《中国语言生活状况报告》中成语与习语的调查与思考*

曾小兵[1] 张志平[1] 刘荣[1,2] 王丽娟[3] 胡竟伟[4]

1 北京语言大学应用语言学研究所 北京 100083　　2 太原理工大学文法学院外语系 太原 030012

3 太原理工大学计算机与软件学院 太原 030012　4 内蒙古河套大学数学与计算机科学系 临河 015000

E-mail：xiaobingzeng@126.com

摘要：成语与习语的调查作为《中国语言生活状况报告》在 2007 年的新增项目，它表明成语与习语使用情况引起了人们更多的关注。成语与习语的研究在语言应用中有广泛而深刻的意义。本文在基于大规模真实语料调查的基础之上，对成语与习语的使用情况做出了“单字差异”等比较，从中发现一些语言现象并提出了自己的思考，以期对汉语语言事实的发现、语言规律的总结、语言词汇的规范化等方面有所裨益。

关键词：中国语言生活，成语与习语，语言规律，词汇规范

The Investigation and Thinking about Chinese Idioms and Idiomatic Phrases in the Chinese Language Situation Report

Zeng Xiaobing

Institute of Applied Linguistics in Beijing Language and Culture University, Beijing 100083

E-mail：xiaobingzeng@126.com

Abstract: As the new item going to the Chinese Language Situation Report in 2007, the investigation of Chinese idioms and idiomatic phrases indicated that people pay more attention to the research of the Chinese idioms and idiomatic phrases. These researches have made an extensive and profound contribution to the applied linguistics. On the basis of the investigation based on the Large-Scale authentic corpora, this paper compares the “separate character’s difference” between Chinese Idioms and Idiomatic Phrases, and finds some language phenomena and puts forward some new ideas, then hopes to have some help in the discovery of Linguistic evidence, the summarization of linguistic rules, the Lexicon standardization and other aspects to some extent.

Keywords: Chinese Language Situation, Chinese Idioms and Idiomatic Phrases, linguistic rules, Lexicon Standardization

《中国语言生活状况报告》是教育部、国家语委在“加强语言文字应用、构建和谐语言生活”的理念下，为加强对现实语言生活的监测而进行的一项持续的调查与研究。自 2005 年的报告出版后，在社会各界引起了强烈反响，对于我们考察中国语言生活状况研究的新进展、进一步做好语言工作、做好语言规划、加强语言战略研究、推动语言状况的健康发展等方面都有深远的意义。对此，李宇明（2007）[1]、王铁琨（2007）[2]、戴庆厦（2007）[3]等老师都有充分的论述。

*本文在最后修改过程中经张普老师、杨尔弘老师阅览并提出宝贵的修订意见，在此叩谢。本次成语与习语调查是《中国语言生活状况报告（2007）》的一部分，杨尔弘老师的承担了《报纸、广播电视、网络（新闻）用字用语调查》总体的统计与分析工作，张志平博士和曾小兵承担了相应的分析工作。

2007 年，经“国家语言资源监测与研究中心”的“平面媒体”、“网络媒体”、“有声媒体”三个分中心联合采集数据并做出统计分析而编成的《中国语言绿皮书·中国语言生活状况报告（2007）》即将由商务印书馆出版，它在大规模的真实语料中进行动态调查与分析，充分而又真实可靠地反映了现代汉字和汉语词汇的使用现状。我有幸参加其《报纸、广播电视、网络（新闻）用字用语调查》（以下简称《调查》）的编写，从中获得一些体会和思考，愿与大家共同探讨。

1．语料的说明

为了便于大家了解，我们有必要对其数据来源与信息处理过程有个大致的说明。本报告的语料来源于“国家语言资源语料库”，它主要由 15 种报纸、9 家电视台、5 家广播电台等主流媒体[†]、以及新浪、腾讯 2 个门户网站 2007 年的全部新闻语料组成，共计 1 363 747 个文本，总字符数（不包括汉字部件、乱码以及无法显示的字符）1 236 120 162 次，总字种数 10 123 个。这些语料由分词软件[‡]切分后得到的字符串总数为 716 021 513 词次，其中词种数有 2 301 553 个。

软件标注为成语的词种数是 5 002 条，总次数 1 923 922 次；标注为习语的词种数为 4 959 条，总次数为 2 120 313 次。经人工校对，确定的成语为 3 892 条，习语为 5 467 条。[4]

由此看来，本次调查的语料规模是巨大的，且注重新闻语料的时效性、大众性、普遍性等特性。正是基于大规模真实文本的这些特征，我们对其的调查分析也就有了动态性，更加关注其在时间轴上的宏观与微观的变化。

2. 成语与习语的界定

2．1 对于成语与习语的认识

对于成语和习语的界定与区分，历来是语言学界争议的焦点，语言学家众说纷纭，尚无定论。

在《现代汉语（增订三版）》[5]中，成语被定义为：一种相沿习用具有书面语色彩的固定短语。而在商务印书馆 2002 年出版的《新华成语词典》前言中，成语的定义为：相沿习用的固定词组或短语，能独立表意，形式短小，一般为四字格式。[6]

对于习语，我们认为有广义与狭义之分，广义的习语即习用语，又叫熟语，是“人们常用的定型化了固定短语”，是“语言中定型的词组或句子。”（《辞海》中“熟语”词条）它包括成语、惯用语、格言和歇后语等；而狭义的习语则不包括成语，是以惯用语、歇后语、谚语等为主的短小定型的词汇单位。习语大多表现力强，在人们的语言生活中运用普遍，在一定程度上反映了人们对与自身生活密切相关的事物的精练概括，有极强的生成能力及丰富的象征意义，充分体现了人们的智慧，体现了人们对事物认识由具体到抽象的动态过程。本文讨论的即是狭义的习语，从而在与成语的比较中发现这种动态过程，以期印证或找到语言规律。

---

† 15 种报纸是（按音序排列）：《北京青年报》、《北京日报》、《北京晚报》、《法制日报》、《光明日报》、《广州日报》、《华西都市报》、《今晚报》、《南方周末》、《钱江晚报》、《人民日报》、《深圳特区报》、《羊城晚报》、《扬子晚报》、《中国青年报》；9 家电视台有：中央电视台、北京电视台、上海电视台、上海东方电视台、天津电视台、重庆电视台、广东电视台、山东电视台、新华电视台；5 家广播电台：中央人民广播电台、北京人民广播电台、天津人民广播电台、山东人民广播电台、深圳人民广播电台。

‡试验使用的是中国科学院自动化研究所的分词标注系统

2．2 在《调查》中采用的方法及其效果分析

我们在《调查》中对习语与成语的界定方法是：遵循“成语从严、习语从宽”的原则，在很大程度上忠实于计算机的标注结果，但在成语部分进行了人工校对，参照商务印书馆 2002 年出版的《新华成语词典》，将该词典中未收录而被软件标为成语的词语全部放到了习语中。

值得注意的是，我们进行的划分与统计并非完全没有错误，受语言信息处理技术和人的知识结构等方面因素的限制，现有分词软件及成语习语的统计方法，其小范围内的错误是在所难免的。如：对于中科院自动化所软件分词或切错词，或是标注错的现象，苏新春、杨尔弘老师在《2005 年度汉语词汇统计的分析与思考》[7]一文中对其做了具体分析，得出结论是：（这种）讹误大多出现在低频范围，对整体数据性质的影响微乎其微。再如：“成语从严，习语从宽”的原则是为了在操作层面便于划分两者而采用的，在实际中可能会有些百密一疏的现象，如：成语“挂羊头卖狗肉”没有归入成语之列而放在习语之中。但在整体的数据分析与统计中，《调查》是符合语言学的原理和实际的。

3．成语与习语的比较及其意义

对其进行数据的统计与分析后，我们得出的成语与习语的差异，在很大程度上是反映了语言稳态部分中的高稳态及相对稳态的区别，是语言稳态的历时性在一定层面上的体现。其主要的意义在于：

3.1 探求语言发展变化的轨迹

语言的发展和变化集中而又迅速地反映在词汇方面。新词的产生、旧词的逐步变化与消失，是语言变化发展的突出外在显现。张普老师则将语言分为“稳态”与“动态”，“（语言的）变化的端倪就隐藏在大规模的真实文本（无论他们是经典的还是非经典的文本）之中，甚至就隐藏在那些非规范现象里。[8]而我们的主要目的是找到这种变化的端倪，即使是不规范不准确的表达，也是我们发现问题、分析问题与解决问题的突破口。

3.2 从中总结发现语言学规律，印证已有的语言学理论

许多语言现象是具随机性与偶然性的，汉语文字具多元性与灵活性。如何在众多的语言现象中，总结与发现一般的语言规律、原则，进而将它们进行总结提炼，进而上升到语言理论的层面，再将它应用于更多语言事实的揭示中，这是我们应用语言学需要解决的突出问题。陆俭明、胡明扬等老师在 2008 年 4 月北京市语言学会第八届学术年会上对此都有着重的强调。因此，就要求“我们的统计分析必须进一步向动态跟踪、检测、监测语言（首先并且主要是词语）的发展变化方向深化。”[9]值得庆幸，我们拥有反映每年语言“实态”的语言资源，依此进行对比统计分析，可以印证语言现象的产生与发展，同时看到更多的语言事实，从而发现更多的新问题。

3.3 可以加强对词汇的规范化研究

在成语与习语的比较中，我们把成语作为一种固定范式，但在与成语密切相关的习语的考察中，我们可以发现其中许多的不规范用法，如一些生造词：休闲养性、开天劈地、心往神驰、井

井有序等，不管是由于书写错误、求新求异心理还是其他，这些不规范的用字用词，对于我们词汇与语言的规范有一定的参考价值。李宇明老师指出，词汇规范的难度很大，原因不仅在于已有的词汇本身就相当复杂，生殖又极快，而且学术界对于词汇的规范规律至今缺乏足够的认识。[10] 戴昭铭老师也指出，对于语言的规范，尤其是细节上的规范，我们还有很多的工作要做。[11]

3.4 为语言的预测提供参考的依据

语言修辞学家王希杰老师在20世纪80年代提出了显语言和隐语言共同构成语言全貌的语言预测观。[12]即可以通过现有的显语言成分，来预测未来语言的变化发展，如“女保姆”一词出现后会不会有“男保姆”出现。对此，在本文的统计结果中，可以看到很多习语都由成语的近义或反义而得来，如：问心无愧到问心有愧。但这些新的用法，能够在多大程度上被人们所接受、是否真正符合语言学的理据，正是我们要在调查分析中量化的。“（语言发展）不时形成一些热点。这些热点，有些需要通过积极引导，促其升温；有些则需要及时妥善地加以处理。”[13]

4. 实验方法及结果分析

4.1 理论基础

首先，语言知识的动态更新理念。我们认为，语言的发展是在稳态的基础上产生相对的变化，而这种变化在一定时期内被人们“约定俗成、逐渐规范”，它最终也进入到稳态的部分。这种稳态——相对稳态（稳态下的变化）——动态——相对稳态（动态下的约定）——稳态的螺旋上升的动态更新机制是语言的发展规律。由于成语与习语都属于一般词汇的范畴，而成语的稳定度高于习语，我们将成语定为高稳态的部分，但相对而言，习语的变化周期更快。

其次，词语具有自我生产性。这种生成性包括两个主要方面：从词汇及句法上讲，语言自身的结构有生产性，即词汇自己生成一些新的要素，包含新词、新语、新义等。如：从成语“心腹之患”生成“心腹大患”、由“如醉如痴”到“如醉如狂”、“如醉如梦”、“如痴如狂”都在结构上反映了词语的生产性；从语用及语境上讲，语言在交际使用过程中保持着一种自我调节的动态平衡。由“任人唯亲”、“任人唯贤”到“任人唯钱”，由“开门见山”到“开门见喜”，更大程度上是反映了社会生活及文化的变化。

最后，对语言现象上升为语言理论需要时间与实践的检验。语言系统是多层次多级别、内部结构富于矛盾而又对立统一的组合。因此我们要在遵循语言发展规律基础上来考虑其运动变化，对其结果要区分对待，不能一味地求新求异而对一些错误、怪异的词语及用法坐视不管。

4.2 数据分析：

在语料的成词与习语词表中，我们查找出相差一字或顺序不同的词，将它们列为一组，如：自食其果、自食其力、自食其言；坐视不管、坐视不理就是我们找出的两组。这些组由2-8个成语或习语组成，其中由2个词语组成的居多，有331组，占总数的89.0%。通过对其意义与使用频率的考察，从定量分析的层面对词语的变化发展情况做探讨。

4.2.1 这些组有的都由成语组成，如：温文尔雅、温文儒雅，有的都由习语组成，如：软磨硬缠、软磨硬泡，有的由成语和习语一起组成，如：全盘皆输、全盘皆活。它们的分布如表1：

组成部分	全是成语	全是习语	成语与习语
组数	79	130	163
所占比例（%）	21.24	34.95	43.82

表 1：组成部分的总体分布表

从中看出，这种动态大部分都存在于“全是习语”、“成语与习语”之中，而在“全是成语”中出现的多是同义词，它们可以在使用中进行替换而不影响其意义的表达。如精疲力尽、精疲力竭等。这也说明，成语和习语中，习语是相对更加活跃的部分，且其中的一些成分，是从成语中借鉴而来的，这部分集中反映在表中“成语与习语”的部分，占整体的比例为 43.82%。

4．2．2 从组内关系看，将 372 组进行观察比较，各组关系主要分为近义、反义、稳定结构、字序交换。另外还有多字词被分词软件切分为两个单元或三个单元的，以下表 2 是其比率及举例：

	切分词	字序交换	近义	反义	稳定结构
组数	15	8	315	15	19
比例（%）	4.03	2.15	84.68	4.03	5.11
举例	智者见智 仁者见仁	斗转星移 星移斗转	走马看花 走马观花	心中有数 心中无数	自食其果 自食其力 自食其言

表 2：各组内容不同关系的比例及其样例

上述表明，在组内关系中，主要以近义生成为主，而这些近义的生成，有些是合理而有效的，有些是偶发而又无所理据的。对这些具体的细节方面，还缺乏规范的标准。“结构紧密、使用稳定”是可以对之进行衡量的两个主要方面。

对于切分开来的词，我们可以进一步考虑其结合的紧密性。在这些多字词的成语习语，多产生于历史典故或者口头语，说明其在一定的历史时期是结合紧密的，而且既然能成为一般词汇甚至基本词汇，说明其在当时是使用稳定的。但是，从另一方面看，这些词开始在很多时候分开说，或者是只说其中的一部分而表达整体的意思。如仁者见仁的频次比智者见智多 38 次，也就是说仁者见仁独用了 38 次。但这里的前提是，我们对那些切分开来的部分要做细致的考察，如百尺竿头更进一步中的“更进一步”在独用的时候，更多是不表达这个短语的整体意思，而是作为一个副词短语，这种情况我们不将它作为此次讨论的内容。其组内间的频次差数见下表 3：

组内词	频次	差数	组内词	频次	差数	组内词	频次	差数
种豆得豆	58	1	有则改之	92	4	此一时	55	5
种瓜得瓜	59		无则加勉	88		彼一时	50	
智者见智	310	38	你一言	65	14	焉知非福	113	45
仁者见仁	348		我一语	51		塞翁失马	158	
知无不言	134	23	听其言	55	16	老骥伏枥	112	48
言无不尽	111		观其行	71		志在千里	64	
言必信	62	33	失之东隅	68	3	百尺竿头	110	2446
行必果	95		收之桑榆	65		更进一步	2556	
千里之行	52	2	深一脚	30	10	千叮咛	9	2
始于足下	54		浅一脚	20		万嘱咐	11	

表 3：切分开来的多字词的频次统计

从中看出，我们在使用过程中，经常用多字成语或习语的部分代表其整体。就其本身来讲，“能指”是简洁明了的，“所指”却是丰富有韵味的，这在显现层面上反映了时代的快捷与迅速发展要求人们在交际中删繁就简。而在隐性层面来讲，正是由于其结合的紧密性决定了其在使用中可以不使用结合体，因为人们一接触“塞翁失马”就知道其要表达“焉知非福”的意义。

4．2．3 纠错性与规范化

在习语的使用中，有些是由成语改一字或结构变幻生成的、使用率很低的，这是我们要规范与关注的，对于这些“端倪”，有必要考察其是否具有更深刻的意义，或者只是误用。在 372 组中，有 125 组中习语的频次少于 10，从中我们选出了频次为 1 的习语，共 34 个。（见下表 4）其中有些是杂糅的，如井井有序可能很大程度上是来源来“井井有条”和“井然有序”的组合；有些是求新求异的，如来源于“晓之以理”的“晓之以利”等。对这些加以规范和纠错，甚至是防范于未然，都是必要的。

坐立不定	群贤毕集	长歌当啸	浩瀚无涯	晓之以利	脱口道出	如醉如梦
坐立不稳	拳拳之忱	从善如登	孤苦伶丁	物极而反	随请随到	百业待举
自觉自醒	情真意长	错落无致	博极群书	贪赃卖法	开天劈地	井井有序
壮志得酬	情深意绵	东奔西窜	退耕还田	千古绝响	休闲养性	竭泽而鱼
只言片字	千难万劫	推三托四	小巫见了大巫	学而不倦	心往神驰	

表 4：各组中频次为 1 的习语汇总

另外，我们发现在反义内容关系的组内，习语的频次普遍低，见下表 5。（表内黑体词为成语，否则为习语，下文同）这也说明了一些生造的习语：如：有隙可乘、从善如登、错落无致等，只是出于人们的求异求新心理，这些词只是在极小的范围内使用。

以上的两种都是我们规范成语习语使用情况的重点，当然，成语习语的规范化远不止这些，这只是一个侧面，一个可以增强我们语言规范化敏感性的方面。

成语习语	频次	成语习语	频次	成语习语	频次	成语习语	频次
壮志得酬	1	力所不及	35	从善如登	1	**曲高和寡**	315
壮志未酬	104	**力所能及**	2027	**从善如流**	123	曲高和众	17
心中无数	54	**后继无人**	253	错落无致	1	**榜上无名**	176
心中有数	1081	**后继有人**	415	**错落有致**	831	榜上有名	2239
无隙可乘	21	**供不应求**	5374	不战而胜	338	违法必究	192
有隙可乘	2	供大于求	1256	不战自败	4	违法不究	106
问心无愧	438	**白璧微瑕**	23	全盘皆活	21		
问心有愧	12	**白璧无瑕**	12	**全盘皆输**	65		

表 5：反义关系的成语与习语频次比较

4．2．4 稳定性和多元化

前面说过，成语作为高稳定的词语，稳定度比习语大，但在其使用过程却出现频次低于习语情况，见表 6。这是否能说明这些习语在与成语的竞争中呈现出一种强势。由于根元、王铁琨、孙述学老师执笔的《新词语规范基本原则》[14]认为：“新词规范的标准是交际值，是交际到位的程度。”这些习语在交际的过程中形成了新的更强的稳态。在特征上更加口语化，更加通俗化。

成语习语	频次	成语习语	频次	成语习语	频次	成语习语	频次
百年难遇	115	**不远千里**	346	**烟波浩渺**	84	**默默无言**	27
百年一遇	567	不远万里	433	烟波浩淼	91	默默无语	94
百废待举	19	**合而为一**	183	**威武不屈**	29	**初露锋芒**	172
百废待兴	217	合二为一	937	威武不能屈	58	初露端倪	362
榜上无名	176	**满腔热忱**	122	**除旧布新**	37		
榜上有名	2239	满腔热情	393	除旧迎新	55		

表 6：习语的使用频次高于成语情况汇总表

即使同是成语，在使用上也是有差别的，如下表 7。而这些现象的出现，应该与人的心理认知、书面语表达、音律和谐等方面有更大的关系。

成语组	频次	成语组	频次	成语组	频次
博闻强记	**37**	**火上加油**	**96**	**精疲力竭**	**485**
博闻强识	**6**	**火上浇油**	**673**	**精疲力尽**	**232**

表 7：同是成语的频次使用情况比较

5. 总结

《中国语言生活状况报告》作为国家构建和谐语言生活的调查研究项目，其对成语和习语情况的关注与研究，有利助于我们在更多层面上对语言的真实使用情况进行比较与分析。

综上所述，目前为止，我们对成语和习语的探讨只是限于上述几方面，在成语与习语的比较中发现了一些问题与规律。最后，希望在更广泛更深入的层面与大家进行学习讨论，能力所限，文中不妥之处还请大家指教。

参考文献

[1]李宇明. 关于《中国语言生活绿皮书》. 语言文字应用. 2007 年第 1 期. 12～19

[2]王铁琨. 语言使用实态考察研究与语言规划. 语言文字应用. 2008 年第 1 期. 15～24

[3]戴庆厦. 中国语言生活状况研究的新篇章. 语言文字应用. 2007 年第 1 期. 25～28

[4]王铁琨主编. 中国语言生活绿皮书•中国语言生活状况报告(2007). 北京:商务印书馆. 2007 年 (待出)

[5]黄伯荣、廖序东主编. 现代汉语（增订三版）(上册). 北京:高等教育出版社. 2002 年 7 月版. 317

[6]商务印书馆辞书研究中心编. 新华成语词典. 商务印书馆. 2007. 前言部分

[7]苏新春、杨尔弘. 2005 年度汉语词汇统计的分析与思考. 厦门大学学报（社科版). 2006 年第 6 期. 86～89

[8]张普. 论（语言的）稳态. 缩简文本载《郑州大学学报（哲学社会科学版)》2008 年第 2 期，总第 194 期

[9]张普. 当前字、词、语量化研究的五个深化方向. 2005 年 12 月报告于中国台北举办的“第三届两岸四地中文数位化合作論壇（CDF)”，并收入会议论文集。见张普动态语言知识更新研究，待出。

[10]李宇明. 词汇规范的若干思考. 厦门大学学报(哲学社会科学版). 2002 年第 2 期. 19～23

[11]戴昭铭. 信息时代的语文规范化问题. 求是学刊. 1994 年第 4 期. 97～101

[12]王希杰. 汉语的规范化问题和语言的自我调节能力. 池州师专学报. 1995 年第 3 期. 9～15

[13]王铁琨. 计算机统计数据与年度语言生活状况报告. 长江学术. 2007 年第 1 期. 2～3

[14]于根元, 王铁琨, 孙述学. 新词语规范基本原则. 语言文字应用. 2003 年 2 月第 1 期. 89～95