

硕士研究生学位论文

**题目：构式视角下的现代汉语四字表达的能产性研究**

**姓　名**：　　　田　骏

**学　号**：　　1301211280

**院　系**：　中国语言文学系

**专　业**：　 汉语言文字学

**研究方向**：　 中文信息处理

**指导教师**：　 詹卫东　教授

二〇一六 年 六 月

**版权声明**

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

A Productive Research on Modern Chinese Four-Word Expression: from a Constructional Perspective

Tian Jun (Chinese Linguistics)

Directed by Professor Zhan Weidong

# ABSTRACT

KEY WORDS:

# 目　录

摘要 I

ABSTRACT II

目　录 III

第一章　引言 1

1.1 问题的提出 1

1.2 选题背景及意义 1

1.3 文献综述 2

1.3.1 不同的术语体系 2

1.3.2 前人研究的阶段性特点 5

1.3.3 前人研究的不足 6

1.4研究方法 7

1.5论文结构安排 7

第二章　现代汉语四字表达概述 8

2.1 四字表达的定义 8

2.2 四字表达的语法地位 8

2.2.1 传统定义的不足 8

2.2.2 构式语法概述 9

2.2.3构式视角下四字表达的语法地位 12

2.3 四字表达的规模 13

2.4 四字表达的能产方式 14

第三章　词库分析 15

3.1 大词库中的格式 15

3.1.1 总体情况 15

3.1.2 前人未收录格式举例及分析 16

3.1.3 有关多继承性的进一步讨论 21

3.1.4 小结 22

3.2 大词库中的变换 23

第四章　语料库分析 24

3.1不经意的错误：误用能产性 24

3.2 巧妙的引用：仿造能产性 24

3.3 言之不足，故复言之：对举能产性 24

3.4 韵律之美：紧缩能产性 24

第五章　相关构式的进一步分析 25

第六章　结语 26

参考文献 27

致谢 28

北京大学学位论文原创性声明和使用授权说明 29

# 第一章　引言

## 1.1 问题的提出

四字表达在汉语中占据着重要地位。粗看下，四字表达可分为两种：成语和常规短语。“人穷志短”是成语，需要收进词典；“中国制造”则用一个偏正式短语规则就能很好地处理。但仔细考察，则会遇到一些困难。例如：“人怂志短”[[1]](#footnote-1)是成语词，还是自由短语呢？

如果看作自由短语，组成这个短语的成分却并不都是词，如“志”一般不能单说；我们也会感觉到这些成分的组合并不是完全自由的，成分间结合的紧密程度也不一样。而如果看作词，这样的词“临时性”很强，不是那种会进入我们的“语言素材库”中、被我们频繁运用的词。

类似的表达数量非常之多，因为它们的表达效果非常强，甚至盖过了它们作为信息载体、传递信息的作用。随着个性的解放和互联网的发展，这种表达非常符合“网络虚拟社区”的交际需求：为了避免常规表述的乏味，引起听话者的兴趣，人们乐意创造并使用它们。

《现代汉语词典》，及一些前人的研究，已经通过 “格式”的概念，如“千…万…”，来规整这些表达，并启发人们创造。但我们认为，“套格式”[[2]](#footnote-2)并不能完全诠释四字表达能产的本质，例如对“悲极而泣”的解释力就极其有限。前人也有意无意提到了其他的一些能产机制，但都是作为补充，而未进行充分考察和量化。这就给我们带来的下述值得研究的问题：

四字表达在今天呈现出繁荣的态势，它到底有哪些能产机制？这些机制的能产程度又是怎样的？

## 1.2 选题背景及意义

未登录词识别是中文信息处理的重要一环，它能够提升自动分词的准确率，从而改善基于分词的任务的运行结果。对于“中国制造”类四字表达，可以很好地被分词程序处理为“中国/n 制造/v”；但对于“人怂志短”而言，很可能得到“人/n 怂/a 志/n 短/a”甚至更离谱的结果[[3]](#footnote-3)。我们的工作能够帮助机器识别一定数目的四字表达，因而能提升相关中文信息处理任务的性能。

（构式语法理论近年来受到越来越多的关注。作为认知学派的理论，它更加鼓励人们去关注语言使用上的细节，比如哪些话更常说，哪些话不常说。……）我们的工作也是一次构式语法理论在汉语中的实践，我们希望借助四字表达来阐述我们的构式观，以起到抛砖引玉的效果。

## 1.3 文献综述

有关四字表达的研究很丰富，光术语体系就分好几套：以（陆志韦，1956）为开端的“四字格”表述，以（文炼，1988）为开端的“类固定格式”表述，以（张国宪，1993）为开端的“对举格式”表述[[4]](#footnote-4)，以（周荐，1999）为开端的“待嵌格式”表述等等。此外，还有一些学者的相关术语也为后人借鉴，如（张拱贵，1985）提出的“语汇格式”，（李宇明，1999）提出的“词语模”，（邵敬敏，2008）提出的“框式结构”等。

从发表时间、研究角度和研究方法上进行比较，这些研究呈现出从定性到定量、从单一的结构主义理论到多元理论的阶段性特点，显示出学者对问题的认识逐渐深入、具体化。在1.3.1小节中，我们首先按术语体系对前人研究进行综述；在1.3.2小节中，我们再依阶段对前人研究进行总结；在1.3.3小节中，我们最后指出前人研究的不足之处。

1.3.1 不同的术语体系

“四字格”表述

（陆志韦，1956）首称所有完整的有意义的四字组合为“四字格”，并把“并立四字格”定义为内部紧凑、形式上为“2+2”并立的四字格。文章主要说明“并立四字格是词”的问题，理由是：并立四字格形式上不可拆分单说，内部结构的句法性减弱，不宜再在词性或句法结构上较真。对于一些半自由格式，作者则认为应特别小心，因为这些格式一旦能产，产出的实例的并立意味难免打了折扣，因为它们并非出于并立的构词方式被构造，而仅仅是一种套格式的行为。换言之，作者认为并立和套格式是两种能产方式。另外，作者还注意到一些并立的四字格的内部汉字顺序可以变换，譬如“开天辟地”能变换为“开辟天地”，从动宾式并立格变为动宾短语，但作者认为后者并非并立格，并未深究下去。

（姚殿芳，潘兆明，1985）采纳（吕叔湘，1963）的定义，把四言短语也纳入四字格范围。作者主要从汉语教学目的出发，分“认识四字格-创造四字格-恰当运用四字格”三步进行介绍，指出四字格生命力强大。作者主要提到两种创造四字格的方式：衬垫虚字和骈立，骈立又分两种，一种是生造，多半得利用古汉语单音节词；一种是套用现成的格式。作者认为，套用格式这种派生方式“体现了能产型骈立格的灵活性”。

（马国凡，1987）也提到了某些四字格的形成途径，如词的AXAB式扩展、词组的紧缩与节缩等，其中最具启发意义的是他称之为的成语的衍化格式。他认为成语的衍化有非固定的和框架格式衍化，前者如从“人云亦云”衍化出“‘帮’云亦云”（指帮派、派性）、“人云不云”（不随便附和）、“人倒亦倒”的说法，后者主要通过向一些“A＿B＿”式的框架中填入临时需要的成分构成。

“类固定短语”表述

（文炼，1988）把类固定短语定义为格式和功能上像成语、却是临时创造的、且含义可根据字面推断的汉语语言形式。作者指出，类固定短语不像固定短语那样能够作为语言材料供遣词造句，但其“组词成句”的规则又是很特别的，掌握这种格式对理解汉语是有帮助的。他具体举出四个包含单一固定汉字的类固定四字格式，并区分了形成该格式的成语和由该格式产生的类固定短语，并没有追求格式本身的意义，较好地诠释了“模仿”能产性的含义。作者还观察到一类特殊的现象，如“水落油出”[[5]](#footnote-5)，他称之为“仿造”，并认为这种仿造的成语必须依附被仿造的成语，才能显示出它的修辞效果。

（齐沪扬，2001）进一步强调类固定短语“结构模式固定”和“意义单一”两个特点，其中结构模式固定是指“类固定短语是是替换了成语中的一些组成成分后构成的”。作者根据替换成分的多少分成了“替换两个成分”和“替换三个成分”两类，前一类是（文炼，1988）没有涉及的部分，其中又进一步分出“保留数词”、“保留相同语素”和“保留同义或反义语素”三类。这种命名方式也体现出与（文炼，1988）一致的“成语-格式-类固定短语”能产方式。

（李传军，2008）同样出于辅助教学的目的，更系统地考察了类固定短语。他沿用齐沪扬的按替换成分分类方式，把类固定短语分为偶位内嵌型（A…C…）、奇位内嵌型（…B…D）、首尾内嵌型（…BC…）、一字支撑型（……C…，…B……）和组块内嵌型（其他型式）。他采用定量方法，“例不十，不立格”，提炼出300个格式，并统计了每个格式的不同实例[[6]](#footnote-6)数目，形成了“型式-格式-实例”三层计量体系，然而，他并未区分出成语，而把成语也看作一种类固定短语。此外，他还注意到同一格式下的类固定短语存在变换现象，例如“越生越穷”和“越穷越生”，并根据二者频率的差异进一步分出自由变换和有限变换。

“待嵌格式”表述

（周荐，1999）观察到《现代汉语词典》中存在这样一类特殊的词条，像“不…不…”、“千…万…”，他在该文的附注中首次采用“待嵌格式”来命名这一类词条，而《现代汉语词典》本身并未对这一类词条做出更多的解释。

（周荐，2001）认为，之所以设定此类条目，是因为其在口语中出现频繁，却不足以设定条目，以“待嵌格式”以纲带目，目的是引导人们用某种架构去类推创造出新的固定语。他认为，四字格是待嵌格式最适宜选择的格式，而且《现代汉语词典》中绝大多数格式都是“一三字显，二四字隐”。他认为，“一三字隐，二四字显”的格式同样存在，《现代汉语词典》中没有收录，应该是无法为之在音序序列中找到一个合适的位置的缘故，属于技术处理上的困难，他将这两类待嵌格式统称为交替显现[[7]](#footnote-7)待嵌格式。此外，他还举出《现代汉语词典》中收录的“自相……”一例，说明这类邻接显现的格式远不如交替显现的格式的能产性强。

（范喜梅，2009）重点分析了由成对方位词构成的待嵌格式。她认为，这些待嵌格式内部结构虽大多并列，却要遵循一种约定俗成的语序，如“上”在“下”前，“左”在“右”前。作者还指出，在并列关系之外，还存在转折、因果、目的等其他关系。此外，作者还为方位待嵌格式拟出了6个释义模板。

（孟祥英，2010）从汉语教学和词典编纂的角度出发，对待嵌格式进行了系统研究。她整理出240个固定字与填充字交替出现的待嵌格式，并根据固定字的词性组合、意义异同分为多个小类。作者指出，固定部分多属于基本词汇范畴，且多具丰厚的文化蕴涵（如“三”表多）。嵌入部分则与固定部分的词性、词义相关，有时候可以嵌入多种词性。作者把格式产生的语言机制归为“对称”、“类推”、“经济”三方面，为我们叙述能产性提供了很好的参考作用。

其他术语体系下的研究

（张拱贵，1985）中首次提出了语汇格式的概念。这篇文章是作者为朱林清《现代汉语“格式”初探》作的序，他指出，朱作中分析的格式都是语法格式，这些格式包含固定词语，一般充作关联词组构成短语，且使用上带有某种习惯性，不能用一般语法规则解释。他认为，朱林清对格式的研究是启发性的，但“‘格式’不仅语法里有，语汇里（成语、惯用语、谚语等）更多”。他举出自己于1963年发表的《成语中的成套格式》一文，说明此类格式在成语中的大量存在。

（冯胜利，1997）从韵律角度讨论了四字格式的成因，他提出了格套化理论，认为某些四字格式的形成源于双音节词的格套化，即将双音节词拆开，中间填入某种具有音缀性质的填充词而成。他认为，双音节词以[[＿A][＿B]]形式拆分居多，而填充词用久用熟以后，便能形成[X＿Y＿]式的格套，从而产生更多的词。

（李宇明，1999）提出了词语模的概念：具有批量产生新词语的框架，其中“模标”指不变的成分，而“模槽”指空位，例如“X盲”中，“盲”是模标，“X”是模槽。他还强调，词语模不仅仅是一种创造词语的模子，也是一种思维模式，模标具有一种“类”功能，例如人们习惯于把有目的、有计划、影响面广、意义重大的工作称为“X工程”，“X工程”便成了一个思维的模式。而他认为，词语模主要分前空型和后空型，中空型的词语模很少，如“吃X饭”。

（黄燕璇，2011）沿用（邵敬敏，2008）提出的框式结构概念分析了方位对举四字结构的特点，以便帮助人们成类地掌握这种结构，“减轻逐个记忆的负担”。她认为，这类四字结构中的方位词大多已虚化，虚化后才能形成“量多”这个结构义。此外，作者还提到成语的凝固性分两种，一种是来自典故的、一开始就凝固的结构，另一种则是经过繁衍、筛选、淘汰后形成的优势结构。因此诸如“东张西望”这样的凝固成语，从历史的角度也属于框式结构。

（韩立秋，2015）用构式语法分析了古汉语的凝固结构，提出了许多前人不曾涉及到的凝固结构，如“不亦…乎”、“何其……”等，作者把这些称为半实体性凝固结构。此外，作者引入“动态浮现语法”，强调语言一直处于一个动态过程中，“一旦临时结构所反映的认知经验得到了认同，并能够引起人们再次使用的兴趣时，凝固结构就很有可能在重复使用中发生变化”。他还提及“重新分析”机制，指出，“随着使用频率的提高，各个成分的本源义就会被遗忘，人们对其内部形式就可能做出重新分析”，我们认为这是非常重要的能产机制。

1.3.2 前人研究的阶段性特点

从定性到定量

“定量分析”是近年来的大趋势，也是青年学者的优势所在。2000年之后，相关的研究基本都带有“语料库”特色，且无一例外地选用了北京大学中国语言学研究中心（CCL）在线语料库，这可能与该语料库提供丰富的查询手段、能很好地查找结构不连续的语法和语汇格式有关。

在定量分析上，李传军的研究最为立体。他提出了“型式的能产性与格式的能产性并不是简单的正相关关系”观点，从“型式-格式-实例”三层对涉及的类固定短语进行了统计，清晰地描述了这类短语的总体分布情况。由于语料的大量性，学者们往往能更仔细、更全面地考察格式特点，也使得文中论点更具有说服力，使读者能更好地掌握格式用法。

从单一的结构主义理论到多元理论

语言理论的发展是另一个不可忽视的因素。较早的时候，由于对结构主义语言学还处于学习、探索的阶段，这一时期的文章多从句法层面立意，如陆志韦主要讨论了并立四字格是不是构词格，而文炼则旨在说明四字表达具有特别的短语性质。

而随着国外先进理论的引入，更多的青年学者开始从寻求更多途径来解释能产性。如孟祥英解释“类推”机制时提到了青年语法学派，韩立秋则引用Langacker定义的重新分析解释句法结构的演变等。多元理论的引入使得观察角度更加多样，使研究的对象变得更加立体，同样增加了结论的说服力。

总而言之，随着研究的深入，人们对四字表达逐渐建立起一些共识，如语法性质更偏向词、语义透明而表现力丰富等，人们开始用更精细的手段去研究它的特点。

1.3.3 前人研究的不足

以上阶段性特点，既是优点，也有不足。

在定量分析上，学者们只把精力放在“套格式”上，而仍用“仿造”、“类推”等术语定性地看待“人怂志短”这种能产方式。另一方面，所有的青年学者都把一个格式的不同实例数量等价于该格式的能产性，而忽略了实例在语料中的频率分布，而后者才是真正的能产性标准。再者，单一的语料来源也使得他们无法关注到更多的语言现象。可以说，前人的定量分析结论是很有限的，有很大的改进空间。

另一方面，前人对构式语法理论存在一定的认识误区。如韩立秋谈及构式时说，“因此，我们认为对于这样一种既非词亦非短语更非句子的特殊结构，最好的解释方法莫过于‘构式’”。孟祥英对构式同样存在类似理解：“无法用常规语法规则和从构成成分的语义来推知其意义”。把构式等价于“特殊的语法规则”，仍然是结构主义的分析模式，而没有体现出构式的认知基础。另外，许多学者都跟孟祥英一样，为了给构式一个独立的构式义，常常把词素甚至语境带来的意义“捆绑销售”，造成不必要的复杂和混乱，这也是没用很好地体会构式概念而产生的盲从做法。

我们认为，前人的不足一方面在于对能产性的认识不足，只考虑了一种能产方式。另一方面，在构式语法的运用上，他们大多只把构式看作一种“带有意义的框架”，本质上还是结构主义的思维，无法很好地说明能产性的本质。

## 1.4研究方法

本文将秉持理论与实际相结合的研究方法，采用定量的分析手段，从经典语料、网络语料两方面入手，探讨四字表达的能产性。

本文借助构式语法理论基于使用的分析方法，动态地看待语言现象，使其更接近心理表征。

本文使用的语料包括：

CCL古代/现代汉语语料库，其在线版本为<http://ccl.pku.edu.cn:8080/ccl_corpus>；

搜狗细胞词库的成语部分，主页为<http://pinyin.sogou.com/dict>；

盘古分词词典，下载地址为<http://pangusegment.codeplex.com/releases/view/31531>；

NLPIR共享平台提供的500万条微博语料及23万条微博内容语料库，下载地址为<http://www.nlpir.org/?action-category-catid-28>。

## 1.5论文结构安排

第二章中，我们首先对四字表达进行定义（2.1节），并讨论其语法地位（2.2节）。讨论中，我们将指出传统分析的不足（2.2.1小节），并概述构式语法的思想（2.2.2小节），然后给出构式视角下的地位（2.2.3小节）。随后，我们将简要分析四字表达的规模（2.3节）和能产方式（2.4节）。

# 第二章　现代汉语四字表达概述

## 2.1 四字表达的定义

本文把满足以下两个条件的汉语言形式称为四字表达：

a. 由四个汉字组成；

b. 意义完整。

a条件说的是形式特征。在“四音节”与“四字”的表述上我们最终选择了后者，因为我们认为以“字”而非“音节”为单位的处理方式更加简明中立。一方面，“奥运健儿”、“鬼魅伎俩”中，“儿”和“俩”都是单音节词素，如果以音节为单位，在中文信息处理中势必要引入消歧工作；另一方面，对于许多南方人和外国人来说，儿化的“儿”和两个的“俩”就是一个音节，在以文字为载体的语言学习和融合的过程中，我们应该宽容一些，客观看待不同人群对这一现象的接受程度。

b条件说的是意义特征。这一条件主要是排除一些“跨意义”的片段，如“一个中年”，显然后面还应该有个中心语，如“男人”，且“中年”跟“男人”结合得更紧密，不能把“男人”单独拆分出去。我们认为，只要满足“内部紧密程度大于外部紧密程度”，不妨都看作四字表达，如《史记》开篇：“黄帝者，少典之子，姓公孙，名曰轩辕。生而神灵，弱而能言，幼而徇齐，长而敦敏，成而聪明。”今天一看，很容易就从这些下划线标示的四字表达中抽象出“X而XX”格式。将范围扩充得大一点，有助于人们积累经验，形成概念，乃至归成一类。

在这两点之上，可能还有一点：

c. 形式紧凑

这一点之所以没有当作标准，是因为“紧凑”是一种经验，并无统一度量，实际上很难取舍。有人可能会介意将典型的由两个双音节词组成的四字短语与成语、类固定短语、待嵌格式等概念相提并论，但面对“智能手机”、“六角扳手”等专有名词时，往往又难统一看法。尤其对于中文信息处理而言，我们认为把四字短语归入四字表达要比排除出去简明清晰。而对于本文分析的能产性而言，这些短语的结构、功能前人也已分析得很完善了，我们可以跳过它们，专注于一些前人不那么关注的能产方式。

## 2.2 四字表达的语法地位

2.2.1 传统定义的不足

非要给四字表达一个传统意义上的语法地位是困难的。

首先，许多四字表达都能看作短语，如“心情舒畅”；其次，也有许多四字表达不能看作传统意义上的短语，如“人穷志短”；此外，还有很多四字表达，有人认为是短语，有人认为是词，如“资本主义”，有人认为“主义”不能单说，也有人认为“主义”内部紧凑、外部自由、意义明确，可以看作新词；又如“你弹我唱”如果是实指，表达“你弹琴，我唱歌”，大家就倾向于归为短语；但如果是虚指，表达“有人弹琴，有人唱歌”，大家就倾向于归为词；即便同为实指，“飞禽走兽”像是词，而“家禽野畜”则又像是短语了。

我们认为，以上的看法都有道理。不能取得一致，是因为传统的词和短语的定义是静态的。“有人认为”就表明，对于大多数表达，人与人的感觉是不一致的。文炼、黄燕旋、韩立秋的文章中都提到了成语界限的动态性和模糊性，说明学者们已经开始注意到这种“一刀切”的分析方式的弊端。而学者们也提出了多种解决方式，如马国凡等许多学者主张“忽略其语法地位，而强调其功能特点”，更多的学者则倾向于用一种特殊的语法单位来弥补这种缺陷，比如“待嵌格式”，或者韩立秋所谓的“构式”。

2.2.2 构式语法概述

基于使用的分析方法

我们先举两个前人论述中没有明确与构式挂钩、但比较接近构式的观点。文炼提到了很重要的一点，就是“语言材料”，并以之区分凝固的固定短语和临时的类固定短语。李宇明则用“基式”的概念来描述一个格式的本原面貌。这两个观点说明了同一种现象：人们在足够多的经验积累下才能得到理性的认识。“语言材料”侧重于强调经验的实体性（substantive），而“基式”则强调经验的本原性（original）。

Steel指出，每个人心里的语法都是不尽相同的。这是因为，每个人所曝露在的语言环境是不同的。人们会根据他们所接触到的语言表达来构建自己的语言体系。这其中并不存在一个唯一的指导标准，如果说有，那就是整个社会的约定俗成。但这个约定俗成也并不是一种抑制的（suppressive）标准，而是一种允准的（license）标准：在真实语境中，很少有人会直接指出表达的“错误”，他们并不会向你提供直接反例（direct negative evidence），他们顶多用一些隐晦的手段（indirect negative evidence）来提醒你。更为极端的例子是，当下人们甚至刻意地利用“错误”，来造成一种焦点，迫使人们关注并理解表达内容，比如“被自杀”（语法错误），或“用脚投票”（语义错误）。

以上理论皆在说明一种基于使用的分析方法（usage-based approach）。

Goldberg等人提出的构式语法理论发展到今天，其主要目的并不是创造一种“普遍性的格式”，而是要说明一种基于使用的方法。许多学者引用Goldberg的“构式是形式和意义的结合体”定义，并举出著名的“双及物构式表转移义”例子，认为只要抽象出一个格式，并给予其一个意义，就是构式的视角。但其实，双及物构式这个例子的主旨在于“转移义”不是跟谓词绑定的，而是由整个表达提供的。构式语法学家认为，以往的论元结构语法把动词和论元分而治之，因此只能把这个义项安放在主要的动词上——但这是不对的，因为只有当动词出现在双及物结构中，才会产生这个义项。这种由于忽略了格式本身的意义而进行的错误的意义指派的做法，并不是一种基于使用的方法。基于使用的方法应该注意还原一个表达的真实、具体的用法，而不只谈抽象的形式理论。使用频率、同功能表达、错误表达的原因[[8]](#footnote-8)等等，都是基于使用的方法乐意关注的焦点。

诚然，在双及物构式的例子中，基于使用的方法让焦点不再聚焦在动词而聚焦在了整个格式上。但如果把构式只看作一个格式，那么它与传统的短语规则语法没有本质的区别——它不像是一套独立于短语规则语法之外的语法体系，而更像是短语规则语法上的一个“补丁”——短语规则原来没有考虑到格式也能表义、也是语法单位、也能参与组合构造新的表达，现在把它算上不就完了吗？为什么要重新用一个“构式”的概念描述呢？

我们把问题问得具体一点：文炼提到的从“水落石出”到“水落油出”，能抽象出一个“水落X出”的构式吗？陆志韦提到的从“开天辟地”到“开辟天地”——如果你觉得这两个词分不清孰先孰后，那么我们拿“开天辟地”与“天开地辟”、“天辟地开” 或者“地开天辟”比较，你觉得能用构式说明这种构造表达的方法吗？**在有具体形式的构式[[9]](#footnote-9)之外，还有哪些构式的存在？**

本文涉及的与构式相关的概念

我们认为，没有具体形式的构式显然是存在的，虽然这可能会为其释义带来一定的麻烦。我们把“水落油出”这一类通过仿造而成的形式称为仿造构式（adapted construction），把“天开地辟”这一类通过变换汉字顺序而成的形式称为变换构式（reordered construction），把“开天辟地”这一类通过对举而成的形式称为对举构式（contrastive construction），把“人艰不拆”这一类通过紧缩而成的形式称为紧缩构式（abbreviated construction）。这四个构式没有具体的形式，类似于双及物构式。我们可以用“X(a→b)Y”，“X(a⇌b)Y”，“X1Y1X2Y2”和“X[…]Y[…]Z[…]W[…]”这种抽象的规则来概括它们。它们的修辞功能要大于表意功能，如果强调释义，应该从修辞角度着手。

需要指出的是，这些构式间彼此有重合。“天开地辟”不仅符合“天X地Y”构式，更符合对举构式。具体把它归入哪一个构式是因人而异的，这取决于人们究竟如何产生出它：甲可能想换种说法表达“开天辟地”，则其对于甲而言属于变换构式；乙可能从“天崩地裂”推及它，则其对于乙属于“天X地Y”构式；更可能的是，丙并没有生产它，而是听到人们使用它——这又分几种情况了，比如，丙是按还原为“开天辟地”来理解它的，则它加深了丙对变换构式的理解；也可能是按对举来理解的，则它加深了丙对对举构式的理解。一般情况下我们认为，当人们接触到一个表达时，它会同时加深人们对所有相关构式的理解，而当人们生产一个临时的表达时，往往也伴随着对这个新表达的理解过程，至于理解/生产过程中每一个相关构式被加深了多少，这就因人而异、不好评判了。

在基于使用的构式视角下，我们更倾向于描述具体实例对人们的刺激作用。我们倾向于认为，像“天…地…”这样的形式并不真实地存在于人们的意识中；但当人们接触到“天开地辟”的刺激时，人们脑海中存储的“天旋地转”、“天长地久”之类的表达也同时获得了不同程度的激活，因而可能使得脑电中的某一块区域变得更加明亮——而这块区域的中心，我们称之为构式。

此外，在人们的脑海中，每一个有印象的实例都是一个构式，都是一个形式和意义的结合体，都能为遣词造句提供经验。而对于“水落油出”这样的刺激，显然“水落石出”这个实体构式的同激活（co-activated）效应要比仿造构式强烈的多，使得人们更容易反应出“哦，这个表达来自于水落石出”，而不是反应出“哦，这个表达很像人怂志短”。仿造构式是一个很抽象、存在感很弱的构式，我们只是从一类现象中把它抽象出来加以描述。后面还会具体谈到。

最后，我们把诸如“天…地…”的构式称为半凝固型构式。在解释这个概念前先谈两点。

一、Goldberg提出了实体性构式（substantive construction）和图式化构式（schematic construction），前者指称那些具体的、没有槽（slot）的表达形式，后者指称那些带有槽（slot）的表达形式。Goldberg还提出一类半图式化构式（partially schematic construction），指称那些既包含固定成分又带有槽的构式，而用图式化构式专门指称一些无固定成分的语法构式，如双及物构式、被动句构式等等。

二、我们认为，虽然构式语法和短语规则语法在方法论上存在差异，但在形式化的表示上有很大的相似性。前面已经讨论过，如果单从格式的角度出发，构式与短语并无本质的差异。构式，具体说是半图式化构式，在短语规则平面上投射成类似于邵敬敏提出的框式结构，一样可以参与到短语的组合中去；而短语规则，也可以看作静态的图式化构式或半图式化构式，是人们基于既有经验概括出的语言组合规律，同样具有心里现实性。在形式化的表示上具有相似性，这一点对于中文信息处理很有意义，这意味着对于自动句法分析而言，人们不必要新建一套体系来适应构式语法，而只需要多增加一些半图式化构式，就可以覆盖那些传统的短语规则不好处理的语言现象。在不依赖大数据的处理任务中，达到改进结果的目的。

回到半凝固型构式的概念。我们之所以不采用半图式化构式来描述“天…地…”，是出于契合短语规则语法的目的。我们认为，一般的半图式化构式，经实例化[[10]](#footnote-10)后产生的表达更像是短语。比如北京大学现代汉语构式数据库中收录的“不是n的n”构式，它实例化后得到的如“不是老师的老师”其实就是语义矛盾的短语。因为语义矛盾而带来了一种特殊的表达含义，这种含义不能绑定在系动词“是”或者复现成分“老师”上，因此是由词语复现和语义错配带来的构式义。说它是短语，是因为它还可以按短语规则扩展成为“不是老师却胜似老师的老师”或“不是老师的好老师”，这两个表达都保留了原表达的构式义，但又不宜看作新的构式，除非我们有证据表明“不是n却胜似n的n”已经具有广泛的“群众基础”，成为了许多人脑子里的语言材料。否则，我们最好还是只把“不是n的n”看作构式，因为（至少从语料看来）它才具有最明显的心里现实性。

然而四字表达中的“天长地久”不能进行这种扩展，主要原因就是“四字”的硬性规定[[11]](#footnote-11)。因此，为了突出四字表达与自由短语不相符的紧凑性，我们不把它看作是半“图式化构式”，而把它看作是半“实体性构式”。由于“实体”这个翻译不能很直白地让人建立概念，我们采取“凝固”的翻译。“半凝固型”这种构式类型，是与“短语型”构式相对应的，后者才是典型的半图式化构式。

2.2.3构式视角下四字表达的语法地位

我们认为四字表达偏向词的性质，但不应该收入词典中[[12]](#footnote-12)。说是词，是出于以下几点考虑：

a. 大量四字表达结构紧凑，无法拆分成各项都完整的词组；

b. 有些四字表达的内部发生的一些微小的变化：顺序颠倒，或者误用一两个字[[13]](#footnote-13)，都不影响其整体意义的理解，这对于强调语序、成分意义必须清晰的汉语短语来说不成立；

c. 有些四字表达可以被仿造，但仿造后的表达必须依赖原表达，即仿造产生的格式并不是自由格式，这在短语中很罕见（短语仿造需要依赖原表达的程度不大，它多是通过一些特征，如语义错配，达到表达目的）；

d. 对举虽然不是四字表达的专利，但把一个双音节单纯词拆开来构成对举的例子，如“左支右吾”，短语中并不常见；

e. 还有一类四字表达，能够拆分成各项都完整的词组，这类表达固然可以处理为短语，但看作词也无可厚非，这取决于其“语言材料化”的程度。

以上五点，a最主要；b, c, d三点都依附a的存在而存在。对于e点，我们不作强制归类，我们只强调把这类词组当成一种凝固体词或谓词的可行性。

不收入词典，与“类固定短语”的考虑一致：排除掉已经成为语言材料的表达——它们一般被称为成语或固定短语，已经当作词处理了——其余的表达则是临时的、低频的。这些表达“收之不尽”，并没有办法启发人们造出更多的表达。

不收入词典，却应该理解它们的能产方式。前面我们提到了几类构式，就是要描述四字表达的能产性。在信息处理中，应该用这些能产方式识别新的四字表达，把它处理成一个整体，然后像词一样参与更大的短语规则组合。

## 2.3 四字表达的规模

四字表达历史悠久，数量丰富。比如《诗经》，绝大多数为四言。在CCL古代汉语语料库接近500兆字节（Mb）2200万句[[14]](#footnote-14)的语料规模中，字数为四的小句数量占到670万条，接近总量的三成[[15]](#footnote-15)。这种文化的积淀使得在当下，人们仍热衷于“四个汉字”的形式特征。

《现代汉语词典·第五版》共收录四字表达5726条，半凝固型构式34个。这里面既有凝固的成语，也有不那么凝固的表达，如“和平谈判”、“国家赔偿”等。搜狗细胞词库中由用户提供的不同成语词库中包含了从4千条到5万条不等的四字表达，盘古分词词典中则包含49218条四字表达。所有网络来源词库包含的四字表达总条目数达到93050条。

然而这个规模仍难覆盖实际生活中的四字表达。雾霾来的时候，所有发/wù//mái/音的成语都有了新的意味，如“十面霾伏”，甚至连出自《易》经的清华大学校训也难逃一劫：“自强不吸，厚德载雾”；北京大学旁边开了一家主打清淡菜品的餐厅，取名“无油无虑”；股市疯狂的时候，“谈股论金”应运而生。这些都是随手收集的仿造构式例子。根据我们的经验，这种例子绝不在少数，其语用修辞上的丰富程度与其句法语义上的残缺不全形成了鲜明的对比，反映出人们“以异代劳”[[16]](#footnote-16)的造词倾向。

再比如今年的金像奖（第35届）主题：“童声同戏”（Forever Young），不对其和谐、优雅与否作出评价，就构词法来说，有强烈的对举意味，一些网络热词如“男默女泪”、“范跑郭跳”、“林貌杨音”，在我们看来也是受到了对举的影响。此外，“人艰不拆”、“飞蝗芜湖”[[17]](#footnote-17)、“好评如潮”，则与（齐沪扬，2011）定义的虚字类固定短语十分类似。

传媒广告历来是新词语的发源地，因为它们需要焦点；网络作为新兴媒体，覆盖面广，传播迅速，成为新词语诞生的温巢。当吸引目光的需求大于传递信息的需求时，人们开始发掘各种能产方式的潜力，而他们的成果在秒级的时间内就能传遍网络、收获反馈、以致迅速凝固成型。这是我们所认为的当下四字表达的规模如此庞大的根本原因。

## 2.4 四字表达的能产方式

后文主要针对2.2.2小节第二部分提到的四个构式进行定量描述。这四个构式也分别代表了四种能产方式，即变换、仿造、对举和套格式。值得一提的是，对于网络热词，它们的主要能产方式表现为“紧缩”（abbreviated），用尽量少的关键字来概括一句原本很长的表达，也算是一种构式。然而如果没有其他能产方式辅助，并不能有效地识别这个构式的实例，因为形式上的特征实在太少。

我们认为变换是一种消极的构词法，在积极的构词法中，仿造能产性最高。

# 第三章　词库分析

根据牛津词典（Oxford Learners Dictionary）和剑桥词典（Cambridge Dictionay）[[18]](#footnote-18)，词典（dictionary）的首要任务是释义，而词汇（lexicon）则注重完整性描述。我们用到的词库概念更接近“词汇仓库”，即lexicon library，可以同称为lexicon。我们对其中词项（item）的唯一要求是“有意义”，即人们会利用它们组合成更大更完整的语言表达。我们更加关注的是词项间的关系（relationship），以预测它们产生新表达的能力。

我们整理了搜狗细胞词库中所有的成语词库[[19]](#footnote-19)。所谓细胞词库，即“相对于默认词库而言、满足用户个性化输入需求的用户词库”[[20]](#footnote-20)，相对于传统的汉语词典，它具有词项繁杂、新鲜度高的实时性特点，更容易发现新词产生的方式。除去少部分不完整的词库，如“非四字成语”、“成语（做人篇）”、“包含动物的成语”等，和部分完全重复的词库，我们总共收集有效词库共15个。此外，我们还将盘古分词词典也考虑进来，以提高对既有四字表达的覆盖程度。针对这16个词库，我们进一步将其中所有的四字表达提取出来，形成16个四字表达词库，它们构成我们最终分析用到的词库集。

我们最开始考虑将所有词库合并为一个大词库，并进行相关能产方式的统计工作，但发现了一些问题。我们于是重新审视了不同词库间的差异，较好地解决了这些问题，并得到了一些新的结论。

3.1节将首先讨论对大词库的格式的统计结果，3.2节将讨论大词库中的变换构词。这两节的定量分析方法是整体的、静态的。3.3节将从词库差异出发进行相对的、动态的统计分析。

## 3.1 大词库中的格式

16个词库所包含的四字表达总数达到93050条。作为一种静态的、直观的描述方法，我们首先从中抽取了足够的格式。相比于前人从格式出发统计实例的分析方式，从实例中抽取格式的方式更加客观全面。

3.1.1 总体情况

我们采取与（李传军，2008）类似的“型式-格式-实例”三层模型进行统计。然而，我们的统计得到的格式数量要远高于他的结论，在型式能产性上也与他的结论有出入。表3.1给出了不同型式、不同实例数量下的格式数量分布，其中行坐标为不同实例数量范围下的格式数量，列坐标为型式类型。例如阴影标示的单元格（第三行第三列）意义为“型式为‘AB……’、且拥有20~49个实例的格式有193个”。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ≥100 | 99~50 | 49~20 | 19~10 | 9~5 | 4 | 3 | 2 |
| …B…D | 2 | 13 | 74 | 225 | 1065 | 929 | 2275 | 7984 |
| A…C… | 5 | 13 | 89 | 262 | 1152 | 950 | 2244 | 7799 |
| AB…… | 1 | 30 | 193 | 660 | 2505 | 1606 | 3161 | 7640 |
| ……CD | 6 | 43 | 251 | 708 | 2155 | 1485 | 2813 | 7044 |
| …BC… | 1 | 2 | 41 | 183 | 999 | 1021 | 2437 | 8495 |
| A……D | 0 | 0 | 2 | 46 | 719 | 917 | 2406 | 9047 |
| ABC… | 0 | 0 | 1 | 31 | 218 | 310 | 1008 | 5262 |
| AB…D | 0 | 0 | 0 | 1 | 65 | 212 | 882 | 4980 |
| A…CD | 0 | 0 | 0 | 2 | 83 | 161 | 762 | 4720 |
| …BCD | 0 | 0 | 0 | 2 | 163 | 279 | 907 | 5217 |

表3.1 不同型式、不同实例数量下的格式数量分布

可以看到，如果按照“例不十，不立格”的方法，至少有千种格式等待筛选。此外，我们发现AB……和……CD的能产性要高于A…C…和…B…D，除此之外，…BC…的能产性也与后者相当，至少没有明显的差距。

3.1.2 前人未收录格式举例及分析

我们主要以（李传军，2008）和（孟祥英，2010）为参照，在每种型式下挑选未被二者收录的实例数量最高的几个格式举例并讨论。

|  |  |
| --- | --- |
| *型式类型* | *格式举例* |
| …B…D | …地…天（48例），…心…力（42例），…世…俗（38例） |
| A…C… | 万…之…（37例），柳…花…（34例），千…一…（32例） |
| AB…… | 不知……（74例），一言……（42例），令人……（34例） |
| ……CD | ……之士（56例），……一时（35例），……学家（26例） |
| …BC… | …而不…（108例），…人之…（59例），…不容…（35例） |
| A……D | 可……性（26例），以……人（16例），在……上（15例） |
| ABC… | 无所不…（21例），不知所…（14例），一无所…（13例） |
| AB…D | (无) |
| A…CD | 一…千金（14例），一…之见（6例），千…万壑（5例） |
| …BCD | …而言之（9例），…然自若（6例），…不可当（5例） |

表3.2 前人未收录格式举例

…B…D

1. “…地…天”没有被收录大概是因为受到了如（黄燕旋，2011）中分析的、对举的两个概念有一个认知的先后关系的思想的影响——因为拥有实例最多的“…天…地”（189例）的存在，使得人们主观上降低了对“…地…天”的接受度。但这个说法其实是有问题的。一方面，“…言…语”（94例）似乎说不上为什么非要按“言语”认知而不是“语言”，另一方面，“…来…去”（47例）比“…去…来”（19例）的实例多，“…来…往”（28例）却比“…往…来”（29例）要低——至少不再具有优势，而这也说不出什么道理。这种所谓的“认知上的先后关系”更多的是一种使用上的倾向，就像有人在陈述事情时喜欢用主动句有人喜欢用被动句一样。但是，不能因为主动句出现的频率要远高于被动句就否认了被动句的语法地位。实际上，动宾对举如“动地惊天”、偏正对举如“福地洞天”都还是很常见的。

2. “…心…力”没有被收录同样可能是因为“抑制”的原因：当人们考虑到“…心…意”（70例）时，可能就忽视了相似的“…心…力”。应当说，这种相关义对举格式数目繁多，正如心意相通、心手相连、心眼相应、心肺相挨，但凡精神肉体上一些相关联的反应都能带上“心”字，如“撕心裂肺”、“心灵手巧”、“触目惊心”等等。但另一方面，这类名词词素对举格式的能产性却是极其有限的，因为所能应用的场景也就那么多，而且能表达的感情也都表达了，很难再有什么创新。前人所举格式中有一大类都是此类，如“…兵…将”、“人…马…”，真正被人们所熟知的不过三四实例，可以说一定程度上夸大了套格式、尤其是套对举格式的能产性。

3. “…世…俗”为人所知的实例不少，如“惊世骇俗”、“愤世嫉俗”、“欺世乱俗”、“避世绝俗”等等。没有被收录体现了前人工作的不完备性。

A…C…

1. “万…之…”没有收录的原因可能因人而异。对于（孟祥英，2010）而言，是否出于它的构式义不完整的考虑？对于（李传军，2008）而言，是否考虑应该统一划在“……之…”格式里？尽管有这样那样可能的顾虑，但二人又同时都收录了“一…之…”格式，这就表现出格式判别上的不一致性。实际上，“万乘之国”、“万全之策”、“万物之灵”都是很常见的表达。类似地，“百世之师”、“百年之好”、“百兽之王”也为人熟知，而“百…之…”也未被收录。我们认为这两个格式与“一…之…”间并无本质的差异。

2. “柳…花…”第一眼可能只能令人想到“柳暗花明”，但举出一些例子，却又让人觉得不难理解，如“柳啼花怨”、“柳巷花街”、“柳绿花红”。花、柳本是娇弱之物，用作女色之词是恰当的类比。这个格式一方面说明了名词词素对举格式的丰富（不仅能实用，还能虚用），另一方面也说明了应用场景的局限。

3. “千…一…”并不是常规的对举格式，它的两个变项并不处于相同的地位，如“千篇一律”；甚至是不同的词性，如“千金一掷”。这个问题不大，我们可以用广义的“大量对小量”来理解它。但是，在李、孟二人的文章中，却都提到了“一…千…”的格式，这再一次反映了认知的不一致性：在这里，“千钧一发”、“千篇一律”乃至“千虑一失”都是“优势构象”[[21]](#footnote-21)，“一…千…”不再占有绝对的优势地位。

AB…… & ……CD

1. “不知……”、“……之士”、“一言……”反映出了多继承性问题[[22]](#footnote-22)。对于“不知其实”，它是来自“不…其…”，还是“不知……”，还是“不知其…”？对于“不羁之士”，是来自“不…之…”，还是“……之士”？对于“一言不发”，是来自“一…不…”还是“一言……”？以上每一个实例都可以看作继承了不同格式的形式和意义。

一个值得注意的问题是，虽然“A…C…”的实例要多一些，但它并不比“AB……”或“……CD”更有意义。究其原因，这些“A…C…”格式最初并不是语汇格式，而是语法格式，许多用法逐渐凝固下来，保留至今。因为格式本身的语义并不明确，我们在造出一些新实例时的生产过程是很复杂的，既包含了对既有格式的套叠和对既有表达的仿造，又包含了一个语法组织和紧缩的过程。

反观“AB……”或“……CD”型式，虽然我们感觉其只是由继承自其它不同格式的实例“偶然地部分重合”形成的：如“不知……”虽有“不知其实”、“不知所措”等实例，但前者更像是继承自“不…其…”格式，后者更像是继承自“不…所…”格式；“……之士”虽有“不羁之士”、“一介之士”实例，但前者更像继承自“不…之…”格式，后者更像是继承自“一…之…”格式。但是，它们的格式义却更直白一些：“不知……”就是“不知道……”的紧缩形式，“……之士”就是“……的人”的紧缩形式，“一言……”就是“一句话……”的紧缩形式。这种更像是通过紧缩而非套格式而来的实例，我们认为将它们归入“AB……”或“……CD”型式也无可厚非。它会同时加强两类型式的印象。

实际上，李传军所谓的“一字支撑型”除了“…B……”和“……C…”之外，也应该考虑“A………”和“………D”型式。我们对93050个实例的统计显示，当按单个汉字及其出现的位置进行统计时，出现在二位和三位上的字并不占多数，尤其当我们排除掉其他格式的干扰时。见下表：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 阈值 | 排名 | | | | | | | | |
| - | 之3 | 不3 | 一1 | 不1 | 无3 | 不2 | 一3 | 心2 | 人2 |
| 5 | 之3 | 不3 | 子4 | 大1 | 人4 | 不2 | 人2 | 无3 | 一1 |
| 2 | 子4 | 大1 | 之3 | 人4 | 不2 | 三1 | 以1 | 金1 | 里2 |
| 1 | 不2 | 大1 | 小1 | 儿4 | 人4 | 子3 | 子4 | 一3 | 三1 |

表3.3 按单个汉字及其出现位置统计排名

其中阈值的意思为实例所属格式[[23]](#footnote-23)所包含的最大实例数，例如“不…其…”格式有33个实例，则其所有实例，如“不知其实”、“不乏其人”，都被排除在阈值为5、2、1的统计范围之外。换言之，设立阈值的目的就是想考察当多固定词素格式都被排除之后，剩余实例对单固定词素格式的支持。

这个统计很好地说明了，除了前人提到的“……之…”“…不……”之外，许多第一字或第四字为固定词素的一字支撑型结构都具有发展成为格式的潜力。如：

大+Subject+Predicate： 大汗淋漓，大江东去

大+Predicate+Object： 大吐苦水，大兴土木

一+Subject+Predicate： 一票否决，一醉方休

一+Predicate+Object： 一改故辙，一决雌雄

我们认为，像“不…之…”、“一…不…”也可以看作同时继承了“不………”、“……之…”或“一………”、“……不…”，这样，由“……之…”半实例化为“……之士”，由“一………”半实例化为“一言……”也就不那么难接受了。

2. “……一时”、“令人……”、“……学家”也兼具紧缩和套格式两种构词方式。它们的实例一般没有多继承的问题，也不能再抽象出单固定词素格式。

…BC…

1. “…而不…”与“不…之…”一样，也可以看作多继承自“…而……”和“……不…”，而且它的格式意义要比“不…之…”清晰，表示一种转折关系。

2. “…人之…”主要有两种结构：一种是2+2结构，如“妇人之仁”、“惊人之举”；一种是1+3结构，如“成人之美”、“乘人之危”。即便如后者，人们出于韵律的原因还是喜欢用2+2韵步去阅读它们。

我们主要讨论后一种结构。若按照2+2韵步阅读，很难把“…人之…”接受为一个格式。但这个格式又能写出释义模板：“……别人的……”，且由于“别人”这个概念很常用，所以格式的能产性也很高，如“拿人钱财，与人消灾”就可以改造为“拿人之财，消人之灾”。

我们举这个例子更多地还是想说明不一致性的问题：从形式看来，“…人之…”的确满足内部凝固、外部自由的特点，从意义看来，它也有明确的应用场景。但是人们还是不乐意承认其为一个格式，而更愿意接受“一…之…”这样的格式。

3. “…不容…”、“…以为…”这样的格式能产性也很高，但都没有被前人收录。这是不完备性的体现。

A……D

1. “可……性”与“在……上”类似，前者更像是语汇格式，后者则是典型的语法格式。之所以没被前人收录，可能是因为它们的实例都太白话、太不像成语了。

2. “以……人”主要是“以+n+v+人”结构，又一次体现出“人”作“别人”解时的能产性。但同样地，它的格式感也十分之弱。

ABC…

“无所不…”、“不知所…”、“一无所…”都包含所字结构，空位中一般填动词或少量形容词。它们没有被收录是不完备性的体现。

A…CD

“一…千金”，“一…之见”，“千…万壑”都可以看作“A…C…”半实例化的结果。

“一…千金”，容易想到“一刻千金”、“一饭千金”、“一掷千金”等实例。它的接受度要高一点。

“一…之见”，除了稍微熟悉的“一己之见”，在细胞词库中还有“一偏之见”、“一得之见”、“一隅之见”等。百度百科认为这些四字表达都是成语，且皆有出处，如“一偏之见”出自《宋史》，为原文[[24]](#footnote-24)；“一得之见”出自《史记》，原文“愚者千虑，必有一得”；“一隅之见”出自《论语》，原文“举一隅不以三隅反，则不复也”。对于后两个四字表达，虽不能验其出处真假，但通过CCL古代汉语语料库，可以查到它们在近代确实是存在的。这个格式由于其实例的不常见性，接受度要低很多。

“千…万壑”，第一眼看上去，总感觉是一个完整的四字表达挖去了第二个字。但究竟是哪个四字表达，又说不清楚。有人能想到“千山万壑”、“千峰万壑”，也有人见过“千岩万壑”、“千沟万壑”，甚至造出“千嶂万壑”、“千谷万壑”。这个格式给人的感觉是“描述‘峰峦叠嶂沟谷迂回’的意味如此清晰，以至于绝不像一个格式，要么把上述实例看作从‘千…万…’而来”，但我们在CCL现代汉语语料库中找到了这样一例：

记忆有时是无味的，却千叠万壑的幽深，让人沉潜其中，不知岁月的流转。

——林清玄散文

“叠”跟“壑”不是对举的关系，看作从“千…万…”而来不妥。我们认为用仿造构式来解释它比较好：即林清玄受到了某个他已知的实例的影响，为了强调“叠”，而下意识地仿造了一个新表达。

…BCD

“…而言之”、“…然自若”、“…不可当”与“千…万壑”类似，给人一种从某个完整表达中派生出来的感觉，比如“总而言之”、“怡然自若”、“锐不可当”。不那么常见的“统而言之”、“泰然自若”、“勇不可当”，像是通过仿造得来的。又如“进而言之”、“昂然自若”、“势不可当”，意义上有了些改变，但结构上还是一样的。

3.1.3 有关多继承性的进一步讨论

“继承”的概念来自于计算机科学中面向对象的设计思维。一个人（Class Human）有胳膊（Class Arm）有腿（Class Leg），但一个人不一定从事某种具体的职业，有些人是学生（Class Student），有些人是老师（Class Teacher）。我们说人有胳膊和腿（Human has Arm and Leg），但我们不说人有学生和老师（\*Human has Student and Teacher），我们说学生和老师是人（Student and Teacher is Human）。“有”（has-a）和“是”（is-a）是对象间的两种关系，其中前者又称为组合关系，而后者又称为继承关系。学生和老师继承了人的基本属性，而他们自己又分别拥有一些新的属性。在继承关系中，学生、老师和人处于同一层级，他们都是人。

短语语法中，语言现象和语法规则间不是继承关系，而是组合关系。一个句子有主语和谓语，主语又由修饰语加中心语组成。由于规则的排他性，组合关系不能很好地解决歧义问题。例如，“根本”是区别词还是副词？句首时间词是主语还是状语？对于前一个问题，前人的解决办法是“兼类”，而对于后一个问题，多数观点是靠能否插入“是不是”来强制判断，也有人认为分析为主语或状语都行。靠“是不是”来判断，容易出现不一致性，而其他两种处理方法，实际上已经打破了体系的严整性。

构式间的关系则更像是继承关系。构式并不独立于语言现象，它是一类语言现象的集合。构式语法对歧义现象是友好的，因为大家都处于同一层级（都是语言现象），所以图式化构式间是可兼容的。从构式角度看来，“根本”既继承了区别词构式作定语、不能受“很”、“不”修饰、不作谓语等属性，又继承了副词构式作状语的属性。但是它没有继承区别词“只作定语”和副词“只作状语”的属性，因为它确实既可以作定语又可以作状语。兼类词是一种多继承关系——假设我们潜意识中的确具有词性的概念。

多继承性在四字表达中尤有体现。格式数量庞大，意味实例间的关系错综复杂。许多实例都对应着若干“潜在的”格式，见下表：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 阈值\继承数 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| 2 | 108 | 1503 | 2597 | 7978 | 4149 | 17068 | 15124 | 7496 | 19201 | 13021 | 4805 |
| 5 | 0 | 5 | 19 | 244 | 164 | 1230 | 3290 | 4020 | 20891 | 35288 | 27899 |
| 10 | 0 | 0 | 0 | 0 | 11 | 107 | 507 | 791 | 8711 | 32306 | 50617 |

表3.4 四字表达多继承统计

其中阈值为格式所需满足的最小实例数目。可以看到，即便按“例不十，不立格”的标准，仍然有超过一万个实例具有多继承性。

在结构主义语法学家看来，从多个格式继承是难以接受的，因为这意味着格式这种特殊的句法规则失去了其指导性。但我们认为，格式不是独立于实例之外存在的，“套格式”本质上是一种仿造，是受到既有实例的影响而产生的，会增强所有与其相关的格式的印象。对于“一言不合”，如果它是由仿造产生的，最可能的仿造对象不是“一毛不拔”（对应“一…不…”格式），也不是“一言难尽”（对应“一言……”），而应该是“一言不发”（对应“一言不…”），因为这种仿造路径最短，最经济省力。但是，它也同时加强了其他两个格式的印象，特别在共时统计中，人们会把它加入到“一…不…”和“一言……”的实例集中。

多继承性本身不是缺点，但放在共时研究中会带来一些问题。最主要的一点是与不一致性的重叠：比如“一言不…”、“千…万壑”，说它们是仿造过程中的固定部分没有问题，但说它们是格式可能为时过早——你怎么知道是实例造就了格式，还是格式造就了实例呢？

3.1.4 小结

静态的格式分析的优点是直观（direct）和大量（substantial），缺点是不一致（inconsistent）和不完备（incompleted）。

从共时层面来看，四字表达类聚成格式的程度很高。其中有少量格式实例众多，容易给人们留下印象，如对举式的“东…西…”、“千…万…”和虚字式的“不…之…”、“一…不…”；也有部分格式结构工整，表义明确，如“…兵…将”、“心…手…”，也容易被人们注意到。这两类格式主要分布在“A…C…”和“…B…D”型式中，成为前人研究的重点。

但是，还有更多未被前人收录的格式——无论实例数目，如“…而不…”、“不知……”，或是工整程度，如“…世…俗”、“…心…力”，都与前人所举格式不相上下，这是不完备的地方。此外，有一些格式如“万…之…”、“…地…天”，也许会引起前人注意，但最终未被收录；或者如“…人之…”、“以……人”一看就不像格式、但又无法说明与其他虚字式格式的本质区别，这是（收录标准）不一致的地方。

不完备性和不一致性不仅在前人的研究中有体现，在我们的研究中亦有体现。一方面，词库中收录的四字表达仍不足以概括现实中四字表达的全貌，仍然存在不完备的地方；另一方面，许多四字表达被收入词库中是偶然的，它们在现实中难以复现，而大多数人对它们并无印象。它们的存在使得潜在格式的数量大大增加，而这些潜在格式到底是不是真的格式，存在很大的争议。换言之，绝大多数格式，包括许多结构工整的“两字交替显现、两字交替隐含”格式，其能产性是有限的。

此外，我们关注到词库中存在多继承性，它要求我们进一步区分仿造和套格式两种构词方法，为此我们需要寻找一些非整体性的研究方法。

## 3.2 大词库中的变换

除了格式研究外，我们还研究了大词库中的变换。互为变换的两个表达拥有相同的汉字组合，只是字序列不一致，它与格式一样，也体现了四字表达间的关系。

3.2.1 总体情况

表3.5展示了93050个四字表达间的变换关系。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变换类型 | 数词 | 叠字 | 二四交换 | 一三交换 | 一二交换 | | 三四交换 | | 一四交换 |
| 组数 | 61 | 293 | 245 | 185 | 150 | | 119 | | 24 |
| 变换类型 | **二三交换** | **前后交换** | **逆序** | **前后分别交换** | | **其他二词组** | | **多余二词组** | |
| 组数 | 56 | 2372 | 12 | 69 | | 94 | | 475 | |

表3.5 大词库中的变换总览

其中“数词”表示互为变换的两个四字表达包含数词，如 [三十二岁，二十三岁] 、 [第七十九，第九十七] ；“叠字”表示两个以上字重复的表达，如 [鼎鼎大名，大名鼎鼎] 、[不徐不疾，不疾不徐] 、 [暮暮朝朝，朝朝暮暮] ；“前后交换”等价于“一三交换且二四交换”；“逆序”等价于“一四交换且二三交换”；“前后分别交换”等价于“一二交换且三四交换”；“其他二词组”表示除前所述所有交换类型之外的、只包含两个四字表达的分组，如 [千姿万态，姿态万千] 、 [泪下如雨，泪如雨下] ；“多余二词组”则表示包含两个以上四字表达的分组，如 [山重水复, 水复山重, 重山复水] 。

包含数词的表达被收录进词库纯粹是基于工程层面的考虑，不是我们研究的范围。包含叠字的表达不好划分交换类型，如 [不徐不疾，不疾不徐] 既可以看作二四交换又可以看作前后交换，应该单独讨论。此外，其他二词组里的交换规则比较复杂，也比较零散，我们统一作为一类变换讨论，而多余二词组里的每一组则可以两两比较，也应该单独讨论。

我们把“数词”、“叠字”和“多余二词组”看作特殊的变换类型，而把剩余的变换类型称为基本变换类型。下面分别对其进行讨论。

3.2.2 基本变换类型分析

二四交换

二四交换主要分两种情况：对等对举和不对等对举。对等对举即交换后不改变原意，如 [东奔西跑，东跑西奔] ，这种情况较多（193组）；不对等对举即交换后意义发生变化，如 [事倍功半，事半功倍] ，这种情况较少（51组）；有1组不是对举关系，即 [一夫之勇，一勇之夫] ，它们形式上呈交换状可看作偶然。

对等对举能发生二四交换的最主要原因是对举工整，由于前后两字内部结构一致（例如同为主谓结构或同为动宾结构），且一三字、二四字意义相近，故交换后原意不变。有一类交换比较特别，如 [此起彼伏，此伏彼起] 、 [春来秋去，春去秋来] ，它们从字面上看来并非对等交换，但其意义是“周遍”的，这个前人已有许多论述，如（侯人渝，2015），我们把它们也看作是对等对举。

不对等对举则分为两类，一类是大家比较能接受的，如 [貌合神离，貌离神合] 、 [事倍功半，事半功倍]，它们对举工整，是典型的对举格式；还有一类，如 [自上而下，自下而上] 、 [量入制出，量出制入] ，它们从逻辑上不属于并列，而属于顺承，但我们认为这类表达刻意选取两个端点进行陈述比较，并不是单纯地顺承叙事，确实利用了对举格式。例如我们常说“自上而下”，而不说“自上而左”；常说“非黑即白”，而不说“非黑即红”，后两者都有成立的语境，但与前两者比起来，对举的意味要弱一些，而前两者对举的意味则是较强的。

我们认为对等对举和不对等对举是对举的两种基本类型。我的导师詹卫东先生认为，把两样事物刻意放在一起说，无非两种情况：要么突出他们的共性，要么则要体现他们的不同。我们把它概括为“求同”和“存异”。“求同”和“存异”都是很常见的，例如“生生死死”，如果解读为生死两别，那就是在“存异”，而若解读为“存者且偷生，死者长已矣”，或者从生死轮回的角度出发，那又像是“求同”了。然而从分布看来，确实存在着“大同小异”的现象，即“求同”居多，“存异”居少。一个可能的原因是，四字表达高度概括，一定的信息冗余既能起到强调的作用，又可以简化生产和理解过程，而“求同”的对举实际上就是一种信息冗余。这对于词语双音节化的现代汉语来讲尤为实用——试想一下，当你要强调“刻苦学习”时，造出一个形式上冗余的“刻学苦习”是不是一下子耳目一新[[25]](#footnote-25)？

此外，在二四交换类型中我们还发现了一个较普遍的现象，即“形式误用”。我们把形式误用构式（form-misused construction）定义为“意义与原表达一致但形式与原表达有所不同的构式”。假借字就是一种典型的形式误用构式。在四字表达中同样存在这类构式，如：a. [春露秋霜，春霜秋露]，b. [如石投水，如水投石] / [两豆塞耳，两耳塞豆]，c. [翘首企足，翘足企首]，d. [每况愈下，每下愈况]。其中a例是绝对误用，“春露秋霜”是不对等对举，以春露喻恩泽而以秋霜喻威严，意为赏罚分明，多用怀念先人，由于感情义差异，虽然“春霜秋露”客观存在，但语用上绝无此说。其余例子为潜在误用，其中b例中每组的后者均为隐含被动表达（“如水（被）投石”、“两耳（被）塞豆”），虽然也能理解，但显然还是前者更加符合原结构[[26]](#footnote-26)；c例中“翘足企首”应该不符合古汉语语义搭配，《说文》谓“企”：“举踵也”，在粤方言中也保留着“企”作为“站立”的用法，而“翘”字虽不专门描述抬头，但它的本义是“尾部长长的羽毛”，即便说“翘足”，也应该是抬起脚尖而不是脚跟，所以“翘足企首”应该是今人的误用，只不过今天人们对“翘”和“企”没有那么多讲究，也就不太在意具体的搭配了；d例中的两个表达都“很古老”，甚至“每下愈况”更古老些，但今人已不用“每下愈况”的本义，

# 第四章　语料库分析

## 3.1不经意的错误：误用能产性

## 3.2 巧妙的引用：仿造能产性

## 3.3 言之不足，故复言之：对举能产性

## 3.4 韵律之美：紧缩能产性

# 第五章　相关构式的进一步分析

# 第六章　结语

第五章说明构式的继承体系和网状结构。

# 参考文献

陆志韦．汉语的并立四字格［Ａ］．//陆志韦．陆志韦集［Ｃ］．北京：中国社会科学出版社，2003: 400-457

吕叔湘．现代汉语单双音节问题初探［Ｊ］．北京：中国语文，1963（1）

郭绍虞．汉语语法修辞新探［Ｍ］．北京：商务印书馆，1979

姚殿芳，潘兆明．说“四字格”［Ｊ］．新疆：新疆大学学报（哲学社会科学版），1985（4）

文　炼．固定短语和类固定短语［Ｊ］．北京：世界汉语教学，1988（2）

齐沪扬．有关类固定短语的问题［Ｊ］．上海：修辞学习，2001（1）

范喜梅．由成对方位词构成的待嵌格式研究［Ｊ］．山东：现代语文，2009（11）

黄燕旋．方位对举四字框式结构研究［Ｄ］．广州：暨南大学，2011

王　娇．现代汉语对举嵌置式四字格习得研究［Ｄ］．南京：南京师范大学，2013

韩立秋．构式语法视野下古汉语凝固结构研究［Ｄ］．吉林：东北师范大学，2015

李传军．类固定短语相关问题研究［Ｄ］．上海：上海师范大学，2008

李宇明．词语模［Ａ］．//邢福义．汉语语法特点面面观［Ｃ］．北京：北京语言文化出版社，1999: 146-157

周　荐．双字组合与词典收条［Ｊ］．北京：中国语文，1999（4）.

周　荐．《现代汉语词典》中的待嵌格式［Ｊ］．北京：中国语文，2001（6）

马国凡．四字格论［Ｊ］．内蒙古师大学报，1987（3，4）

张拱贵．语法格式与语汇格式——《现代汉语“格式”初探》序［Ｊ］．吉林：汉语学习，1985（5）

孟祥英．汉语待嵌格式研究［Ｄ］．山东：山东师范大学，2010

# 致谢

# 北京大学学位论文原创性声明和使用授权说明

**原创性声明**

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

**学位论文使用授权说明**

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日

1. 语出作者某同学微信朋友圈状态：“兴致冲冲跑去听大神报告，结果一瞄会议室里坐满了人还欢快地聊着，突然无端地脑袋一抽，拉开隔壁实验室的门搬砖去了……真是人怂志短，一念之间。” [↑](#footnote-ref-1)
2. 格式有两种，一种是具体的格式，即包含固定词素的格式，如“千…万…”；另一种是抽象的格式，例如“对举”格式，可以概括为“X1Y1X2Y2”，其中X1Y1、X2Y2的语法结构相同。这里说的“套格式”主要是针对前一种格式而言。 [↑](#footnote-ref-2)
3. 张华平博士提供的NLPIR的分词结果为“人/n 怂/vg 志/n 短/a”，其中vg代表“动词词素”，应该与“怂恿”相关，说明分词程序不理解“怂”作为形容词的含义。 [↑](#footnote-ref-3)
4. 张国宪这篇文章并不针对四字表达进行讨论，但后人研究多以其文为参考，沿用“对举”概念。另外，“对举格式”只是四字表达中的一部分，不能概括四字表达全貌，这与其他派系不同。 [↑](#footnote-ref-4)
5. 例句如下：“龙二井有油和水的矛盾，这是它的特殊性。周队长说，要促使矛盾转化，就要捞水，把水捞干。我们想一不做，二不休，搞它个水落油出。” [↑](#footnote-ref-5)
6. 原文采用的是“语例”的说法。 [↑](#footnote-ref-6)
7. 原文中称“两字交替显现，两字交替隐含”，我们简称为交替显现；同样地，后面的邻接显现，原文为“两字邻接显现，两字邻接隐含”。 [↑](#footnote-ref-7)
8. 构式语法中一般不用“\*”来标示一个错误的表达，而顶多用“??”来标示该表达的不常用性。 [↑](#footnote-ref-8)
9. 即一个带有固定成分和空位的形式，如“…兵…将”（eg. 虾兵蟹将）、“天…地…”（eg. 天长地久）、“吃……饭”（eg. 吃大锅饭）、“……之…”（eg. 大喜之事）等等。 [↑](#footnote-ref-9)
10. instantiation，即向构式槽中填入适当的成分形成一个完整的表达。 [↑](#footnote-ref-10)
11. 当然，有些如“一发而不可收拾”和“一蹴而就”确实都又像是“一…而…”的实例，这时候就看具体的任务需求来决定如何权衡了。从使用的角度我们认为这两者都是实体构式，后者加强了“一…而…”的构式感，而前者除了加强“一…而…”的构式感外，还加强了“一……而不可……”的构式感，所以后者在短语规则屏幕宜看作词，而前者宜看作短语。 [↑](#footnote-ref-11)
12. 当然，作为信息处理提高匹配效率和匹配准确率的手段，对于这种表达“见一个收一个”也无可非议。 [↑](#footnote-ref-12)
13. 误用的例子很多，见后文具体讨论；此处仅举两组对比：“时过境迁”常被误用为“事过境迁”或“时过景迁”，这是一种意义还算正确的误用；“虚无缥缈”常被误用为“虚无飘缈”，这可能就属于一种想当然的、意义并不正确的误用了。当然也可以把“事过境迁”看作一种“刻意的误用”，也就是仿造了。 [↑](#footnote-ref-13)
14. 古汉语语料中的标点多为后人所加，但符合我们对句子的认知，可以进行相关统计。我们断句的原则是逢标点则断，这意味着逗号、引号等符号也被当做分隔符。这样分出的小句更方便统计，且同样代表了人们对语言的组块化认知，具有实际意义。 [↑](#footnote-ref-14)
15. 6708243 / 22169755 ≈ 30.26% [↑](#footnote-ref-15)
16. 这个新造的词有两层仿造，第一层是“以逸‘代’劳”，是一种半误用半仿造的表达，形容“偷懒不劳作”；我们再改造一次，变成“用异构既有表达的方式省去了重新创造的麻烦”。 [↑](#footnote-ref-16)
17. 原词“非黄勿护”，这里经过了一次“加密”式仿造，详见后文。 [↑](#footnote-ref-17)
18. 全部释义参考牛津初学者词典在线版本（<http://www.oxfordlearnersdictionaries.com/>）及剑桥词典在线版本（<http://dictionary.cambridge.org/>）。 [↑](#footnote-ref-18)
19. 语料下载地址：<http://pinyin.sogou.com/dict/cate/index/54> [↑](#footnote-ref-19)
20. <http://pinyin.sogou.com/help.php?list=6&q=1> [↑](#footnote-ref-20)
21. 一个化学分子可能存在多种空间构象，其中最稳定、出现几率最高的一种称为优势构象。这里，“千钧一发”相对于“一发千钧”来说更容易被使用，故为优势构象。 [↑](#footnote-ref-21)
22. 继承这个词来源于计算机科学的面向对象设计思维，指一个具体的对象会从一个抽象的对象中继承属性和方法，如“三角形”继承了“形状”的可绘制性，而“等边三角形”又从“三角形”处继承了“三个角”的属性。 [↑](#footnote-ref-22)
23. 这里的格式不包含单固定词素格式，特指表3.2中所列10种包含两个及以上固定词素的格式。 [↑](#footnote-ref-23)
24. “安石任一偏之见，改立新议，以害天下大公。” [↑](#footnote-ref-24)
25. “刻学苦习”并不是真的信息冗余。“学”和“习”在古代汉语尚且是两个近义词，“刻苦”则未见分而用之的例子。但把“刻苦”、“学习”拆开，就像是由一个词变成了两个同义词素一样，造成了一种冗余的假象。 [↑](#footnote-ref-25)
26. “如石投水”源自“以石投水”；而“两豆塞耳”则在原文中与“一叶障目”对仗。 [↑](#footnote-ref-26)