

硕士研究生学位论文

**题目：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

四字表达在汉语中历史悠久，形式丰富。在使用中不仅积累了许多成语，当下更涌出了大批类似成语的表达，例如“人怂志短”。这些表达形式紧凑，应该看作词；但临时性强，一方面频率很低，另一方面条目众多，现有词库中既没有收录，也难以收尽。我们应该找到它们能产的规律，从而更好地为中文信息处理中的未登录词识别提供支持，也是对前人理论研究的补充。

前人的研究多着眼于既包含固定成分又带有空位可接受其他成分填充的格式，他们不仅应用了定量分析的手段，还引入了多种先进理论辅助论述。然而，前人定量分析是“从格式到实例”，即先收集一批格式，再到语料库中匹配实例，虽然利用了语料的客观和大量的特点，但缺点则体现在对格式判别的不一致性上——如收录了“天…地…”格式却不收录“地…天…”格式，和内省的不充分性上——如“…世…俗”这样实例众多的格式却没有被前人注意到。

我们则采用了“从实例到格式”的分析方法，从分词词库中提取包含十个以上实例的格式。这样一定程度上避免了以上缺点，但仍然受限于词库收词的有偏性，不可能完整地提取所有已存在的格式。更重要的是，大量四字表达无法被格式覆盖，如从“人怂志短”无法抽象出“人…志短”或“人…志…”，因为这两个格式实例数量太少。因此，我们在格式套用的基础上基于前人提到的一些现象提出了变序、仿造、对举三种能产方式，从词库——它收录了明确标出了边界的四字表达，和语料库——它蕴藏着大量还未被识别的四字表达，两方面分析了其存在性和能产性，覆盖到了更多格式照顾不到的四字表达，也发现了更多应用格式套用所无法发现的四字表达。

具体地，我们把变序定义为“只改变字的顺序”的能产方式，例如“开天辟地→开辟天地”，把仿造定义为“只改变一个字”的能产方式，例如“水落石出→水落油出”，把对举定义为“前后对举且一三字、二四字分别对举”的能产方式，例如“开天辟地”，其中“开天”与“辟地”对举，且“开”与“辟”、“天”与“地”分别对举。在词库分析中，我们得出了仿造能产性高于对举能产性和变序能产性之和的结论；而在语料库分析中，我们则发现当把对举的字限定在实词范围内时其识别准确率大大提升，且能产性也要高于仿造。我们认为，具有十个以上实例的“典型格式”的数量是很有限的，而仿造和对举是对不足十例的“潜在格式”的很好补充，其中仿造补充了三固定字潜在格式，而对举补充了“A…C…”和“…B…D”型潜在格式。

此外，我们发现一个实例往往可以归入多个格式，或通过多种能产方式解释其存在。例如“不毛之地”，它既来自于“……之地”格式，又来自于“不…之…”格式；“地辟天开”，既可以看作对举，又可以看作变序。我们通过基于使用的构式语法理论分析了这种范畴重合现象。它反映出人类语言知识的网状储存特点以及认知的主观差异。通过这个分析，我们阐述了我们的构式语法观，并试图在着眼于知识的高层抽象的形式语言学和着眼于知识的底层表征的认知语言学乃至神经语言学间建立可靠的联系。

关键词：四字表达，格式，变序，仿造，对举，构式语法

A Productive Research on Modern Chinese Four-Character Idiomatic Expressions

Tian Jun (Chinese Linguistics)

Directed by Professor Zhan Weidong

# ABSTRACT

Four-character idiomatic expressions abound and have a long history in Chinese. Many four-character idioms appeared during time, and new idiom-like expressions are still in production, such as “Ren Song Zhi Duan”. These expressions have a tight form, which should be considered as words. However, they are so temporary and rare that collecting them in dictionary is hardly useful. It’s more important for us to find their laws in production, so that we can have a better performance in Unregistered Word Recognition task and have an improvement in earlier researches.

Earlier researches often focused on form pattern which both contains constants and variables. Researchers not only took a quantitative analysis, but also introduced various theories to help discussion. However, these researches take a "pattern to instance" method. Researchers first collected a bunch of patterns, then match their instances in corpus. Although the corpus is intuitive and substantial, but these researches have disadvantages such as inconsistency, that researchers only considered “Tian … Di …” as a pattern but left “Di … Tian …” in back, and incompleteness, that such a pattern like “… Shi … Su” wasn’t taken into consideration.

We change the method to "instance to pattern", extracting patterns that hold more than ten instances from segmentation dictionary. It avoids the disadvantages mentioned above, but is still limited to the bias of the dictionary. Moreover, many four-character expressions cannot be covered by patterns, such as “Ren Song Zhi Duan”, it cannot be abstract into “Ren … Zhi Duan” or “Ren … Zhi …” due to rarity in similar expressions. Therefore, we propose three kinds of productive way of four-word expressions: Reordering, Adapting, and Contrasting, which based on some phenomena that researchers mentioned. We analyse their productivity in two ways: in dictionary, which collects certain four-character expressions, and in corpus, which contains plenty of unrecognized four-character expressions. Thus more expressions that form pattern cannot cover are considered and found.

Precisely, we define “Reordering” as a productive way that only changes ideographes’ order in the expression, such as “Kai Tian Pi Di → Kai Pi Tian Di”; we define “Adapting” as a productive way that only changes one ideograph in the expression, such as “Shui Luo Shi Chu → Shui Luo You Chu”; we define “Contrasting” as a productive way that the former part and the latter part of the expressions constrast as well as the 1st and the 3rd ideographes contrast and the 2nd and the 4th ideographes contrast, such as “Kai; Tian | Pi; Di”. In dictionary analysis, we make the point that Adapting is more productive than Contrasting plus Reordering, while in corpus analysis we find that Contrasting features are more reliable when adding the restriction that ideographes should only be nouns, verbs or adjectives, it is also more productive than Adapting. Since the "typical patterns" that have more than ten instances are rare, we conclude that Adapting and Contrasting are good suppliment for "potential patterns" that have less than ten instances. Adapting supplies the patterns with three constant characters, while Contrasting supplies “A … C …” and “… B … D” types of patterns.

In addition, both the patterns that we find and the productive ways that we define can be overlap. For example, “Bu Mao Zhi Di” can either be from “… … Zhi Di” or be from “Bu … Zhi …”; “Di Pi Tian Kai” can either be considered as a contrastive expression or a reordered expression. We discuss this phenomenom through the usage-based construction grammar. It shows the network structure of memory of our language knowledge and the subjective cognition distinct between individuals. By this discussion we state our view of construction grammar, and we are trying to build reliable connection between formal linguistics which focus on high-level abstraction of knowledge and neural linguistics which focus on base-level representation of knowledge.

KEY WORDS: four-character idiomatic expression, form pattern, reordering, adapting, contrasting, construction grammar.

# 目　录

摘要 I

ABSTRACT III

目　录 V

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 前人研究的特点和不足 5

1.4研究方法和特点 6

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

2.3 四字表达的规模 10

2.4 四字表达的能产方式 11

第三章　四种能产方式在词库中的描述和比较 12

3.1 概述 12

3.1.1 定义 12

3.1.2 分析方法 12

3.2 词库中的格式 15

3.2.1 总体情况 15

3.2.2 前人未收录的格式的分析 16

3.2.3 前人收录的格式的分析 19

3.2.4 小结 22

3.3 词库中的变序 23

3.3.1 总体情况 23

3.3.2 基本变序类型举例及分析 23

3.3.3 小结 27

3.4 词库中的仿造 27

3.4.2 仿造能产方式与其他能产方式的能产性比较 29

3.4.3 小结 33

3.5 词库中的对举 33

3.5.1 对举能产方式与其他能产方式的联系 33

3.5.2 对举能产方式与其他能产方式的能产性比较 34

3.6 本章总结 35

第四章　语料库分析 36

4.1 工作概述 36

4.1.1 分词工具的选择 36

4.1.2 辅助识别的方法 36

4.1.3 语料的选择和处理 37

4.2 利用格式套用发现新表达 38

4.3 利用变序发现新表达 39

4.4 利用仿造发现新表达 41

4.5 利用对举发现新表达 42

4.6 本章总结 43

第五章　构式语法视角下的范畴重合现象的解释 44

5.1我们对构式语法的理解 44

5.1.1 构式语法是基于使用的理论 44

5.1.2 构式不脱离于实例单独存在 45

5.1.3 构式可以没有具体形式 46

5.2 构式视角下的范畴重合现象的解释 47

5.2.1 构式视角下的多继承性 47

5.2.2 构式视角下的能产方式的重叠 48

5.2.3 构式视角下的形式误用 49

第六章　结语 51

参考文献 52

致谢 54

北京大学学位论文原创性声明和使用授权说明 55

# 第一章　引言

## 1.1 问题的提出

四字表达可分为两种：成语和自由短语。“人穷志短”是成语，需要收进词典；“中国制造”则可以看作一个偏正结构的自由短语。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，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。而如果看作成语，它的临时性很强，并不是那种会进入我们的“语言素材库”中、被我们频繁运用的单位。

类似的表达数量非常之多，它们的表达效果非常强，甚至超过了它们作为信息载体传递信息的作用。信息时代里，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…”，来规整一部分四字表达，并启发人们创造。但我们认为，“格式套用”并不能完全诠释四字表达能产的本质，例如对“人怂志短”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：

四字表达出现频率越来越高，它到底有哪些能产方式？这些能产方式的能产性又如何？

## 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**表现**（performance）。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”，很可能得到“人/n 怂/a 志/n 短/a”甚至更离谱的结果[[2]](#footnote-2)。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。

另一方面，构式语法理论自提出以来受到越来越多的关注。作为功能学派的理论，它更加鼓励人们去关注语言使用上的细节。然而，在已有的汉语语法的相关研究中，构式只是作为特殊的、不连续的语法结构，对短语规则进行补充，本质上还是形式学派的方法。我们的工作也是一次构式语法理论在汉语中的实践，我们希望借助对四字表达的研究来阐述我们所理解的构式方法，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[[3]](#footnote-3)，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、从单一的结构理论到多元理论的特点，显示出学者对问题的认识逐渐深入、具体。我们首先按术语体系对前人研究进行综述，再总结前人研究的特点和不足。

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。对于一些半自由的格式，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，而仅仅是一种格式套用的行为。换言之，作者认为并立和格式套用是两种能产方式。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。他具体举出四个只包含一个固定字的类固定短语格式，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，显示出二者的差别。作者还观察到一类特殊的现象，如“水落油出”[[4]](#footnote-4)，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”先后顺序。

（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，把类固定短语首先按“型式”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[[5]](#footnote-5)数目，形成了“型式-格式-实例”三层计量体系。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“两字交替显现，两字交替隐含”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“两字邻接显现，两字邻接隐含”的格式远不如前者的能产性强。

（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”。

（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 前人研究的特点和不足

前人研究的特点可以从分析方法和理论指导两方面进行总结，其中也体现了一些不足的地方，下面具体说明。

分析方法上体现从定性到定量的特点

“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。

然而，学者们过于注重建立一个完整的格式系统或子系统（如单论方位词对举格式），而忽略了前人总结的能产方式。例如陆志韦认为，并立的大前提是前后并立，而后人则把格式中的固定字与非固定字分别比较，从而变成了“间隔并立”，且产生了一种先来后到的错觉，例如（王娇，2013）把“大同小异”中的“同”、“异”称为“对举词”，而把“大”、“小”称为“嵌置词”，这样的区分是没有什么道理的；又如文炼、齐沪扬都区分了形成类固定格式的成语和由类固定格式生成的类固定短语，但后人却淡化了二者的差异。后来的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，却忽略了格式是怎么产生的，等于从一个共时的层面（当前语料中存在的实例）去解释一个历时的概念（格式的能产性），这样的定量分析结论是很有限的，有很大的改进空间。

从单一的结构理论指导到多种先进理论指导的特点

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（1956）主要讨论了并立四字格是不是构词格，而文炼（1988）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。如孟祥英引用青年语法学派相关理论阐述“类推”机制，韩立秋则引用Langacker定义的重新分析解释句法结构的演变等。当然，也有许多学者注意到了构式语法理论并引而用之。多种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。

然而，前人对构式语法理论的认识存在一定偏见。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”，仍然是形式学派的分析方法，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。

## 1.4研究方法和特点

1. 前人的定量研究是从格式到实例，即先通过总结前人著作或内省得到一批格式，再在语料库中查找该格式的实例，这样很容易遗漏一些格式。我们则从实例出发提取格式，相比于前人的工作更加的客观、完善。

2. 格式要有一定的实例支持。对于那些表面统计上实例不足的“潜在格式”，我们则用变序、仿造和对举等多种能产方式对其实例进行覆盖，比较其能产性的差异，并利用其形式化特征，帮助发现更多的新表达。

3. 在描述这些能产方式的基础上，本文进一步利用这些能产方式的形式化特征在语料库中进行未登录四字表达的识别工作，具有参考和应用价值。

4. 前人多采用单一的CCL语料，而本文将更加注重语料的实时性，引入了搜狗细胞词库和微博语料等网络语料，进一步扩大四字表达的覆盖面，尤其是对临时的、低频的四字表达进行收集和发现。此外，也注重比较不同语料的差异性，更立体地分析四字表达的能产性。

5. 本文还阐述了对构式语法理论的认知层面的理解，并用之对一个表达可归入多个范畴的现象进行了解释，是对认知理论的实际应用。

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，我们首先概述现代汉语四字表达，包括其定义 （2.1节），其语法地位（2.2节），其规模（2.3节）和我们要研究的几种能产方式（2.4节）。

第三章中，我们在词库分析中定义了几种能产方式并比较了在词库扩展假设中其能产性差异。我们首先介绍搜狗细胞词库的基本情况和两种不同的分析方法（3.1节）；其次，将首先在3.2节具体讨论格式套用能产方式的优缺点，说明其对四字表达的覆盖力的不足，并在后面几节讨论其他几种能产方式，包括变序（3.3节）、仿造（3.4节）和对举（3.5节）。

第四章中，我们在语料库中应用几种能产方式进行新四字表达的发现工作并比较了其识别能力的差异。我们首先介绍语料库和分词工具的大致情况和分析方法（4.1节），其次，分别利用格式套用（4.2节）、变序（4.3节）、仿造（4.4节）和对举（4.5节）进行四字表达的发现工作。

第五章中，我们从基于使用的构式语法视角出发分析了范畴重合的现象。包括对基于使用的理论视角的介绍（5.1节）和对多继承性（5.2.1小节）、能产方式的重合（5.2.2小节）、形式误用（5.2.3小节）的分析。

# 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客观。一方面，虽然大量的“儿”字音节并不独立、“俩”字则为双音节，但在“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，则还要进行消歧工作；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，在以文字而非语音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不完整的语言片段（fragment），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。

在这两点之上，可能还有一点：

c. 形式紧凑凝固

这一点之所以没有当作标准，是因为“紧凑凝固”是一种主观经验，并无统一度量，实际上很难取舍。我们在后文中会用“**典型**”这个饰词来区分形式紧凑凝固和形式松散自由的四字表达，但在此处我们认为，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。”在古代，这些下划线标示的四字表达可能都是松散的自由短语，但今天一看，很容易就从中抽象出“…而……”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。在今天，一些2+2式自由短语的结构、功能前人已经分析得很完善了，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

非要给四字表达一个整体性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一般意义上的短语，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”像是成语词，而“家禽野畜”则又像是并列短语了。

我们认为，以上的看法都有道理。不能取得一致，是因为真实语言是不断发展变化的。“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，如马国凡等一些学者主张“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”。

我们将在第五章中具体讨论构式语法的视角。

在形式语法视角下，我们认为四字表达偏向词的性质，但不应该收入词典中[[6]](#footnote-6)。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[[7]](#footnote-7)，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”[[8]](#footnote-8)的程度。

a, b, c, d四点都是在前人的基础中概括提炼出来的，其中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，我们只强调把这类表达看作词的合理性。

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。实际上，a, b, c, d四点分别代表着四种能产方式。在中文信息处理中，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[[9]](#footnote-9)的语料规模中，字数为四的小句数量占到670万条，占到总量的三成[[10]](#footnote-10)。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，四字格式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。例如：雾霾来的时候，所有发/wù//mái/音的成语都有了新的形式，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利用仿造能产方式构建新表达的例子。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”[[11]](#footnote-11)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”、“非黄勿护”、“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

2.4.1 四字格式的套用

本文所谓四字格式（后文简称格式），即既包含固定字，又包含空位可填入其他汉字的四字结构，如单固定字格式“……之…”，双固定字格式“一…之…”，以及三固定字格式“一…之见”[[12]](#footnote-12)等等。与四字表达类似，我们也用“典型”与否来描述格式的被接受程度，一般地，我们称具有十个以上实例的格式为典型格式。

格式与一般的语法结构不太一样，前者的子项不一定处于同一层级。例如“的”字定中结构可以用重写规则描述为np → np1 的 np2，其中np1、“的”和np2是在同一层级上的，但“一…之…”的内部结构为 [np [np 一 …] 之 …]，“一”和前一个空位处于同一层级，而“之”和后一个空位则处于更高一个层级。值得一提的是，受到韵律的影响，人们可能会对“一…之…”作出 [ [一 …] [之 …] ] 的分析，从而认为它具有与“千…万…”相同的结构层级，但这是错误的。无论前人定义的“待嵌格式”或“框式结构”，其内部结构都是可以忽略的，人们往往强调其整体意义。因此，格式的概念并不像是“syntactic structure”，而更偏向“pattern”。

前人多数把格式的套用看作一种仿造，但我们还是把格式套用当作一种有别于仿造的能产方式。一方面，正如李传军区分的“语例仿造”和“格式仿造”，有些仿造的结果还不足以形成格式，如“水落油出”不足以形成“水落…出”；而有一些实例很多、也都很为人熟知的格式，在生成新的实例时很难说它到底仿造了哪个已知的实例。另一方面，格式不仅与仿造有关，与变序和对举也有很大关系，我们将在3.2.3小节中详谈。总之，我们认为格式套用是一种独立的，与下面定义的仿造、变序和对举地位相同的能产方式，当人们想要文绉绉地表达数量众多的时候，他们就会想到“千…万…”这个格式并填补其中的两个中心语空位。

2.4.2 本文新定义的其他几种能产方式

我们把2.2节提到的四种能产方式总结为紧缩、变序、仿造、对举。它们在前人的文章中都有提及。其中，① 四字表达形式紧凑是所有学者的共识，而孟祥英进一步称其为“经济机制”。② 两个四字表达形式上互为变序，这个在陆志韦和李传军文章中都有提及，陆文关注的是由对举格（即并立格）变序成非对举格，如 [开天辟地，开辟天地] ，而李文则关注两个对举表达间的变序，如 [越生越穷，越穷越生] 。③ 仿造最先由文炼提出，他所谓的仿造是那种必须依赖原表达的实例，如“水落油出”，而后人则把标准放宽，如李传军、孟祥英文章中都用“类推机制”来概称“从实例或格式推衍”这种能产方式。④ 对举最早见于陆志韦文，他特别区分了对举（即并立）和格式套用两种能产方式，以说明对举不同于格式套用之处；另外，以冯胜利“音缀说”为代表，许多学者也用“扩展”来描述把双字词通过对举的形式变成四字表达的过程。

以上四种能产方式，紧缩能产方式蕴含在其他三种能产方式之内，虽然它能够独自作为一种能产方式，来解释最近出现的一批网络热词，如“人艰不拆”、“十动然拒”，然而如果既没有其他能产方式辅助，频率又不突出，很难识别紧缩的实例，因为形式上的特征实在太少。所以对于紧缩能产方式本身，我们并不作太多的分析。

此外，我们还将在后续分析中提到一类特殊的能产方式，我们称之为误用能产方式。我们并不把它与上面的能产方式相提并论，因为从形式上看，仿造和变序都有可能是误用造成的，其中有的误用已经凝固了，有的还没有凝固，使用频率上存在差异。因此，我们把误用能产方式看作是与以上能产方式正交的一类特殊能产方式，并会在5.4节专门分析。

# 第三章　四种能产方式在词库中的描述和比较

## 3.1 概述

3.1.1 定义

词库是信息处理中的概念。它收录的词条并不一定有完整的释义，但比较齐全，覆盖面比较大。词库是中文信息处理中自动分词任务的辅助工具，也适合用来进行相关的统计工作。

细胞词库（cell dictionary），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”[[13]](#footnote-13)，是由网络用户自行上传提供的。相对于传统的汉语词典，它们具有词项繁杂、新鲜度高的实时性特点，更容易发现新词产生的方式。

我们整理了搜狗细胞词库中所有的成语词库[[14]](#footnote-14)。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，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，它们构成本文分析用到的词库集。下文如无特别注明，“词库”均指称四字表达词库。

3.1.2 分析方法

不同的能产方式需要用到不同的分析方法。要描述词库中存在哪些格式，应从整体上对格式进行提取，这可以看作一种静态的分析方法；而若要描述仿造能产方式，则必须区分孰先孰后，故需要模拟一种动态的环境。变序和对举能产方式，从静态或动态分析均可，如果用动态方法分析，则可以比较它们与仿造能产方式的能产性高低。

前提：词库差异

在描述静态和动态方法之前，我们首先要提及词库差异。不同词库收录的词条数量存在差异，而我们希望了解这些差异的程度，以便决定如何选取最合适的词库进行静态或动态分析。

我们用“d”（意为dictionary）加上其词条数量代表一个具体的词库，其中d46238和d49218分别对应搜狗官方推荐成语词库和盘古分词词典，我们给它们加上特殊的后缀标识。不同词库间的差异见图3.1。

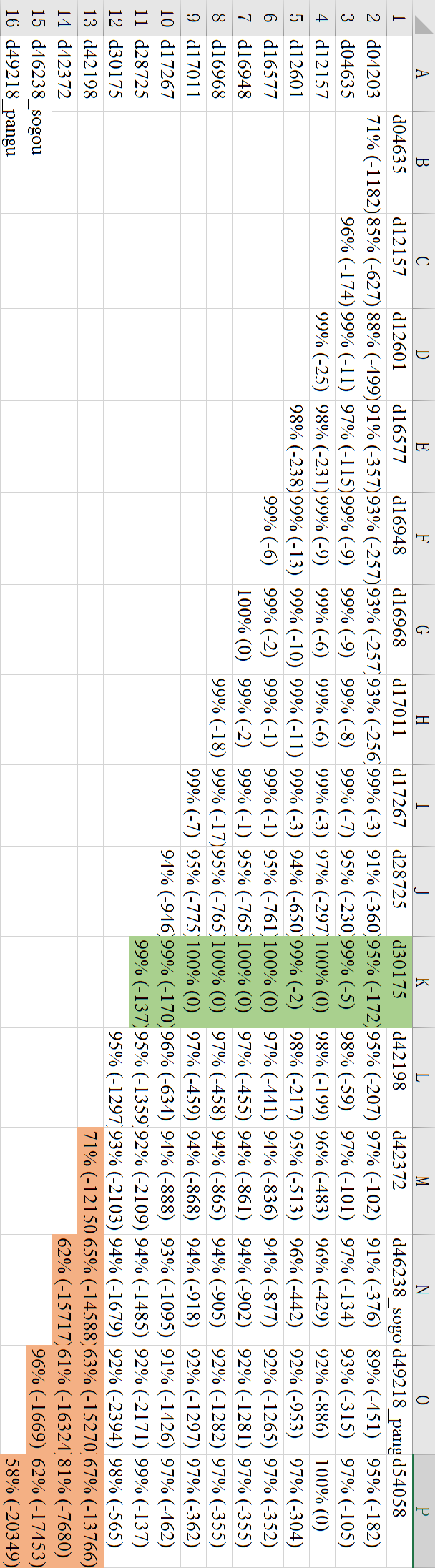
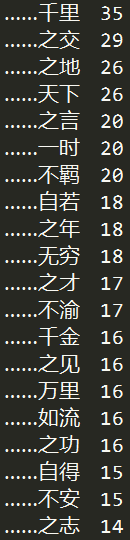


图3.1 词库差异比较

图3.2 新旧格式统计对比

（左为小词库，右为大词库）

“……CD”型格式

“AB……”型格式

该Excel表中每一行代表了一个较小的词库被其他较大的词库包含的比例，括号内的负数则是未包含的条目数量。例如B2格表示“d04203词库中有71%的词条被d04635词库包含，而有1182条未被包含”。对应地，该表中每一列都表示一个较大的词库包含其他较小的词库的比例。

我们首先看第K列，该列显示，d30175词库对较小的词库包含程度非常高，除了d4203外均在99%以上；其次看K列左侧的三角形区域（B2-J2-J10区域），该区域显示，条目数量小于30175的词库间差异都不大，尤其是C3-I3-I9区域，基本在99%左右（d4203的被包含程度和d28725的包含程度要稍微低一点）；最后我们看K列右下方的三角形区域（M13-P13-P16区域），该区域显示，条目数量大于30175的词库间差异几乎都很大，只有搜狗官方推荐成语词库和盘古分词词典间的差异稍微小一点。另外通过第M、N、O、P列可以发现这四个大词库对较小词库的包含情况也不是很好。

这给了我们两个提示：1. 在静态分析的时候单用d30175的表现可能比用大词库更好，因为比d30175大的词库间的差异很大，很可能出于工程目的而各自混入了大量四字短语；2. 在动态分析的时候应该分两个阶段，规模小于等于d30175的词库可以直接两两比较，而大于d30175的则应该统一与d30175比较，以减小词库差异带来的损失。

静态方法：单一词库法

静态方法很简单，即对单一词库进行相关统计分析。单一词库既可以在16个词库中选择，也可以把16个词库合并为一个大词库。

在格式分析时，我们最初的想法是将16个词库合并为一个大词库，它对四字表达的覆盖程度应该更广更全面一些。然而我们发现大词库中包含着许多2+2式四字短语，它们按前两字或按后两字聚类的结果就是一个双字词（见图3.2），这样的格式收录的意义不大，因为这些四字短语用普通的短语规则就能很好地处理。我们最终选择了d30175词库进行了提取格式的工作，它有效地避免了四字短语格式的出现。

而在变序分析时，由于变序分析同样适用于四字短语，不像格式分析那样受到干扰，故为了更全面地分析变序现象，我们把16个词库合并为一个大词库作为统计对象。

动态方法：词库扩展法

我们设计的动态方法又称为词库扩展方法，它基于词库规模的不同。在一般认识中，较大的词库应该蕴含较小的词库[[15]](#footnote-15)，且规模越小的词库中词条也越具有典型性[[16]](#footnote-16)，而大词库与小词库的差集[[17]](#footnote-17)则要临时一些、边缘一些。在这样的认识下，我们便可以假设较大的词库是由较小的词库派生扩展而来，从而获得了一个先后关系，依据该关系，我们可以动态地分析各种能产方式的能产性高低。

## 3.2 词库中的格式

孟祥英研究的240个格式来源为《现代汉语词典》、姜德梧编著的《汉语四字格词典》和张斌主编的《新编现代汉语》，她只研究了“A…C…”和“…B…D”两种型式；李传军虽然没有说明其格式来源，但通过与孟祥英的格式的比较，发现在以上两种型式下的格式基本一致[[18]](#footnote-18)，加之他亦参考了姜德梧和张斌的以上书作，可以断定他研究的格式的主要来源与孟祥英类似。

相比于前人从既有词典、教材中获取格式，继而统计实例的分析方式，从词库实例中抽取格式的方式更加客观全面。我们将用实例说明，前人有哪些格式没有收集、有哪些格式收集得不好，也将讨论格式在概括四字表达能产方式的不足之处。

3.2.1 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。表3.1给出了不同型式、不同实例数量下的**潜在格式**[[19]](#footnote-19)数量分布，其中行坐标为不同实例数量范围下的潜在格式数量，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6个”。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ≥100 | 99~50 | 49~20 | 19~10 | 9~5 | 4 | 3 | 2 |
| …B…D | 1 | 4 | 33 | 87 | 383 | 335 | 803 | 2568 |
| A…C… | 1 | 7 | 27 | 118 | 366 | 322 | 820 | 2656 |
| AB…… | 0 | 0 | 6 | 72 | 520 | 483 | 1166 | 3270 |
| ……CD | 0 | 0 | 7 | 66 | 471 | 477 | 999 | 3101 |
| …BC… | 0 | 1 | 4 | 35 | 260 | 339 | 846 | 2870 |
| A……D | 0 | 0 | 0 | 3 | 141 | 261 | 771 | 3155 |
| ABC… | 0 | 0 | 0 | 2 | 54 | 83 | 372 | 1903 |
| AB…D | 0 | 0 | 0 | 0 | 33 | 83 | 290 | 1937 |
| A…CD | 0 | 0 | 0 | 1 | 16 | 47 | 229 | 1628 |
| …BCD | 0 | 0 | 0 | 0 | 37 | 77 | 279 | 1669 |

表3.1 不同型式、不同实例数量下的潜在格式数量分布

按照“例不十，不立格”的方法，我们共抽取格式475个[[20]](#footnote-20)。与李传军相比，我们在“A…C…”（即其所谓“偶位内嵌”）型式的格式数量要少于他的结论，而在其他型式的格式数量上要多于他，尤其是“AB……”、“……CD”和“…BC…”上。

3.2.2 前人未收录的格式的分析

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 |
| 1）…B…D | …地…天（31例），…心…力（24例），…世…俗（27例） |
| 2）A…C… | 百…之…（14例），柳…花…（27例），千…一…（27例） |
| 3）AB…… | 不知……（40例），一言……（14例），进退……（47例） |
| 4）……CD | ……之地（26例），……不渝（17例），……分明（10例） |
| 5）…BC… | …而不…（61例），…人之…（25例），…不容…（14例） |
| 6）ABC… | 无所不…（21例），不知所…（14例），一无所…（13例） |

表3.2 前人未收录的（潜在）格式举例

1）…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，是一种主观上的不一致性。实际上，动宾式对举如“动地惊天”、偏正式对举如“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，正所谓心意相通、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。

另一方面，这类名字[[21]](#footnote-21)对举格式的能产性却是极其有限的，因为所能描述的对象基本被限定住了，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，还远远达不到一个格式所应具备的自由度。

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。就实例的数量和质量来说，“…世…俗”不比“…心…意”差太多。但由于前人更多的是通过内省的方式来寻找格式，很可能没有注意到这个格式的存在。

2）A…C…

1. “百…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，这也是一种格式判别上的不一致性。实际上，“百世之师”、“百年之好”都是很常见的表达。类似地，“万全之策”、“万物之灵”也为人熟知，而“万…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字对举格式的丰富（不仅能实用，还能虚用），另一方面也说明了应用场景的极其有限。

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，这再一次反映出不一致性：在这里，“千钧一发”、“千篇一律”乃至“千虑一失”都是“优势构象”[[22]](#footnote-22)，“一…千…”不再占有绝对的优势地位。

3）AB…… & 4）……CD

1. “不知……”、“……之地”、“一言……”反映出了多继承性问题。对于“不知其实”，它是来自“不…其…”，还是“不知……”，还是“不知其…”？对于“不毛之地”，是来自“不…之…”，还是“……之地”？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，以上三个“A…C…”型格式的实例数量不少，但它并不比相关的“AB……”或“……CD”型格式更有意义。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用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”型格式，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；“……之地”虽有“不毛之地”、“一席之地”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，“……之地”就是“……的地方”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。

2. “进退……”、“……不渝”、“……分明”则反映出一二字对举这种特殊的部分对举格式的能产性。其中“进退……”是一二字固定的部分对举格式，而“……不渝”、“……分明”则可以分别向空缺部分填入两个近义或反义对举字构成四字表达，如“忠贞不渝”、“黑白分明”。我们在3.3节还要提到这种部分对举的实例。

5）…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”。我们主要讨论后一种结构。这种结构是能写出释义模板：“v别人的n”，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：前人都把“一…之…”列作格式，但几乎所有人遇到“…人之…”的第一反应都会怀疑甚至拒绝其成为一个格式。事实上前者的两个固定字在句法结构中离得更远[[23]](#footnote-23)，且释义模板更加抽象。再比如“…而不…”格式，“而”、“不”严格来说同样不处于一个完整的子结构中[[24]](#footnote-24)，但人们会非常自然地将之看作一个整体——对于“…人之…”则不然。我们似乎只能把这种不一致性归咎为“语感”。

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。

6）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。

3.2.3 前人收录的格式的分析

除了未收录的格式外，前人研究中也有一些格式的收录是值得商榷的。我们举本节开头包含“心”字的七个格式进行讨论。

|  |  |
| --- | --- |
| 格式形式 | 举例 |
| …心…胆 | 倾心吐胆,剖心析胆,同心共胆,同心合胆,吐心吐胆,开心见胆,忠心赤胆,悬心吊胆,惊心吊胆,惊心悼胆,惊心破胆,惊心裂胆,提心吊胆,放心托胆,熊心豹胆,琴心剑胆,碎心裂胆,铜心铁胆 |
| …心…肺 | 刿心刳肺,狼心狗肺,驴心狗肺,鼠心狼肺 |
| 心…气… | 心平气和,心平气定,心浮气盛,心浮气粗,心浮气躁,心粗气浮,心高气傲 |
| 心…神… | 心到神知,心弛神往,心往神驰,心怡神旷,心悦神怡,心旷神怡,心旷神恬,心旷神愉,心旷神飞,心照神交,心荡神怡,心荡神摇,心荡神迷,心醉神迷,心领神会,心领神悟,心驰神往 |
| 心…手… | 心慈手软,心慕手追,心摹手追,心灵手巧,心狠手毒,心狠手辣,心辣手狠,心闲手敏 |
| 心…意… | 心乔意怯,心低意沮,心劳意冗,心劳意攘,心回意转,心孤意怯,心忙意乱,心忙意急,心慌意乱,心慌意急,心慵意懒,心满意得,心满意足,心灰意冷,心灰意懒,心灰意败,心烦意乱,心烦意冗,心猿意马,心领意会 |
| …心…意 | 一心一意,三心两意,三心二意,专心一意,二心三意,二心两意,五心六意,低心下意,依心像意,像心像意,像心如意,像心称意,像心适意,全心全意,可心如意,同心一意,同心合意,回心转意,好心好意,安心乐意,安心落意,实心实意,快心满意,快心遂意,有心有意,正心诚意,歪心邪意,灰心丧意,用心用意,真心实意,真心真意,真心诚意,称心如意,称心快意,称心满意,粗心大意,诚心实意,诚心正意,诚心诚意,趁心像意,趁心如意,输心服意,逞心如意,遂心如意,遂心快意,遂心满意 |

表3.4 前人收录的格式举例

与仿造能产方式的重叠

以上例子中有一类有争议的现象，例如 [惊心吊胆，惊心悼胆，惊心破胆，惊心裂胆] ，后三个表达的典型程度显然不如第一个，乍看之下像是由第一个仿造而来。虽然没有确切的证据说明这种仿造关系，但如果这样的例子都能分开计数，对于格式实例数量的说服力显然大大降低，因为这种例子显然可以造出太多，以至于随便一个格式都能够轻松拥有多个实例，而这与我们的认知是相悖的。

我们虽然不知道李传军和孟祥英都找到了哪些实例，但我们相信这些格式所拥有的典型实例数量是及其有限的，因为它们都是围绕着某一类具体的名物进行描述的，尤其是这些实例都包含了两个意义相似的名字，这实际上进一步限制了其描述范围。而且能够被描述的场景几乎也都已被描述了，几乎不可能再产生什么创造性的新实例，而剩下的就只是对原有实例进行的一些细节改动，以增加新鲜度——而这个显然更偏向仿造能产方式的功能。

值得一提的是，“惊心悼胆”这个表达不像是一个语义上合理的表达。关于它的一个合理的猜测是它从“惊心掉胆”而来，而后者又从“惊心吊胆”而来。同样地，我们没有确切的证据，但我们在更大的词库（d49218\_pangu及d54058）中找到了“惊心掉胆”这个表达，它虽然也不常见，但至少能在语义上站得住脚，“掉”字可以类比“失魂落魄”中的“落”字，是可以说得通的。

我们把这两个不常见的表达称为**形式误用**，更具体地，把“惊心掉胆”称为音近误用，把“惊心悼胆”称为（相对于“惊心掉胆”）的形近误用。形式误用有一定的主观性，比如“惊心掉胆”，我们无法说造出这个表达的人确实是用错了——因为它语义上确有合理性，可能人家确实是有意识地进行仿造；又比如2.3节所举的“十面霾伏”，这种句法语义上皆不甚合理的表达却更像是精致的仿造。但是，它多数情况下确实是误用。此外，还有一种误用为近义误用，比如“惊心悬胆”，它是不是真的误用就很难判断了，毕竟语义上是合理的。

与变序能产方式的重叠

我们的第二个问题是，为什么只有“心意”横跨“A…C…”和“…B…D”两个格式？一方面，在我们的统计中，“…心…气”有11例，“心…胆…”有10例，都应该算作格式；另一方面，甚至还有“胆…心…”（11例）这样的“C…A…”型格式[[25]](#footnote-25)。它们未被收录可能是由于语料的不完备性造成的，但更可能是由作者主观上的倦怠这种不一致性造成的。

我们认为，如果收了“…心…胆”，就应该收“心…胆…”和“胆…心…”。这三个格式反映出格式与变序间的关系：“A…C…”既可以变序为“…A…C”，也可以变序为“C…A…”，甚至可以变序为“…C…A”，虽然统计结果中“…胆…心”只有8例，但把表3.4中“…心…胆”的所有实例都前后变换顺序（例如“倾心吐胆”变换为“吐胆倾心”），新的表达都是能说的。

与对举能产方式的重叠

我们的第三个问题是，还有多少“A…C…”或“…B…D”型格式没有被发现并收录？我们的答案是很多很多。因为这类格式中有一大类是对举格式，而对举格式中又有一大类是近义对举，说白了就是把语义相近结构相同的两个双字词放在一起。如果上面七个格式以这样的实例都能算作格式，那么要“创造”一个格式实在不是难事。在表3.1中可以看到，具有5-9个实例的“A…C…”和“…B…D”型格式数量庞大，它们都具有成为典型格式的潜力。

3.2.4 小结

前人利用了语料的直观和大量特点对格式进行了定量考察，但存在不一致、不充分和有限性等缺点。

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…”；另外许多格式结构工整，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，这是从格式到实例这种内省方法不充分的地方。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。此外，“A…C…”和“…B…D”型格式数量庞大，但绝大多数的能产性是有限的。

不充分和不一致的缺点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，仍然存在不充分的地方；另一方面，一些四字表达并不十分典型，工程师们出于更快速、准确地分词这个工程目的而将它们收录进来，而大多数人对它们并无整体的印象。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。

此外，我们关注到词库中存在多继承性。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。最主要的一点是不一致性，比如我们在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

说“千山万壑”、“千沟万壑”是对举格式的实例没有问题，但“叠”跟“壑”不是对举的关系，看作从“千…万…”而来不妥。就“千…万壑”而言，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？

总之，格式对四字表达的覆盖力是有限的，既然不能把表3.1中所有潜在格式都看作格式，我们就应该寻找更多能产方式去覆盖这些不在格式中的实例。

## 3.3 词库中的变序

互为变序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。本节将对大词库中的整体情况进行分析，在3.4节中我们将从词库扩展的角度对比变序与仿造和对举间的能产性高低。

3.3.1 总体情况

表3.5展示了大词库共93050个四字表达间的变序关系。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变序类型 | 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 |
| 变序类型 | **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 大词库中的变序总览

其中“数词”表示互为变序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序讨论，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”看作特殊的变序类型，而把剩余的变序类型称为基本变序类型。由于“叠字”和“多余二词组”又可以两两归入到基本变序类型中，下文只对基本变序类型进行讨论。

3.3.2 基本变序类型举例及分析

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，把两样事物刻意地放在一起说，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，则存在着“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”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[[26]](#footnote-26)？

此外，在二四交换类型中也有形式误用，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[[27]](#footnote-27)；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例中的两个表达产生的时间都很早，甚至“每下愈况”更古老些[[28]](#footnote-28)，但今人已不用“每下愈况”的本义，而只用“每况愈下”的意义，今文出现的“每下愈况”，多为“每况愈下”之义，也应看作是一种误用。

一三交换

一三交换与二四交换情况类似，对等对举共161组，不对等对举共24组。对等对举中也包含字面不对等而意义周遍的如 [出文入武，入文出武] 、 [走南闯北，闯南走北] ，不对等对举中也包含各种误用，如 [兔死狐悲，狐死兔悲] 、 [薪尽火传，火尽薪传] 、 [眼高手低，手高眼低] 、 [白山黑水，黑山白水] 等，其中“兔死狐悲”本来是对等对举，原语境中狐狸兔子是同阵营的，但今人因“狐为兔敌”而将之曲解为“虚情假意”，变成了不对等对举，从而“狐死兔悲”就显得比较莫名其妙，甚至有人生硬地解释为“没有天敌而倦怠”，这也体现了形式误用能产方式和仿造能产方式的重叠性。

此外，“火尽薪传”是明末寓山居士的“巧讹”，把积极的原意故意误用为消极的新义，比喻前功尽弃，一切皆空。“手高眼低”也像是故意误用，比喻贪得而不识货，空有技艺而无大局观。“黑山白水”则见仁见智了，因为“白山黑水”一指“长白山和黑龙江”，至于“黑山白水”指的什么地方不得而知，但二者可能同有穷山恶水、草莽乱世之意。

一二交换

一二交换主要有两种类型：（一二）对举交换和（一二）非对举交换[[29]](#footnote-29)。对举交换即交换二字为对举成分，共84组，如 [往来如梭，来往如梭] 、 [死生有命，生死有命] 。非对举交换共66例，又可分几类：a. 主谓（偏正）交换，如 [叶落归根，落叶归根] 、 [胆大包天，大胆包天] ，交换后多变成偏正结构，也有些偏正结构整体变成了一个谓语，如 [血流成河，流血成河] 、 [心甘情愿，甘心情愿] ；b. 述宾交换，如 [琢玉成器，玉琢成器] ；c. 虚字，如 [一决雌雄，决一雌雄] 、 [自不量力，不自量力] ；d. 其他，如 [七乱八糟，乱七八糟]。

主谓交换与偏正交换重合，上面已有举例，然而并不是说主谓一定是原形，如 [覆水难收，水覆难收] ，就应该看作偏正为原形；主谓交换与述宾交换有时也难区分，例如 [陷身囹圄，身陷囹圄] 、 [云开见日，开云见日] 。我们并不刻意区分这些非对举交换小类，它们中占优势的一方很多都是约定俗成的，并没有什么确凿的句法上的证据证明其的确优于另一方。

所有一二交换的例子中，每组的两个四字表达的意义是相同的。

三四交换

三四交换同样分（三四）对举交换和（三四）非对举交换，但非对举交换数量要多，共68组，对举交换则要少，共48组。一大原因是受到否定句的影响，如 [万无一失，万无失一] 、[岁不与我，岁不我与] 。此外，有几组四字表达意义并不相同，如 [世界名著，世界著名] 、 [人文学科，人文科学] 、 [有效字数，有效数字] 。这些都是今天的表达。其余例子中，每组的两个表达意义也都相同。

一四交换

一四交换全部属于（一四）不对等对举，其中多“不”、“而”等虚字，如 [华而不实，实而不华] 。

二三交换

二三交换绝大部分属于上文“二四交换”中提到的对等对举，如 [拘文牵义，拘牵文义] ，其中又有一些是许多人不能仔细区分的，如 [飞短流长，飞流短长] 、 [进荣退辱，进退荣辱] ，后两者字面上更和谐，但也更难揣测原意，因此误用也不在少数，如CCL现代汉语语料库中可查到1994年《人民日报》有这么一句：“但张瑞芳想的并不是个人的进退荣辱，而是艺术的发展与进步”，这与原意是有差别的。

此外，还有一些四字表达也能发生二三交换，如带“可”、“能”字表达：[大可有为，大有可为] 、 [力所能及，力能所及] ；或带“如”字表达：[性如烈火，性烈如火] 。

所有二三交换的例子中，每组的两个表达意义都是相同的。

前后交换

前后交换的例子占了所有交换例子的六成以上。其中绝大多数都是对举，也有部分如 [满目疮痍，疮痍满目] 、 [蜻蜓点水，点水蜻蜓] 这样的非对举例子。

当然，由于大量现代汉语四字短语的混入，如 [学习理论，理论学习] 这样的交换也比比皆是。把这些例子去掉，其余每组的两个表达意义都是相同的。

逆序交换

所有逆序交换的例子都属于对举结构，且每组两个表达意义相同。

前后分别交换

所有前后分别交换的例子都属于对举结构，且每组两个表达意义相同。

其他二词组

剩余的包含两个四字表达的例子中，情况比较复杂，如 [千姿万态，姿态万千] ，后者可以看作前者先后经历二三交换、一二交换、前后交换得来，而意义一直保持一致；[不言而信，言而不信] 则没有办法等价于其他多个交换的组合，且其意义发生了变化。但总体说来，两个表达意义相同的例子有90组，不相同的只有4组，前者占了绝大多数。

3.3.3 小结

变序在四字表达中比较常见，尤其是对举或部分对举的四字表达。互为变序的例子中，意义相同的例子占大多数，其产生的原因一方面是求新，还有一种可能性就是形式误用。

变序的方式有很多种，而且往往可以同时作用于一个表达。多余二词组虽然没有讨论，但其中的例子全部都属此类：同一组中的所有四字表达意义均相同；在基本变序类型中的许多例子，也还可以发生更多的变序，如 [千姿万态，姿态万千] ，还可以变序为 [万态千姿，万千姿态] 甚至 [千态万姿，万姿千态] 等等。

然而，大词库下的变序分析也有一个孰先孰后的问题。虽然同一组变序中，会存在一个出现频率占优的“优势构象”，但要说由“优势构象”产生了“劣势构象”，似乎还缺少了点证据。我们需要一种更有说服力的假设来区分变序的方向。

## 3.4 词库中的仿造

2.4节和3.2.3小节中已经提到，格式与仿造间存在密切关联。我们可以认为，对于那些已经拥有大量实例、且许多实例已经足够凝固的格式，它们通过格式套用而产生新表达；而对于那些实例既不多又不够凝固的格式，它们的实例中一定有一个比较典型，而其他非典型实例，包括新造实例，皆为仿造该典型实例而成。

我们将在3.4.1节中首先给仿造一个明确的定义，再分析词库扩展中仿造的能产性。

3.4.1 仿造的定义

本文中，我们把仿造进一步限定为只改变一个字而形成新表达的能产方式。限定在只改变一个字，一方面是保证了仿造的可靠性；另一方面，则是由于这类仿造在数量上占有绝对优势。

可靠性

对于四字表达，改变一个字所带来的结构和意义变化远不及改变两个字那么大。例如，在只包含两个实例的“A…C…”和“…B…D”型格式中，可以找到如下例子：人…之…（人中之龙，人情之常），…亭…立（亭亭玉立，渊亭山立），这两个格式的实例间几乎没有相关性，自然也难以归为仿造。但在三固定字格式中，这样的例子是很少的，因为固定的三个字基本已经将语法结构限制住了。

当然，对于“A…C…”和“…B…D”型格式，其最主要的特征还是对举。对于大部分实例数较少的格式，把实例归为对举要比归为仿造可靠得多。

可靠性这一点对于自动识别新表达尤其重要。在词库分析中，被分析的表达都确实是完整的四字表达；但在真实语料中，没有这样一个明确的边界，待分析的表达只不过是一个句子中连续的四字字串，有很大的可能并非一个完整的四字表达。它们与既有的四字表达恰好有两个字相同的概率要远高于有三个字相同的概率，因此把改变字数限制在一个字，也能显著提高识别正确率。

数量优势

把仿造限定在只改变一个字，更重要的原因是因为，改变一个字（即形成三固定字格式）的情况比改变两个字（即形成双固定字格式）要多得多。我们上文已经讨论过，许多格式的实例数量是偏高的，因为其中许多实例并不是典型的实例，而是通过其他典型实例仿造而来，例如“…心…胆”格式中，蕴含了“…心吊胆”、“惊心…胆”和“同心…胆”等格式。（cf. 表3.4）

另一方面，虽然在表3.1中，双固定字（潜在）格式数量比三固定字（潜在）格式看起来要多，但其中大部分都是“伪格式”，例如“指桑说槐”和“指桑骂槐”两个实例，既可以形成“…桑…槐”潜在格式，也可以形成“指桑…槐”甚至“指桑……”、“指……槐”等潜在格式。这四个潜在格式中“…桑…槐”最像一个格式，因为它具有对举的性质，但实际上它并不是一个格式，因为它除了“指桑骂槐”和语义基本相同、只是换了个动字的“指桑说槐”、“说桑骂槐”之外再无其他实例。比较符合认知的是“指桑说槐”仿造“指桑骂槐”而来，而它们在仿造的过程中形成了“指桑…槐”这个潜在格式。然而在统计过程中，“指桑…槐”这个三固定字潜在格式进一步地抽象出了“…桑…槐”、“指桑……”、“指……槐”三个双固定字潜在格式，所有四个潜在格式都只有两个实例，而后三个双固定字潜在格式都是伪格式。

我们对不同型式下仅拥有两个实例、实际为三固定字格式的双固定字格式的比例进行了统计，结果如表3.6所示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 型式 | AB…… | A…C… | A……D | …BC… | …B…D | ……CD |
| 比例 | 1870/3270 | 1688/2656 | 2590/3155 | 2303/2870 | 1730/2568 | 1641/3101 |

表3.6 不同型式下仅拥有两个实例、实际为三固定字格式的双固定字格式的比例

可以看到，大多数仅拥有两个实例的双固定字格式其实都是三固定字格式。

综上，我们认为，虽然一些改变两个字甚至三个字的能产方式也包含了仿造的成分，但出于可靠性及数量优势的考虑，我们还是只把仿造能产方式限定为只改变一个字。

3.4.2 仿造能产方式与其他能产方式的能产性比较

我们用词库扩展的假设分析了仿造与变序、对举三种能产方式的能产性高低。在16个词库中，d16948与d16968、d16968与d17011间的实例数量差距很小，我们因此先人工分析了这两个词库的差异；其他词库，我们则用程序自动进行分析。

人工分析结果

1. d16968相对于d16948增加的20个四字表达非常有特点：它们要么包含“抃”字，要么包含“敲”字，可以猜测是在d16948的基础上定向补充包含此二字的表达而成。

抃[biàn]，《说文》：“拊手也”，鼓掌、击、顿踏的意思。包含“抃”字共4个表达：喜跃抃舞、扛鼎抃牛、抃风舞润、抃风儛润，皆来自古文。古文表达不好评价能产方式，但从古文句法上看，都可以归为对举；

包含“敲”字表达中：

① 敲冰戛玉（戛玉敲冰），敲金戛玉（戛玉敲金）是变序[[30]](#footnote-30)；

② 敲冰索火（敲冰求火），敲榨勒索（敲诈勒索）是仿造，此外“敲膏吸髓|敲骨剥髓|敲骨取髓|敲骨榨髓|敲髓洒膏”五个表达皆可看作从“敲骨吸髓”中仿造而来；

③ “敲山震虎|敲山振虎|敲牛宰马|敲锣打鼓|敲锣放炮|敲锣放砲”为对举，此外还有一个“敲冰玉屑”，也是来自古文，也可以看作“敲冰”对“玉屑”前后对举。

因此，变序能产方式共2例，仿造能产方式共7例，对举能产方式共11例。但是，对举例子中有5例为古文，且“敲山震虎|敲山振虎”、“敲锣放炮|敲锣放砲”、“抃风舞润|抃风儛润”和“敲冰戛玉|敲金戛玉”中也存在仿造关系，实际上仿造的能产性不比对举要低。

2. d17011比d16968多出43例。其中：

① 变序的例子有：作福作威（作威作福），子孝父慈（父慈子孝），心怡神旷（心旷神怡），羞花闭月（闭月羞花），滴水石穿（滴水穿石）；

② 仿造的例子比较多，有只改一个字的：一枕黄粱（一枕黄梁），作奸犯罪（作奸犯科），做人做事（做人做世），做好做恶（做好做歹），养家糊口（养家活口），古调独弹（古调单弹），坐拥书城（坐拥百城），做张做智|做张做致（做张做势），（坐而待旦，坐以待旦），字字珠玉（字字珠玑），尊师贵道（尊师重道），恣意妄行（恣意妄为），恣肆无忌（恣睢无忌），纵横交贯（纵横交错），纵横开阖|纵横开合（纵横捭阖），罪不胜诛（罪不容诛），罪业深重|罪恶深重|罪逆深重（罪孽深重），罪应万死|罪当万死（罪该万死），罪有攸归（责有攸归），行侠好义（行侠仗义），走马到任（走马上任）；

也有改两个字的：纵横天下（纵横……（3例）|……天下（17例）），纵虎出柙|纵虎出匣（纵虎归山），罪责难逃（劫数难逃），舍小取大（舍短取长），尊己卑人（尊古卑今），尊贤使能（选贤举能）；

这些例子中，做人做事（做人做世），做好做恶（做好做歹），做张做智|做张做致（做张做势）也可以看作对举，但由于“事”和“世”既为同音，又有一定语义上的相关性，“恶”、“歹”也是近义字，而“智”、“致”、“势”在一开始就是模糊两可的，故它们看作仿造更合理些；而改动两个字的，确实也都有仿造的倾向，但根据我们的讨论，在统计时不计入仿造能产方式内。

③ 对举的例子有：嘴快舌长，尊年尚齿，纵曲枉直，纵风止燎，走花溜冰，钻坚仰高；

④ 此外，“作小服低”（服低做小）和“罪恶贯盈”（恶贯满盈）可以看作变序+仿造；“作辍无常|夜色阑珊|左右图史”则都是部分对举，其余的则不能归入以上能产方式中。

在这一组词库扩展中，仿造能产性要高于对举能产性，而变序能产性仍然最低。

自动分析结果

在人工分析中存在一个两可的问题，即一个表达可以同时归入几种能产方式中。我们在第五章中还要具体讨论这个问题，现在则按可靠性对不同的能产方式设立一个优先级。

我们认为，变序能产方式的优先级是最高的，其次是仿造能产方式，而对举能产方式则是最低的。这主要是按照固定字多少进行的可靠程度排序，因为固定字越多，代表着句法结构和语义上变化的可能性越小，则新旧表达间的关联程度越高。因此，我们设计了如图3.3所示的算法流程。

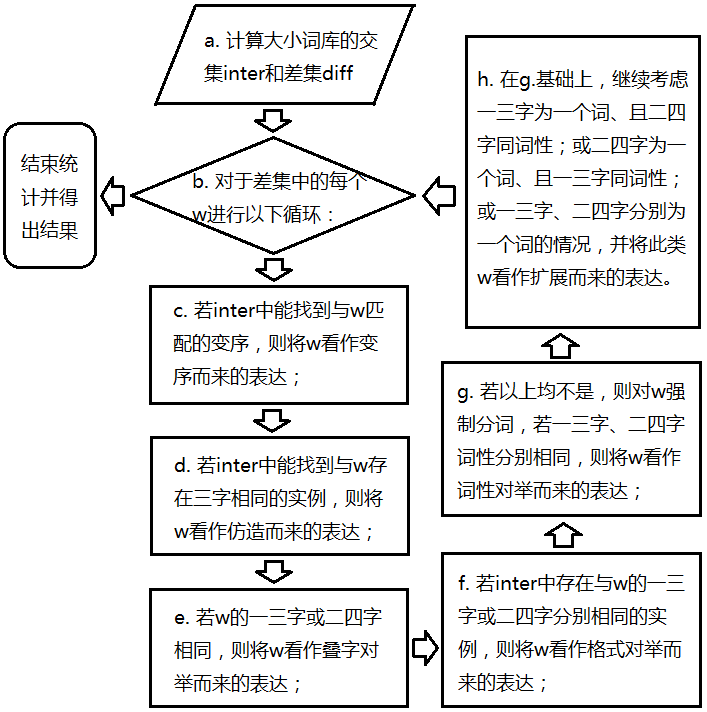


图3.3 词库扩展自动分析流程图

其中我们将对举能产方式进一步按可靠性拆分为多种情况，而并非只按g步所示的词性对举进行统计。这是因为一个字可能对应多种词性，而自动分词软件并不能像人一样清晰地为每一个字匹配上在该表达下最适合的词性，即会产生分析错误，而降低可靠性。要指出的是，f步的“格式对举”，只是一种近似处理，是我们对大量实例较少的“A…C…”和“…B…D”格式总结出来的特点，把所有这两个型式的格式都看作对举格式，一定程度夸大了格式对举的数量，但比例不大。我们认为这种误差应该比g, h步依赖自动分词而进行的词性判断误差要小。

统计的结果如表3.7所示。其中有一列“短语”，指的是能够切分成两个词的四字表达，这种表达应该首先被排除在外，因为它不应计入能产方式的统计中，且会对统计造成影响。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 变序 | 仿造 | 叠字 | 格式 | 词性 | 扩展 | 短语 | 其他 |
| d04203/d04635 | 27 | 84 | 58 | 199 | 111 | 139 | 202 | 356 |
| d04635/d12601 | 193 | 683 | 239 | 1242 | 563 | 403 | 831 | 1383 |
| d12601/d16948 | 420 | 1295 | 108 | 712 | 153 | 107 | 468 | 268 |
| d17011/d17267 | 22 | 116 | 2 | 19 | 3 | 6 | 47 | 9 |
| d17267/d30175 | 1014 | 4297 | 211 | 2465 | 478 | 269 | 1100 | 740 |
| d30175/d42198 | 9 | 339 | 22 | 440 | 431 | 953 | 6299 | 2694 |
| d30175/d42372 | 849 | 3544 | 241 | 1993 | 337 | 288 | 2281 | 1151 |
| d30175/d46238 | 10 | 101 | 119 | 404 | 222 | 413 | 13149 | 1904 |
| d30175/d49218 | 56 | 462 | 227 | 689 | 328 | 657 | 14383 | 2515 |
| d30175/d54058 | 1181 | 5937 | 646 | 3635 | 780 | 545 | 3357 | 2341 |

表3.7 词库扩展分析统计结果

从表3.7中可以看到，仿造能产方式的总体占比是非常大的。在一半的词库扩展比较中（d12601/d16948，d17011/d17267，d17267/d30175，d30175/d42372，d30175/d54058*）*，它的数量比变序和四种对举的总数还要多，更何况数量排名第二的格式对举数量还有一定程度的夸大。

而在另一半的词库扩展比较中，d04203/d04635和d04635/d12601是小词库扩展，两个词库均为假设中的核心词库，其差集中的实例仍具有较高的典型性，对举能产性表现出高于仿造是情有可原的，因为仿造能产性虽然高，但产出的新实例却不一定能凝固下来，例如“水落油出”，应用的场合相当有限，自然不如对举实例那样容易成为典型表达。另一方面，d46238和d49218分别是搜狗和盘古的推荐分词词库，它们的主要目的是加速匹配和提高命中率，因此新收录的很多表达都不典型，例如二者均收录了“一个叫做”的表达，这个表达并非我们定义的四字表达，在其他词库中也都没有出现。此外，这两个词库的短语数量都特别大。因此，它们在d30175的基础上并没有增加许多仿造、对举的例子。

我们最后来看一下仿造和格式套用的能产性比较，虽然我们认为后者不应该置于与前者相同的层面相提并论。表3.8给出了在不同阈值下的格式套用数量。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 仿造 | 格式套用(阈值=5) | 格式套用(阈值=10) |
| d04203/d04635 | 84 | 73 | 14 |
| d04635/d12601 | 683 | 447 | 157 |
| d12601/d16948 | 1295 | 831(6) | 370 |
| d17011/d17267 | 116 | 79 | 32 |
| d17267/d30175 | 4297 | 3021(15) | 1363(4) |
| d30175/d42198 | 339 | 424(4) | 128 |
| d30175/d42372 | 3544 | 3783(50) | 1755(3) |
| d30175/d46238 | 101 | 617 | 188 |
| d30175/d49218 | 462 | 1082(6) | 411 |
| d30175/d54058 | 5937 | 6756(119) | 3451(10) |

表3.8 仿造与不同阈值下的格式套用产生的实例数量比较

阈值即一个格式所需具有的最小实例数量，而括号内数值代表套用三固定字格式产生的实例数量，之所以标示出来是因为这部分实例同时也计入仿造能产方式的统计中。由于这部分实例所占比例很小，不影响仿造与格式套用能产性的直观比较。

可以看到，仿造能产性不亚于格式套用能产性，尤其对格式的实例数量要求比较严格的时候（阈值=10）。另一方面，格式套用的能产性也随着基础词库的扩大而增加，并在3万左右的量级处超过了仿造能产性。这个很容易理解，因为实例多了，自然就有更多的格式达到阈值标准，而匹配到更多实例。但是，我们也应该看到，一个人的词汇量是有限的，而很少有人仅四字表达的识记量就达到3万条。因此我们认为，从形式上来说，格式固然能产，但仿造，包括对举和变序，也都是重要的能产方式，更充分地利用这些能产方式，才能更好地发现更多的四字表达实例。

此外，表3.8还反映出一个现象，就是格式套用的能产数量与仿造相仿，在仿造能产性较低的几组中，格式的套用也相应减少。这既验证了这些词库扩展的异常，又说明了仿造和格式套用的相关性。显然，格式套用与格式对举间也存在更强的相关性。

3.4.3 小结

仿造的能产方式本身并不限于改变的字数。但由于改变一个字产生的新表达数量较多，分歧较小，可靠性较高，因此我们只研究这一类仿造的能产性。

在一般场景中，既有四字表达通过仿造的能产方式生产新四字表达的倾向是很高的，并不亚于对举或格式套用的能产性。

## 3.5 词库中的对举

在本章前面几节中，我们已经涉及到了许多对举能产方式。我们所定义的对举主要是3.4.2小节中提到的词性对举，它满足：① 前后对举；② 一三字、二四字分别对举。分别对举的主要参照就是词性，而前后对举似乎没有太好的形式化检测手段——除非它能发生前后交换。

本节中，我们首先再总结一下对举与其他能产方式间的联系，其次再回顾一下它的能产性强弱。

3.5.1 对举能产方式与其他能产方式的联系

与格式套用的联系

“A…C…”和“…B…D”两个型式下的绝大多数格式都是对举格式。对举很容易形成格式，正如孟祥英所说，待嵌格式的固定字都是基本词汇，越基本，被反复多样化描述的概率就越高，越能形成格式。但我们同时也已指出，名字对举格式的能产性是有限的，就像陆志韦认为并立是一种构词格一样，这种对举现象本身很丰富，但很难成格式，甚至一旦成了格式，还要加倍小心，防止任意地套用而降低对举性。常用的对举格式大部分都是一定程序虚化了的格式，如黄燕旋研究的方位词对举格式。而这样的格式数量则是很有限的。

与变序的联系

一三交换、二四交换、前后分别交换和逆序四种变序所涉及的全部实例和前后交换的大部分实例都是对举实例；二三交换的一方也都是对举实例；一二、三四和一四交换则中既有部分对举（一二对举、三四对举和一四对举）的实例，也有特殊对举的实例。例如“乱七八糟”、“甘心情愿”，它们虽然不满足一三字、二四字分别对举，但却是前后对举的。此外，没有研究的叠字变序显然也全部都是对举实例。

可以说，变序中的绝大多数实例都是对举实例。它们以两种方式进行着变序：① 同功能字交换，如交换（前后结构中的）饰词或谓词；② 结构变换，如从主谓结构变成偏正结构或述宾结构，前者如 [龙神马壮，神龙马壮]，后者如 [云开见日，开云见日]。同功能字交换反映出语义上的相似性；而结构交换则体现出汉语语法的兼容性（例如受事作主语）。

与仿造的联系

对举与仿造间存在很大的重合性，如3.4.2小节人工分析结果中举出的“做人做事”（做人做世），“做好做恶”（做好做歹），“做张做智|做张做致”（做张做势）等。（它们又可以看作格式套用的实例）。有关这三者的重合性，第五章还要进行讨论。

3.5.2 对举能产方式与其他能产方式的能产性比较

在3.4.2小节中，我们已经详细呈现了词库扩展分析方法及其统计结果。在四种对举能产方式中，叠字对举的可靠性最高，但数量最少；其余三种对举都有一定误差，但数量较多。

我们分别用格式对举和词性对举+扩展对举的方式估计了一下d30175的对举表达数量，其中包含在“A…C…”和“…B…D”中的实例总数为21775个，而词性对举加上扩展对举统计总共只发现了10773个实例。这说明格式对举可能高估了对举实例的总数，而词性对举+扩展对举则可能低估了对举实例的总数。

但无论从以上对对举实例总数的估计还是从词库扩展分析中得到的能产性比较来看，对举能产方式都占有着重要地位。实际上，在4.5节中我们将看到，对举表达在真实语料中数量非常庞大，利用它发现的新表达数量在其他能产方式之上。

## 3.6 本章总结

词库分析是本文的重点章节。它的目的是定量地描述2.4节定义的四种能产方式。

首先是前人的研究重点：格式。在3.2节中可以看到，格式不仅数量有限，而且质量也参差不齐：有的格式实例数量多，意义却不大，如“一…之…”；有的格式对举意味明显，但实例数量和质量却值得掂量，如“…心…肺”。格式对四字表达的覆盖力有限，故用格式套用进行四字表达的发现甚至生产，其能力是十分有限的。

变序不改变表达中的字形和字义，只交换了字的顺序，故其形式化特征是很可靠的，除了 [一夫之勇，一勇之夫] 是偶然相同而非变序关系外，其他经形式化特征筛选出的例子间都具有同义或反义的联系，可以看作变序。

仿造限定了四个字中的三个字，且字序不变，意味着新表达与原表达的语法结构也相似，其形式化特征也是可靠的。在词库扩展分析中，仿造而来的表达能产性非常高，这与表3.1节中的“长尾现象”吻合[[31]](#footnote-31)。另外，有相当一部分仿造实例其实是形式误用而来。

对举的形式化特征并不是很可靠，这主要受限于分词工具的准确性，因此我们又分了几个小类。叠字对举其实是格式对举的特例，即由“A…C…”变为“A…A…”，这一形式特征也使得它具有较高的可靠性；而词性对举和扩展对举都依赖分词，它们的数量不及格式对举，可靠性不高，需要进一步加以限制才能应用。

仿造和对举是格式的重要补充。对于那些“例不十”的潜在格式，其中的三固定字格式都可以归入仿造能产方式，而“A…C…”和“…B…D”型格式也有一大部分可以归入对举能产方式——这并不是说超过十个实例的典型格式的实例就不是通过仿造或对举而来的了，只不过这样的实例既可以归入格式又可以归入仿造或对举，总之是有个归属了——我们定义的仿造和对举能产方式，主要是照顾那些没有“格式归属感”的四字表达，让所有的能产方式能覆盖到更多的实例。

# 第四章　利用四种能产方式在语料库中识别新表达

## 4.1 工作概述

识别未登录的四字表达对于中文信息处理中的自动分词任务意义重大。现行自动分词任务基本都要依赖分词词表，像（顾森，2013）那样完全依赖信息熵进行的无指导分词不常见，表现也不好。本文开头已经提及，许多四字表达不典型，具有很强的临时性和低频性，收进词表对分词帮助不大。如果能利用形式化的能产方式辅助识别这些四字表达，将会很好地提高分词任务的准确率。

4.1.1 分词工具的选择

当下比较著名的自动分词工具包括NLPIR平台提供的ICTCLAS[[32]](#footnote-32)和基于条件随机场模型（CRF）的CRF++[[33]](#footnote-33)；前者主要基于分词词典，可以看作静态的分词工具；后者则依赖对训练语料的字序列标注，是动态调整的分词工具。

我们权衡后仍然采用ICTCLAS进行基准的分词工作。因为像变序和对举能产方式，它们很难转化为字序列的表现方式为CRF++提供经验，而格式套用和仿造，也很难完整地标出一个单位的开始（B）和结束（E）。

4.1.2 辅助识别的方法

在ICTCLAS的基础上辅助识别四字表达有两种方式：在分词之前识别和在分词之后识别。前者的优点是一旦识别成功会显著提高句子其他部分的分词的准确率，缺点是一旦识别失败则会影响到其他部分，使整体准确率变得更糟糕。

我们选择了后者的识别策略，特别地，我们仅对ICTCLAS分词结果中的连续单字进行识别，挑选出其中可能存在的四字表达，如1.2节提到的“人/n 怂/vg 志/n 短/a”。这样的好处是不影响分词工具的整体分词策略——例如它也可应用于加入了上下文信息的CRF++的分词结果上，同时对句子的其他部分影响较小。

4.1.3 语料的选择和处理

我们选择了两份语料：CCL现代汉语语料库和微博语料。前者虽然也包含了部分网络语料，但篇幅较少，更多的是经典语料如《人民日报》、新华社文章等等。后者则是典型的网络语料。两份语料规模相近，其中CCL现代汉语语料库规模1.1千兆字节（Gb），微博语料在删除了特殊字符（包括标点符号、转载、点名、表情、网页链接等所有非汉字字符）后规模为566兆字节（Mb）。但行文风格有较大差别。我们希望在这种不同的语境中检验我们提出的几种能产方式的存在性和能产性。

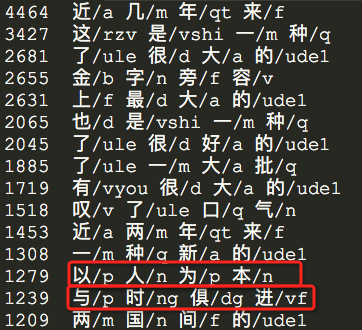
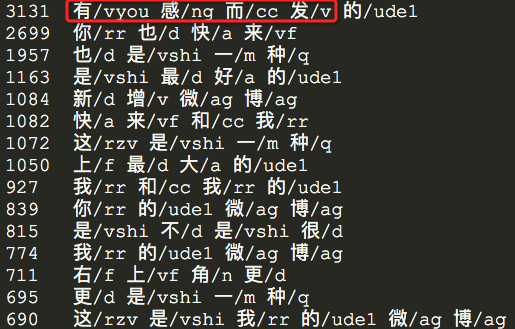
我们对两份语料都进行了分词，并从结果中整理出连续的单字序列，按其出现的次数逆序排序。图4.1展示了两份语料中出现次数最多的15个单字序列。

图4.1 出现次数最多的15个单字序列（左图：CCL现代汉语语料；右图：微博语料）

其中每个表达左侧的数字代表其在语料中出现的总次数，而以红圈圈出的就是待发现的四字表达。可以看出，在高频的序列中，四字表达出现的概率并不大，而且多包含虚字。这些表达由于频率很高适合收入词库，而我们要找的是那些具有临时性和低频性的不典型的四字表达。

CCL现代汉语语料（以下简称CCL语料）中共发现不同的单字序列420万条，微博语料中共发现不同的单字序列127万条，这其中的差异要大于两个语料库规模上的差异[[34]](#footnote-34)，这是因为微博语料中存在大量转发性微博，其原始微博内容都是一致的。此外，单字序列中包含的单字数量可能大于4，这意味着潜在的四字表达数量还要更多[[35]](#footnote-35)。我们接下来就要利用不同的能产方式的形式化特征来从这些单字序列中发现更多的四字表达。

## 4.2 利用格式套用发现新表达

格式套用采用d30175提取的拥有10个及以上实例的格式。其中三固定字格式只有“一无所…”、“无所不…”和“一…千金”三个，我们放在4.4节仿造能产方式中统计，剩余的都是双固定字格式。用格式来发现实例的算法很简单，用一个简单的正则表达式就能完成。其主要困难在于划定正确的表达边界。

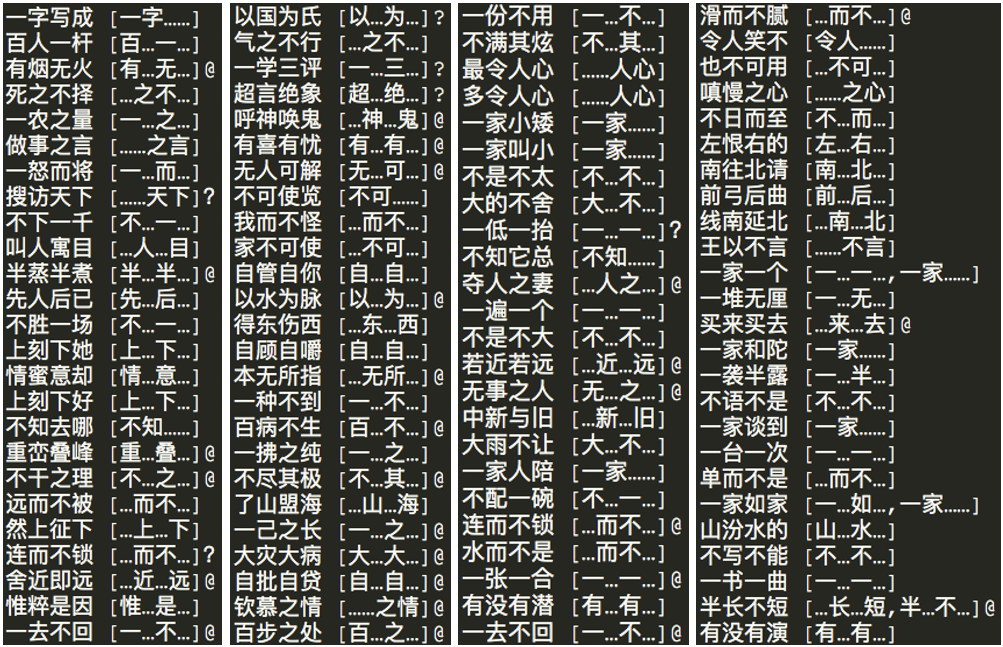
在CCL语料中，共获得73408个不同的潜在四字表达实例；在微博语料中，总共获得18197个不同的潜在四字表达实例。其中确有不少四字表达，例如CCL语料中的“有烟无火（有…无…）”、“半蒸半煮（半…半…）”，微博语料中的“大涨小跌（大…小…）”、“夺人之妻（…人之…）”等等，然而错误率也很高。我们各自随机选取50个实例进行校对（如图4.2），发现CCL语料中只有18个可以被看作四字表达[[36]](#footnote-36)，还有5个疑似四字表达，例如“连而不锁”在CCL语料中确有例子[[37]](#footnote-37)，但乍看之下又比较难理解，倒像是“……连|而不|锁……”的错误截取。另外，微博语料中正确和疑似的四字表达分别只有9例和1例。

图4.2 随机选取50个潜在四字表达中正确和疑似的例子

（左二列：CCL语料，右二列：微博语料；@代表正确的例子，？代表疑似的例子）

用格式套用的能产方式去发现新的四字表达需要在划定表达边界上花大力气。相比于仿造能产方式，前者光靠两个固定字很难约束其整体边界。另一方面，可能能从格式型式、固定字搭配或多继承性上找到突破口，例如“AB……”和“……CD”型式在正确率上远不如“A…C…”和“…B…D”，因为前者的固定字作为一个连续体更可能与外部成分发生联系，也更难约束空位一边的边界；又比如包含虚字或常用字的格式可能更容易出错，因为它们的搭配情况更多更灵活。但目前看来找到一个稳定突破口的困难很大，可能更多的优化策略是概率的、工程性质的，没有什么特别的语言学道理。

## 4.3 利用变序发现新表达

变序以d30175中的所有实例为参照。由于大部分实例本身在词库中就存在变序，因此想要找到新的变序实例就比较困难。但变序能产方式发现新实例的准确率很高，这说明我们在词库扩展分析中依据固定字多少判断可靠性是正确的选择。

CCL语料中，新的变序实例type数为1642个，token数有3166个[[38]](#footnote-38)；微博语料中，新的变序实例为292个，token数有497个。

新的变序实例基本都是四字表达，出错的几个例子几乎都是边界问题。例如（括号内为词库中的既有表达）：

**CCL语料（前500例中的误例）：**

了无休无（无了无休，无休无了）

一报还一（一还一报）

不知无所，知无所不（无所不知）

来无踪去（来去无踪）

计走为上（走为上计）

所不能无（无所不能）

在无所不（无所不在）

得一无所（一无所得）

能一无所（一无所能）

**微博语料（所有误例）：**

擒七纵七（七擒七纵，七纵七擒）

人做事做（做人做事）

不当正不（不当不正）

在其位不（不在其位）

一说是一（说一是一）

高一尺道（道高一尺）

入趁虚而（趁虚而入）

慈大悲大（大慈大悲）

杯中物之（杯中之物）

花遍地开（遍地开花）

远万里不（不远万里）

屋及乌爱（爱屋及乌）

言不知所（不知所言）

有一失必，失必有一（必有一失）

惧危而不（危而不惧）

是百无一（百无一是）

以上例子中绝大多数都是“复句型对举”[[39]](#footnote-39)例子中的一部分，例如“不知无所”和“知无所不”都是“无所不知无所不晓”中的片段。它们被变序能产方式发现是因为“回文”的特点，即通过把原表达的最前一个字补到最后或把最后一个字挪到最前形成，比较好进行纠错的工作。

此外，也有极少结构变化的例子，如：有大人在（大有人在），或误用的例子，如：不文一名（一文不名，不名一文）。还有些实例虽然是四字表达，但十分不常说，例如“名闻迩遐”[[40]](#footnote-40)，虽然“遐”“迩”是对举，但语序上已经趋于凝固，形式上已经成为整体，用作“迩遐”，实在牵强；“之乎也者”也是类似。这二者都出现在CCL语料中，而未在微博语料找到类似结构，反映出部分经典语料“转文”的语言风格。、

变序发现新表达的准确率非常高，且绝大多数变序仍为对举型变序，即一三交换、二四交换和前后交换，它们在CCL语料和微博语料中的type数分别为 [162, 226, 362] 和 [26, 37, 55] ，此外，叠字也分别有149例（CCL语料）和55例（微博语料），这四种变序类型牢牢占据所有变序类型的前四名。

但要指出的是，这是在“先分词、后挑四字表达”的模式中，如果在未分词的原始语料中用变序的方式发现四字表达，虽然表达的总数要多得多，但准确率也明显下降。例如在原始语料中，排名第二和第四的变序类型分别是“YAXB”和“XBYA”[[41]](#footnote-41)，说明“回文”现象明显增多，且还有大量如“以为是自（自以为是）”、“进一步更（更进一步）”这样的明显包含双字词（以为）和三字词（进一步）的例子。这里也体现出了前一种模式对于发现变序更加友好。

## 4.4 利用仿造发现新表达

仿造也以d30175中的所有实例为参照。通过仿造能产方式发现的实例数不及通过格式套用发现得多，但准确率比格式套用高得多。

CCL语料中，形式上类似仿造的实例共有13003个type，61876个token；微博语料中，形式上类似仿造的实例共有3024个type，9756个token。表4.1列举了几个实例数量最多和最少的仿造框架。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **分组** | | **既有表达框架** | **type: token** | **举例（括号内为该例的token数）** |
| **高频** | **C**  **C**  **L** | 不得不X(X=尔) | 285: 1492 | 不得不去(77)，不得不停(57)，不得不说(43) |
| 说X是一(X=一) | 34: 365 | 说这是一(98)，说的是一(58)，说它是一(46) |
| **微博** | X有人在(X=大) | 39: 128 | 总有人在(25)，会有人在(16)，天有人在(7) |
| 一X之力(X=臂) | 7: 96 | 一己之力(66)，一人之力(24)，一已之力(2) |
| **低频** | **C**  **C**  **L** | 正：一表狗才（一表人才）[[42]](#footnote-42)；得钱回朝（得胜回朝）[[43]](#footnote-43) | | |
| 误：古夕之年（古稀之年）；干难万险（千难万险）；心像鹿撞（心头鹿撞） | | |
| **微博** | 正：萌态横生（逸态横生）[[44]](#footnote-44)；辣气回肠（荡气回肠）[[45]](#footnote-45) | | |
| 误：不绝于而（不绝于耳）；肆无忌弹（肆无忌惮）；有种难言（有口难言） | | |

表4.1 利用仿造能产方式发现的实例数量最多和最少的仿造框架（低频部分括号中为既有表达）

可以看到，高频框架中有些已经趋近于格式，如“不得不…”，“一…之力”，而其他结构则多误例，如“说…是一”、“…有人在”。当然，趋近于格式的框架仍然可能产出误例，如“不得不停”，可能在现代汉语中不太能这样说，而是要说成“不得不停下来”。这些结构都可以挑选出来，要么算作格式，要么就放入“黑名单”，不对它们进行识别。

而低频的框架则具有较高的准确率。在我们的低频例子中，正例都给出了原文，可以看到这些仿造都明显地依赖原表达，且十分形象生动。而误例中第一例都是音近误用，第二例都是形近误用，第三例才是真正的错误结构。在真实语料中，我们不应把前两例看作错误，而也应将它们识别乃至还原。

粗略估计下，仿造能产方式发现的低频实例准确率在9成以上。对于用仿造能产方式发现新表达的优化，主要还是工程性地筛除一些误例较多的高频框架，此外准确识别边界仍然能大幅地提升识别性能。

## 4.5 利用对举发现新表达

利用对举能产方式发现新表达并不依赖于知识库，且在某些限制下准确率非常高。“某些限制”指的是对词性的限制，由于许多干扰项中都包含各种虚字，且对举表达中多为实字，故我们只针对对举词性为名词、动词、形容词进行了统计；只要两个词性的第一个字母相等，我们就认为二者为同词性[[46]](#footnote-46)。

我们按间位是否为叠字对统计结果进行进一步划分，其中axax代表一三字、二四字分别相同，axay代表一三字相同，axbx代表二四字相同，每个单元格的第一个数字（较小）代表type数，第二个数字（较大）代表token数。结果如表4.2所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | axax | axay | axbx | axby |
| CCL | 752 / 1445 | 7714 / 15539 | 3878 / 5059 | 74835 / 94443 |
| 微博 | 421 / 901 | 2420 / 4377 | 1679 / 2227 | 22606 / 26885 |

表4.2 利用词性对举能产方式发现的实例的统计情况

可以看到，通过这种方式发现的潜在四字表达数量庞大；另一方面，四字表达的token数和type数比值接近1，这说明对举型四字表达确实很临时且低频[[47]](#footnote-47)。

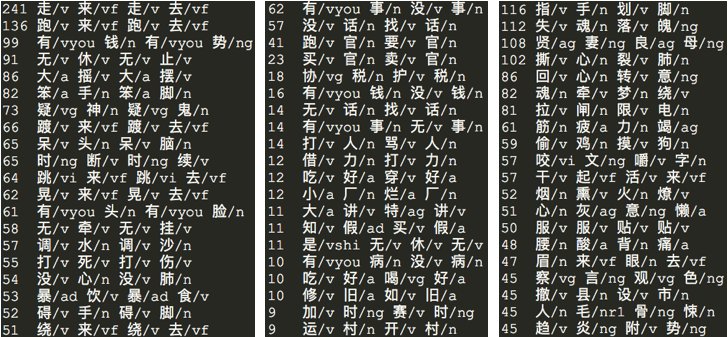
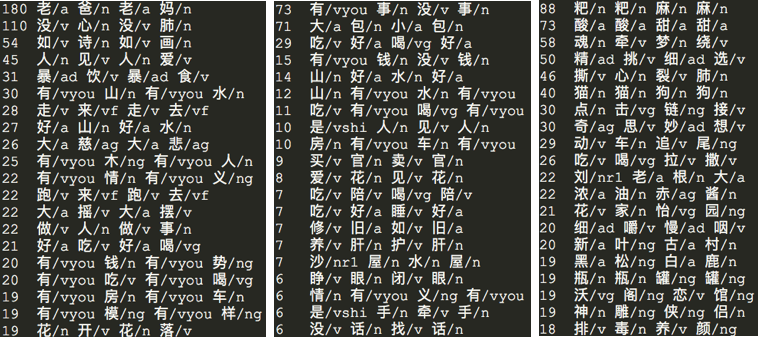
图4.3展示了CCL语料和微博语料中axay, axbx和axby类型的最高频的20个实例，每个实例左边的数字是其在语料中的token数。

图4.3(a) CCL语料中的对举实例

图4.3(b) 微博语料中的对举实例

从结果中可以看出，axay的准确率是最高的，axby次之，其中微博语料由于年代较新，里面有许多词如“动车”、“点击”还没有被词典收录，因此造成axby中有部分误例，而在CCL语料中的准确率是很高的。对于axbx类型，准确率稍低，可能需要限制一下x为名词，这样可以避免类似4.3节中提到的“回文”情况，如“（陪）吃陪喝陪（睡）”、“（有）情有义有（担当）”等。

## 4.6 本章总结

未登录四字表达的发现体现了本文研究的应用价值。在其他分词工具的分词结果上，本文进一步对其中连续的单字序列进行了识别工作，发现了许多新的四字表达。

从单纯的统计结果上看，除了变序的能产性稍低之外，其他三种能产方式都从千兆字节级别的语料中发现了数万个的潜在的四字表达实例。但进一步人工筛查发现，通过格式套用能产方式发现的实例中，误例众多，占到了半数以上，必须通过更多的约束限制才可能投入实际应用之中；而其他三种能产方式的准确率都很高，许多误例也有较为明显的“回文”特征，容易筛除，可以投入应用中。

综合看来，仿造和对举这两种能产方式具有很高的能产性，而变序则不够能产，因为其可变程度是有限的。至于格式套用，虽然有一定能产性，但误例太多，能应用的环境较为严苛。这再一次加强了我们的论点：格式对四字表达的覆盖力是有限的，我们应该通过更多能产方式研究、发现和生产四字表达。

值得一提的是，如果结合多种能产方式进行判定，包括近音或近形两种形式误用能产方式，还能进一步肯定被发现的确实是新的四字表达。

# 第五章　构式语法视角下的范畴重合现象的解释

四字表达在形式语法框架下已经能进行很好的描述，四种能产方式也都有形式化的特点可以利用以发现新的表达，唯一还值得继续探讨的是某些范畴重合现象。例如，“一言不发”是来自“一…不…”还是“一言……”？“天辟地开”是来自变序还是对举？“以逸代劳”是仿造还是误用？

诚然，以上三个范畴重合现象：四字表达的多继承性、能产方式的重叠和误用与否，对于形式学派来说都“无关痛痒”，甚至根本不在形式语法的谈论范围之内。但我们想表达的是，它们确实影响到了人们对一个具体四字表达的认识。形式学派可能不关注一个表达的来历，也不关注一个概念如何在人脑中形成，因为这些都不影响这个表达的形式表征——除非把“一言不发”归入“一…不…”或“一言……”中会产生歧义。但是，如果把对语言现象的描述更深一层，就可以进行更细致、更精密的语言运算，也有助于探索语言生产这块黑域。

构式语法是近年来热门的语法理论，但是在中国，人们对它的认识还过于狭窄。我们认为，要么就在原有的形式理论层面进行讨论，这样反而没有对新术语的认知和记忆负担；如果要使用构式的概念，就应该真正抓住它的认知核心，去回答一些形式学派不怎么关注的问题。这样，才能使形式语法在其应用范围内更加完善、而在更底层的认知现象中找到合适的补充理论。

## 5.1我们对构式语法的理解

5.1.1 构式语法是基于使用的理论[[48]](#footnote-48)

我们先举两个前人论述中没有明确与构式语法挂钩、但比较接近构式语法的观点。文炼（1988）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（1999）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）。

Steels（2011）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）模型，而是一种允准的（licensed）模型[[49]](#footnote-49)：在真实语境中，很少有人会直接指出一个表达是“错误”的，他们并不会向你提供直接反例（direct negative evidence），而是试图理解你的意思，顶多用一些隐晦的手段（indirect negative evidence）来提醒你[[50]](#footnote-50)。更为极端的情况是，人们甚至刻意地利用“错误”，来造成一种焦点，迫使他人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）。

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理论。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。这种由于忽略了在使用环境中才产生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。

5.1.2 构式不脱离于实例单独存在

构式不是一个格式框架，而是一类实例的集合。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…”这样的格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——这是一种所谓的启动效应（prime effect）[[51]](#footnote-51)——而这块区域的中心，我们称之为构式。

构式不仅仅是一类实例的中心。实际上每一个脑海中存储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。这正是Goldberg（2006）描述的构式的不同层次：从一个实体词或一个习语表达这样的实体性构式（substantive construction）到部分填充（partially filled）的半图式化构式（semi-schematic construction），再到被动句、双及物构式这样的图式化构式（schematic construction）。

当然，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。实体构式就是具体短语，图式化构式就是一般的短语规则，而半图式化构式，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。反观短语规则，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。

5.1.3 构式可以没有具体形式

正如Steels（2012b）中打的比方：“形式学派会把锤子描述成一个圆柱体和一个铁块儿以一种特定的方式组合在一起，而功能学派则会描述成一个手握的柄加上一个锤击的头。柄和头既可以是典型的圆柱体和铁块儿，也可以是一只鞋的鞋帮和鞋底——它们具体是什么不重要，重要的是它们提供了什么功能。”基于使用的理论不排斥任何一种形式的语言表达——只要它是真实的。但基于使用的理论不满足于只找到形式上的某种一致性，而是要找到功能上的某种一致性。我们把功能看作一种“形式的形式”，也就是一种抽象的形式。实际上，本文提到的四种能产方式都可以分别看作一个没有具体形式的构式。

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）[[52]](#footnote-52)，把“天开地辟”这一类通过交换汉字顺序而成的形式称为变序构式（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）。这三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”这种抽象的形式来概括它们。它们的修辞功能要大于表意功能，如果要对其释义，应该从修辞角度着手。

最后，我们把诸如“天…地…”的格式统称为半凝固型构式。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 “天…地…”不能进行这种扩展，主要原因就是“四字”的硬性规定。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。

## 5.2 构式视角下的范畴重合现象的解释

5.2.1 构式视角下的多继承性

本文中“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 or 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。而在组合关系中，上层类别比下层类别处于更高的层次，且上层对象对下层对象拥有绝对的管辖权，例如人可以有两条胳膊两条腿，且一个人只能且完全能控制他自己的胳膊和腿。

短语语法中，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。由于上层对象拥有对下层对象的绝对管辖权，低层次的规则或底层的表达不能同时进入多个高级规则中，因而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。靠“是不是”来判断，容易出现不一致性，而兼类的方法，实际上已经打破了体系的严整性，而且在一个具体的语境下，兼类词仍然需要体现出其中一类特征，而隐藏另一类特征。

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），所以不同的构式间是可兼容的。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。半凝固型构式数量庞大，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”半凝固型构式，见下表（基于包含93050个四字表达的大词库统计）：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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5.1 大词库中的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。

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。我们对大词库（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 | | | | | | | | |
| - | 1) 之3 | 2) 不3 | 3) 一1 | 4) 不1 | 5) 无3 | 6) 不2 | 7) 一3 | 8) 心2 | 9) 人2 |
| 5 | 1) 之3 | 2) 不3 | 3) 子4 | 4) 大1 | 5) 人4 | 6) 不2 | 7) 人2 | 8) 无3 | 9) 一1 |
| 2 | 1) 子4 | 2) 大1 | 3) 之3 | 4) 人4 | 5) 不2 | 6) 三1 | 7) 以1 | 8) 金1 | 9) 里2 |
| 1 | 1) 不2 | 2) 大1 | 3) 小1 | 4) 儿4 | 5) 人4 | 6) 子3 | 7) 子4 | 8) 一3 | 9) 三1 |

表3.3 按单个汉字及其出现位置统计排名

其中阈值的意思为不纳入统计的格式[[53]](#footnote-53)所包含的最小实例数，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字[[54]](#footnote-54)格式的实例都被排除之后，剩余实例对单固定字格式的支持。

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字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+Verb+Object： 大吐苦水，大兴土木

一+Subject+Predicate： 一票否决，一醉方休

一+Verb+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。

形式学派其实是排斥多继承这种关系的，因为形式规则是高于语言现象的，如果一个实例能够应用多个规则分析，那就说明产生了歧义。但多继承性却又是没有歧义的。因此形式语法学家可能会不承认这些规则，例如他们会说“一…不…”并没有很明显的意义，不应该搞出这样的规则。

而如果用“多个实例的中心”这种构式观点来看待，我们大可以把这种意义不大的格式也算作构式。一方面构式间是不冲突的，而另一方面，当接触到的“一…不…”的实例多了之后，人们确实会对“一…不…”产生印象。构式在这里起到的就是联系的作用，把一些形式上相似的表达联系在一张网络中。

多继承性体现了语言知识的网状储存特点：在受到一个具体实例的刺激时，所激活的其他实例很可能来自不同的构式区域，因而激活了许多构式；而当该实例扎根于认知后，它又会进一步“点亮”这些构式区域。构式是主观建立或划定的概念，但神经刺激和激活效应确是真实存在的。构式只不过是对这些脑电反应的一种形式化描述而已。

5.2.2 构式视角下的能产方式的重叠

能产方式的重叠其实也是一种多继承性。“天开地辟”不仅符合“天…地…”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序构式；乙可能从“天崩地裂”推及它，则其对于乙属于“天…地…”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，则它加深了丙对变序构式的理解；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，至于理解和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。

当然，我们认为不同的能产方式的使用倾向是不一样的，影响因素我们称之为“省力原则”。比如我们认为使用倾向最高的是误用，因为它是一种被动的、无意识的使用，自然很“省力”，许多人在网络聊天时会直接使用输入法最前面的汉字代替正确的汉字，甚至专门使用同音的数字如“88”或者“666”[[55]](#footnote-55)，这也是一种省力。又比如我们认为仿造的使用倾向比（双固定字）格式套用高，因为仿造只用改动一个字，而格式套用却要改动两个字，显然前者更加省力。这也是我们倾向于把“千叠万壑”看作仿造而不看作由“千…万…”来的原因之一：“叠”和“壑”既不对举，“壑”字又如此不常用，硬生生从“千…万…”造出个“千叠万壑”，多费劲儿啊！

5.2.3 构式视角下的形式误用

形式上的误用是形式学派回避的一个问题，或者说，他们对这个问题持抑制的态度：人们就应该生产出正确的表达；句法分析的时候，也应该按正确的表达来分析。然而事实上，人们出于各种原因会进行误用，例如（Steels，2012a）提到的“外语环境”：

A: Bring me the Salmon.

B: (Foreign student, didn’t know what Salmon meant, so she took the SALT instead.)

A: No, (point to the fish on the dish) I mean SALMON.

B: (Give A the right object.)

计算语言学家会特别注意真实语境下的错误，因为这涉及到消歧工作，而Steels则将之上升为理论层面，他认为语法没有唯一性，有些人可能一辈子都在使用错误的语法，譬如在道尔顿之前，所有色盲患者都没有意识到他们看到的颜色有多么不同。他把形式误用看作是一种正常的现象，并试图让机器将之还原为正确的表达——像人一样——例如通过寻找与该表达形近或音近的相似表达[[56]](#footnote-56)。

形式误用其实就是一种“重新分析”：当一个表达的整体意义被人们正确或错误地习得后，人们通常便不关心其部分的形式甚至根据自己的理解为之安排合适的形式，比如变换一下语序、或忘记一个字的字音或字形。我们应该承认这种误用而来的表达的地位，因为它正这样被许多人自以为正确地使用着。譬如“以逸代劳”，我就是这样理解并使用它多年，而且我相信与我一样的人不在少数。当你遇见如此一个语音形式而没有人告诉你它的正确意义时，你并不是束手无策，而是会根据语境或其字母组合猜出一个意义并为之分配合适的形式。

值得一提的是，在四字表达的分析中我们还遇到一类形式误用，即繁体误用。说其为误用并不是指把简体误用为繁体，而是在多个繁体间误用了异体而非正体字，例如“說”和“説”，前者是《康熙字典》里的正字，也是国家标准（GB）中明确定义的繁体字；而后者在《康熙字典》中是找不到的，严格来说连异体字都算不上，只是一种笔画上的翻转。然而后者（包括前者）却被收入日本大修馆出版的《大汉和辞典》中，并被收进日本工业标准（JIS）中。Unicode为了兼顾与JIS字符集的编码转换，不得不将两个“说”字都收入Unicode编码。

而就是这样一种情况下，无论在词库还是在语料库中，异体“説”却比正体“說”具有高得多的频率。我们猜测原因可能是人们掌握了简繁转换的基本规则：按照偏旁部首把一个简体汉字一部分一部分地转化为繁体，却不晓得古文里“兌”才是正体。这就是我们所谓的“错用一辈子”的例子，而且你甚至无法指责人们这样用是错的，因为有人（日本人）这样用了，而且这样的用法得到了Unicode的“支持”。

# 第六章　结语

四字表达是一个很宽泛的概念，其中既包含了形式紧凑凝固的成语，也包含了形式松散自由的短语。从整体上不好给它一个确定的语法地位。前人的相关研究多从格式出发，对一部分四字表达进行了整理和概括。

我们也对四字表达中较典型的、形式紧凑凝固的、应该看作词的那一部分感兴趣，但我们认为，光用格式不足以覆盖这些四字表达。更重要的是，四字表达还在不断地产生、繁衍，那些修辞意味浓厚、给人耳目一新的表达都是临时创造的，我们想要发现的是这些表达，这就要求我们在共时的、静态的格式描述之外，探索新的动态的能产方式，一方面覆盖更多既有的四字表达，另一方面也发现更多的未登录的四字表达。

我们因此总结了前人有意无意间提到但未给予充分重视的现象，并将之总结为变序、仿造、对举三种能产方式，在词库和语料库两方面对比了其能产性，很好地覆盖了格式顾及不到的实例。其中变序的形式化特征可靠性最高，但能产性也较有限；仿造和对举如果加以限制，例如仿造只限制为改变一个字，对举只限制为实词对举，则其可靠性也足以支持其投入实际应用，且二者能产性非常高，能够发现大量新的四字表达。

而在实际生产过程中，无论是不同的格式间还是不同的能产方式间都存在着重合性，包括一种特殊的能产方式：误用，它有时候也与仿造和变序难以区分。我们认为，从使用的角度来看，多种能产方式的重合是正常的，它反映出语言知识的网状储存特点和认知的主观性，是对经验知识的更底层的描述。基于使用的构式语法理论，则在着眼于高层抽象的形式语言学和着眼于底层事实的认知语言学乃至神经语言学间建立了可用的关联，使我们能从更多维的角度观察和分析语言现象。

本文的研究限于时间，还存在很多待填补的部分，例如文炼和齐沪扬特别提到的单固定字格式，我们就没有对其可靠性进行分析。此外，在指导应用和真正投入应用间还有很多细活儿要做，例如研究不同能产方式（包括误用）间的重合率，从而评价结合多种能产方式对识别新四字表达的准确率的提高程度。但我们认为，我们的重点是提出了三种（加上误用是四种）形式化特征，并证明了它们是有效的新词发现手段。在理论上，则是对格式的很好补充，例如仿造主要补充了三固定字潜在格式，而对举则主要补充了“A…C…”和“…B…D”型格式。

# 参考文献

**中文文献**

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）

顾　森．基于点阵字体的汉字字形相似度比较［Ｄ］．北京：北京大学，2011

顾　森．基于大规模社交网络语料的新词发现和热词挖掘［Ｄ］．北京：北京大学，2013

郭绍虞．汉语语法修辞新探［Ｍ］．北京：商务印书馆，1979

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

侯人渝．现代汉语对举型反复义构式研究［Ｄ］．北京：北京大学，2015

黄燕旋．方位对举四字框式结构研究［Ｄ］．广州：暨南大学，2011

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

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）

马国凡．四字格论［Ｊ］．内蒙古师大学报，1987（3，4）

孟祥英．汉语待嵌格式研究［Ｄ］．山东：山东师范大学，2010

齐沪扬．有关类固定短语的问题［Ｊ］．上海：修辞学习，2001（1）

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

文　炼．固定短语和类固定短语［Ｊ］．北京：世界汉语教学，1988（2）

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）

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）

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）

诸桥辙次．大汉和辞典［Ｍ］．台湾：北一出版社，中华民国76 [1987]

**英文文献**

Allen J. D. *Unicode Standard (Version 8.0)* [M]. California: the Unicode Consortium, 2015

Beuls & Van Trijp. Diagnostics and Repairs [A]. *Language Grounding in Robot*. New York: Springer, 2012: 215-234

Boyd, Goldberg. Learning What Not to Say [J]. *Language*. Washington: Linguistic Society of America, 2011: 55-83

Bybee J. L. Usage-based Theory and Exemplar Representations of Constructions [A]. *the Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 2013, Chapter 4

Michaelis L. A. Making the Case for Construction Grammar [A]. *Sign-based construction grammar*. Stanford: CSLI Publications, 2011: 653-676

Pickering, Branigan. the Representation of Verbs: Evidence from Syntactic Priming in Language Production [J]. *Journal of Memory and Language*. 1998: 633-651

Steels Luc. How to Make Construction Grammars Fluid and Robust [A]. *Design Patterns in Fluid Construction Grammar*. Amsterdam: John Benjamins, 2011:301-330

Steels Luc. Grounding Language through Evolutionary Language Games [A]. *Language Grounding in Robot*. New York: Springer, 2012: 1-22

Steels Luc. Design Methods for Fluid Construction Grammar [A]. *Computational Issues in Fluid Construction Grammar*. New York: Springer, 2012

# 致谢

本文在导师詹卫东先生的悉心指导下完成，本文中对构式语法理论的许多独到见解也得益于詹老师每周一次的讨论班课。三年时间内在他的督促之下完成了对许多西方构式语法理论的阅读，其中对认知学派的主张十分感兴趣，故有了文章的第五部分。而在认知学派与形式学派的站队问题上，也吸取了他的建议，即不要无谓地造成冲突，而应该以解决问题为主。故本文绝大多数篇幅均未采用“构式”的表述，因为我认为四字表达的能产性比较和未登录词的识别这两个话题，在形式语法框架下完全能够解决。“求同”大于“存异”，这样的做法在论文之外的许多地方也都是有益的。在此对詹老师深表感谢。

本文的完成还要感谢我的妻子刘红萍女士。在完成论文最关键的两个月里她给予了我莫大的鼓励，包括许多顿丰美的晚膳。让我在实习之余有充足的精力完成论文的写作。

最后，我所实习也是签约的公司，在完成论文最关键的两个月里并未安排太多的任务，可以说十分清闲，也给了我足够的时间利用公司的网络借助谷歌查阅了许多国外资料。在此也对小组产品经理、我的mentor、小组leader及所有小组成员表示感谢。

# 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. NLPIR平台提供的ICTCLAS的分词结果为“人/n 怂/vg 志/n 短/a”，其中vg代表“动词词素”。 [↑](#footnote-ref-2)
3. 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-3)
4. 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-4)
5. 原文采用的是“语例”的说法。 [↑](#footnote-ref-5)
6. 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”、“以空间换时间”的处理方式也无可非议。 [↑](#footnote-ref-6)
7. 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-7)
8. 例如“六角扳手”、“智能手机”，它们显然明确地指向一类客观实体。我们顶多说其为偏正构词，而不倾向于将其看作自由短语。 [↑](#footnote-ref-8)
9. 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-9)
10. 6708243 / 22169755 ≈ 30.26% [↑](#footnote-ref-10)
11. 这个新造的四字表达经历了两次仿造，第一次是“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-11)
12. 该格式有实例“一己之见”、“一偏之见”、“一孔之见”、“一隅之见”、“一得之见”等。 [↑](#footnote-ref-12)
13. <http://pinyin.sogou.com/help.php?list=6&q=1> [↑](#footnote-ref-13)
14. 下载地址：<http://pinyin.sogou.com/dict/cate/index/54> [↑](#footnote-ref-14)
15. 即对于任意在小词库中出现的条目，它也应该在大词库中出现。 [↑](#footnote-ref-15)
16. 典型性可以理解为该四字表达的实体性或“语言材料”性。四字短语就不是一种典型的四字表达，下面还会提到一类形式误用的表达也不是典型的四字表达。 [↑](#footnote-ref-16)
17. 即大词库比小词库多出来的那部分条目 [↑](#footnote-ref-17)
18. 李传军多收了“长…短…”，“道…不…”（方言），“既…且…”，“…里…气”，“宁…不…”，“人…人…”，“添…加…”，“朝…夕…”，“这…那…”，没有收“…而…之”，重合率达到230 / 240 ≈ 96% 。 [↑](#footnote-ref-18)
19. 即具有固定形式，但需要进一步由实例数量和形式特点来判断其典型性的格式。 [↑](#footnote-ref-19)
20. 实际上，我们只采用了d30175一个词库进行了统计，如果用包含93050个表达的大词库进行统计，格式数量将达到1000多个。 [↑](#footnote-ref-20)
21. 即单音节名词词素 [↑](#footnote-ref-21)
22. 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”来说使用更频繁，故为优势构象。 [↑](#footnote-ref-22)
23. [np [np 一 …] 之 …] v.s. [vp … [np 人 之 …]] [↑](#footnote-ref-23)
24. [vp … 而 [vp 不 …]] [↑](#footnote-ref-24)
25. 这只是一种主观表述，假设“胆…心…”由“心…胆…”而来，但实际上显然并非如此。 [↑](#footnote-ref-25)
26. “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-26)
27. “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”对偶。 [↑](#footnote-ref-27)
28. “每下愈况”出自《庄子·知北游》：“夫子之问也，固不及质，正获之问于监市履稀也，每下愈况。”比喻越深入就越能接近本质；“每况愈下”则出自南宋文学家洪迈的《容斋续笔・蓍龟卜筮》：“人人自以为君平，家家自以为季主，每况愈下。”比喻境况越来越糟糕。 [↑](#footnote-ref-28)
29. 此处“对举”仅针对第一字和第二字而言，而非整个四字表达。三四交换、一四交换与此类同。 [↑](#footnote-ref-29)
30. 括号中为相对应的原表达，下同。 [↑](#footnote-ref-30)
31. 即真正的格式数量很少，大量的潜在格式都只有两三个实例，且大量的双固定字格式其实是三固定字格式。 [↑](#footnote-ref-31)
32. http://ictclas.nlpir.org [↑](#footnote-ref-32)
33. https://taku910.github.io/crfpp [↑](#footnote-ref-33)
34. 420 / 127 ≈ 3.3 > 2 ≈ 1.1 \* 1024 / 566，其中1Gb = 1024Mb [↑](#footnote-ref-34)
35. 例如“这是我的微博”就包含三个潜在表达：“这是我的”、“是我的微”和“我的微博”。 [↑](#footnote-ref-35)
36. 当然具体是不是还要根据具体语境判断，例如“无人可解”可能是“无人可解开”的一部分。 [↑](#footnote-ref-36)
37. “他认为，企业之间如果只是规模上的合并，形成‘连而不锁’的状况，并不能形成强大的进攻平台。”（网络语料） [↑](#footnote-ref-37)
38. “古香古色”这个四字表达本身是一个type，它在语料中的每一次出现则是一个token。 [↑](#footnote-ref-38)
39. 即两个单句间形成的对举。 [↑](#footnote-ref-39)
40. “阳澄湖蟹之名闻迩遐，还与毗邻上海有关。”（《市场报》1994年） [↑](#footnote-ref-40)
41. 我们以“AXBY”代表原表达，“YAXB”相当于把原表达最后一字提到最开头，而“XBYA”则相当于把最开头一字放到最后。 [↑](#footnote-ref-41)
42. 那小黄狗长得小巧玲珑，一表狗才，一身亮油油的黄毛紧裹全身，人见人爱，狗见狗喜。（韩仁均 杂文） [↑](#footnote-ref-42)
43. 夜深人静，食客散去，经营者得钱回朝。（1994年报刊精选） [↑](#footnote-ref-43)
44. （@点名某人）萌态横生连波波老师都忍不住在他脸上（[表情]） [↑](#footnote-ref-44)
45. 带出辣气回肠的好滋味 [↑](#footnote-ref-45)
46. 每一个词类下面还分了一些小词类，例如名词又分为名词素（ng）、人称名词（nr）等等。 [↑](#footnote-ref-46)
47. 许多高频的四字表达已经被分词工具收入其词典，不会分解为四个连续单字词，故不在我们发现的范围之内。 [↑](#footnote-ref-47)
48. 基于使用的理论（Usage-based Theory）主要见于Bybee等人的文章中，如（Bybee，2013）。 [↑](#footnote-ref-48)
49. （Michaelis，2011） [↑](#footnote-ref-49)
50. （Boyd & Goldberg，2011） [↑](#footnote-ref-50)
51. （Pickering & Branigan，1998） [↑](#footnote-ref-51)
52. 之所以用adapted而不用borrowed或imitated，是想强调仿造构式不仅借鉴了原表达，还在其基础上进行了一定的适应，使之满足新的表达需求。 [↑](#footnote-ref-52)
53. 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-53)
54. 也就是具有两个以上字母的型式。 [↑](#footnote-ref-54)
55. “88”的原话是“拜拜”；“666”的原话是“溜溜溜”，意思是“（玩游戏玩得）很溜”。 [↑](#footnote-ref-55)
56. （Beuls & Van Trijp，2012） [↑](#footnote-ref-56)