性	频数
名+缀	233
组织机构+名	120
名+简称	114
区别词+名	68
形+单音名	56
地名+单音通名	34
处所+名	29
形+缀	22

五、余论

本文通过抽样语料库的二元相关性统计分析，探讨三字格结构的准词语知识挖掘的可行性。结果显示，频次因素并不是知识挖掘的唯一可靠指标。将词长相关性和词类相关性相结合，并把未登录词语的发现和识别锁定在“名词相关性”上，具有结构合理性和现实可行性。

本文写作过程参考了前辈时贤的有关论著，恕不具列大名，谨此一并致谢。

不当之处，敬请批评指正。

参考文献

[1] 盛玉麒．现代汉语词汇系统动态评估模型的理论与实践[A]．澳门科技大学“第二届海峡两岸现代汉语问题学术会议”，2006 年

[2] 盛玉麒．基于语料库的基础教育规范成语表的选词原则[A]．香港岭南大学“第三届海峡两岸现代汉语问题学术会议”，2007 年;《海峡两岸现代汉语研究》[C]，2009 年

[3] 盛玉麒．短语结构树形图练习题库的理论与实践[A]．韩国大田又松大学“第六届中文教学现代化国际会议”，2008 年

[4] 盛玉麒．基于语料库的汉语词汇知识挖掘研究[A]．台湾师范大学“第四届海峡两岸现代汉语问题学术研讨会，2009 年

[5] 盛玉麒．基于语料库的 HSK 知识挖掘研究[A]．温哥华“加拿大汉语教学国际会议”，2010 年