现代汉语半习语性构式的能产性研究（开题报告）

A Research on the Productivity of Chinese Semi-Idiomatic Construction

作者：田骏

专业：中文信息处理

指导老师：詹卫东

开题日期：2015年11月11日

目　录

一、研究对象

二、研究意义

三、研究概况

四、使用语料

五、研究方法及文章结构

六、小结

附录

参考文献

**一、研究对象**

Goldberg（1995）正式将构式语法（construction grammar）的研究方法应用于论元结构中，该语法将不同尺寸的语言单位统一地看作形式和意义的结合体[[1]](#footnote-1)，也就是构式。本文研究的对象是现代汉语中的某一类四字构式，以《现代汉语词典》中收录的“千…万…”词条[[2]](#footnote-2)为例：它们带有一定数量的“槽”，可供实词填入，且本身具有比较完整的释义模板；另一方面，它们经实例化（instantiation）后的完整形式，还往往给人一种紧凑、凝练的感觉。这类四字构式以其丰富的表现力让人们乐于使用它们，并在使用中进一步凝固、虚化，同时给予使用者更多启发，使他们创造更多实例和更丰富的同类构式。由于该类四字构式形式紧凑，习语感强，且多由成语衍化而来，故我们以半习语性构式（semi-idiomatic construction）称谓它。

**二、研究意义**

前人关于此类四字构式已做了许多研究，意义主要有以下三点：1、研究现代汉语特点；2、建立现代汉语规范；3、指导现代汉语教学，尤其作为外语教学。

然而，前人的研究缺乏一个“完备性”的考察，他们更多的是从理论层面去内省那些典型的构式和典型的实例，他们缺少的是系统地考察这类构式及实例在语料中的分布，尤其是那些“非典型”的构式，如“X百出”，或者典型构式的非典型实例，如“千仰万慕”；又如孟祥英（2010），虽然考察了很多构式和实例，但是并没有把数据整合在一起，形成一个整体的印象和推断。

较之理论意义，我们更加关注其实践意义。构式语法的提出和发展很大程度上是受到一种基于使用（Usage-Based）的研究方法的影响，这种方法更加关注真实用例，而我们希望做的，是借助充分的语料和多元的统计更加客观、系统地描述这一类构式，在前人的基础上回答这么两个问题：

a. （内涵）什么样的结构可以归入我们定义的这一类构式？它的所有实例都可以当作一个词处理吗[[3]](#footnote-3)？

b. （外延）人们对这类构式的使用情况如何？人们为什么倾向于使用这类构式？

如果用“能产性（productivity）”去重新描述这两个问题，则可以转化为：

a. 数量规模；

b. 演变趋势。

答好这两个问题，不仅能对前三点理论意义形成充分、客观的补充，对于未登录词的识别及相关自然语言处理课题也具有一定的启发。

此外，该研究成果对于构式理论应用于现代汉语的研究将具有指导意义，因为半习语性构式是最简单的半图式化构式（semi-schematic construction），且其构式感较强，难以用传统的短语结构去分析。对于它的研究，尤其是实验研究，将有助于将构式理论在现代汉语上的铺展与应用，使汉语研究者多一种手段对汉语进行考察与描述。

**三、研究概况**

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，可惜由于时间较远，我们并未能找到这篇文献。

马国凡（1987）中论述了四字格式的外延与内涵问题。他提出，从修辞学的角度看，可以排除四字格究竟是词还是词组的困扰，而强调它们具有的共同的修辞效果。他给出了几种四字格形成途径，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，也就是下文中所谓的待嵌格式。

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。

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[[4]](#footnote-4)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。

孟祥英（2010）对待嵌格式进行了系统研究，他综述了前人对此类格式的称谓[[5]](#footnote-5)，认为不应将之定义为词或短语，而又要突出固定成分与填充成分的关系，因此沿用周荐的“待嵌格式”称谓。他从原型范畴理论角度出发，将待嵌格式看作特殊的词汇单位，且限定为固定字与填充字交替出现的四字格。孟祥英首次在此类格式的研究中应用了构式语法理论，且结合语料数据对待嵌格式的实例进行了统计，虽然我们对其数据感到疑惑，但定量方法的引入仍值得肯定。

我们认为，格式概念对于四字格式的系统化研究十分重要，尤其是语汇格式的提出；马国凡避免讨论词和词组的定界，使四字格式的研究摆脱了早期西方语言学研究理论框架带来的困惑；冯胜利的研究可谓巧妙地结合了先进理论与汉语实际，使“偶语易安”的讨论可以延续到现代；周荐的待嵌格式的提出则规范化了这一格式的称谓，其后多位研究者均采用这一术语；孟祥英的博士论文则博采众长，且引入了定量方法和构式理论，无疑是一部承上启下的力作。

从前人研究中我们可以概括出以下几点四字格式的特点：

1、实例众多；

2、韵律整齐，定字以交替显现居多；

3、具有明显的习语性。

但是，正如前文所述，他们的研究中缺乏如下考察：

1、四字格式中究竟包含哪些具体的格式？

2、四字格式究竟属于哪种语法单位？哪些实例应该被纳入习语（即词）的范畴，哪些只能算作短语？[[6]](#footnote-6)

3、所有的四字格式都具有相同的能产性吗？

最后，能产性研究应该问的一个终极问题是：

4、四字格式是如何创造、又如何被接纳的？

可以这样总结，前人的研究多从理论探索的角度出发，语料多用来帮助他们阐述理论；而我们的研究恰恰从语料出发，引入的构式语法理论是用于帮助解释语言现象的。

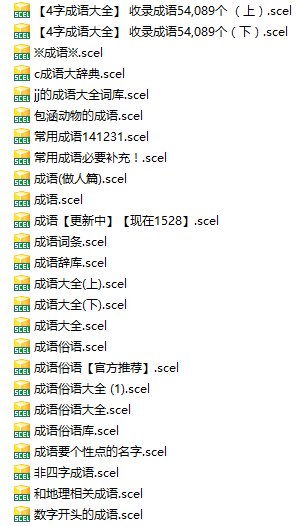
**四、使用语料**

与前人研究不同，比起规范语料，我们更加关心口语中人们都在如何使用和创造这些构式的。对于前者，一来前人已有较充分的研究，二来，从原型范畴理论[[7]](#footnote-7)角度出发，这些习语感较高的构式和实例属于该类中的典型——我们更加关注非典型，因为它们刻画了半习语性构式的边界。对于预测一类构式的能产性，或曰生命力，网络语料显然比经典语料更具有说服力。

我们将借助搜狗细胞词库里的成语词库来进行词库的比较及用于自动发现的模型的训练。搜狗细胞词库是由网民提供的词库，经搜狗编码后形成了搜狗细胞词库格式（.scel文件），可手动导入搜狗输入法，获取地址为：<http://pinyin.sogou.com/dict>；此外，还有一份来自搜狗官方的词库，获取地址为：<http://www.sogou.com/labs/dl/w.html>，我们也将其中的四字词语提取出来，作为一个词库放入词库集中。

我们将使用CCL现代汉语语料库作为经典语料来源，同时以微博语料作为网络语料来源，比较两份语料的差异，并从中寻找有价值的半习语性构式。其中，CCL现代汉语语料库由北京大学中国语言学研究中心提供，网址为：<http://ccl.pku.edu.cn:8080/ccl_corpus>，其收录了近现代包括小说、散文、新闻等不同体裁风格的语料，语料规模1G左右；微博语料由自然语言处理与信息检索共享平台（网址：<http://www.nlpir.org/?action-category-catid-28>）提供，共两份微博内容语料计523万条，语料规模1G左右。

以下是搜狗细胞词库的截图：



**五、研究方法（文章结构）**

我们首先在第二章陈述我们选择构式语法作为研究手段的原因。由于本章已初步成型，故摘取供老师们评判。

1. 为什么用构式方法

Michaelis（2012）中比较了基于允准的语言模型（licensing-based model）和基于抑制的语言模型（suppresion-based model），前者接纳正确的结构，如构式语法；而后者则将不正确的结构排除在外[[8]](#footnote-8)，如生成语法。作者认为，虽然后者更加的清晰、直观，但却更加的依赖动态的操作，且由于追求描述的完备性而牺牲了精确性，无法解释诸如“基于同一底层的两个不同的表层结构具有统计频率上的差异”这样的实际语料中发掘的问题；她认为，构式语法采用的是基于允准的模型，这是一种表层的、静态的模型，具有更大的心理现实性，且更加关心实际的语言现象而非语法现象。

Wulff（2013）则专门从习语的角度探讨了构式语法较之生成语法的长项，她开篇即举Croft and Cruse（2004）及Chafe（1968）[[9]](#footnote-9)说明“习语的存在和假设语法与词汇间有着严格的区分界限的语言理论（比如早期的转换生成语法）背道而驰”。之后，她将分析重心放在了习语性的定量界定上，而她的理论出发点是：习语不是简单的、固定的词语搭配，它们较凝固的词语而言可以发生形式上的变化，而相对于自由短语而言变化又往往是极其有限的。

我们认同以上两位学者的观点。如2.1节总结，前人对于四字格式的研究多少受制于语法框架，他们对于该类格式究竟是归于词还是词组一直含糊不清。传统语法的研究模式使前人仅关注（甚至过于关注）“什么格式是对的”、“哪些实例是习语、哪些只是普通短语”，而不关注“这些格式有多少实例”、“这些实例分布如何”[[10]](#footnote-10)。构式语法忽略了不同层次的语法单位间的差异，而更加关注语言现象本身，这对于四字格式的界定是十分有帮助的。

另一方面，正如Michaelis所说，传统语法在完备性和准确性的选择上更倾向前者。我们不否认前者是一套完善的语法理论的必要性质，但在理论尚未完备之时，迫切地追求完备性可能适得其反。例如语法树的分析方法，似乎暗示了句子中的词语只能相邻地产生联系，且对产生联系的顺序要求严格。但事实上，汉语使用者无论对于“在……之上”这一类的框式结构，抑或“…了个…”这一类的离合词嵌入成分，都表现出十分的认可和接受，对于后者，除了经典的“洗了个澡”这种嵌入动宾式离合词用法之外，甚至还出现了“高了个兴”这类嵌入单纯词的用法，足以证明该成分的凝固程度及人们对其的接受程度。此外，“今天我吃了个苹果”和“我今天吃了个苹果”这类形式可变换（alternative）的句子用相邻且有序的联系关系并不好分析：要么承认有两种不同的语法树（两种不同的底层），要么承认一种结构由另一种结构转换而来，两种解释都不能令人满意，前者的问题在于，没有证据表明这两句话存在语义上的差异；而后者的问题在于无法断言其中一个表达是另一个的原型[[11]](#footnote-11)。

Michaelis提到了构式在允准语言对象的过程中相互作用，换言之，一个构式实例可以同时实现多个类型。对于“我今天吃苹果”的例子，我们可以把“今天”看作一个副词性时间构式，把“我吃苹果”看作主谓构式，后者允准前者填充在主语前后[[12]](#footnote-12)。这样的分析方法避免了将“今天”与具体的片段“吃苹果”绑定，而是与整个表达“我吃苹果”绑定。类似地，我们还可以往句子中加入副词性地点构式，例如“我今天在教室吃了个苹果”或“今天我在教室吃了个苹果”。值得注意的是，与语法树不同，构式并不存在“嵌套”的说法，一个实例能够实现出两个以上的构式纯粹是因为这些构式间彼此允准了对方，它们间的允准并不存在一个先后关系。正如Steels等人构建的FCG系统实现的那样，将构式方法应用于自然语言处理时，对具体构式的应用（apply）完全是并行的。至于各个构式对应的形式（form）在语句中有序（嵌套亦是一种有序）出现，那只是一条允准规则——当我们读到“我”字时，我们已经对后文作出了期待，期待后面是一个行为动词（主谓构式）或状态形容词（主表构式），或者一个副词性的槽。

总结起来，构式方法的优势在于，它一定程度摆脱了传统语法研究中重内省轻实际的方法，避免了在形式规则间做复杂的运算，更注重研究并适应客观事实；同时，它允许构式形式带有可填充的构式槽，并行地考察各构式间的允准关系，避免了迭代嵌套带来的困惑。

2. 相关术语

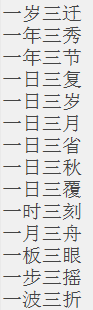
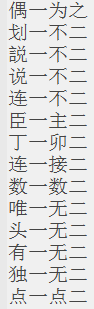
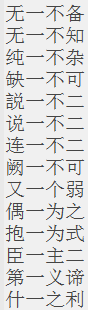
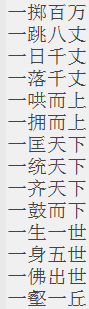
我们把所有未经填充的构式称为构式（construction），而把这些构式经过填充，也就是实例化（instantiation）后所形成的完整的表达称为构式实例或简称为实例（token）。这意味着：一个实体构式（substantive construction）从形式上看既是构式也是实例，因为该类构式的实例即为其本身；而如“千…万…”这样的图式化构式（schematic construction）只能是构式，其对应的实例为“千言万语”、“千差万别”等。

我们把构式中的成分称为常项（constant），而把需要实例化的部分称为变项（variable）。我们将变项用英文字母代替，例如“千X万Y”。一般地，两个变项如果用相同字母代替意味着它们必须实例化为相同的成分，如“一X再X”；而若用不同字母代替，则不一定非要实现为不同的成分，例如“千X万Y”既可以实例化为“千恩万谢”，也可以实例化为“千谢万谢”。

此后，我们的研究主要分为两个部分，分别对应前面提到的能产性的两个方面：

1. 构式发现

构式发现包括从词表中发现和从语料中发现两个子任务。从词表中发现构式任务较简单，主要通过比较词项的相同成分发现构式，例如对【四字成语大全】，分别按第一字、第三字、第二字、第四字顺序（以下简称1324序）、2431序、2314序、1432序排序，得到如下序列：

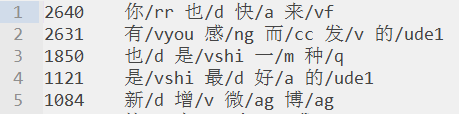
这里可以看出，“一X三Y”、“X一Y二”都像是构式，也被前人收录了。但“X一不Y”是不是构式？“一日三X”是不是构式？这些就有待于设计方法进行判定了。我们希望的结果是，可以纯粹地通过形式进行判定，而尽量少地引入意义标准。这受到一种基于使用（Usage-Based）的观点的影响，代表人物有Joan Bybee、Stefan Th. Gries[[13]](#footnote-13)[[14]](#footnote-14)等人。例如Gries（2014）中所说：“语料库语言学的工作必须基于以下基本假设：一个元素的分布特征（distributional characteristics of an element），反映了许多乃至绝大部分（many if not most）有关它的句法结构、语义和语用特征。”具体实现上，我们可以参考自然语言处理中的一些统计分词方法，例如最大熵、条件随机场等等。例如顾森（2013），用的是无指导的方法去发现人人网语料中的热词，其中用到的就是一种信息熵的方法，它对一个中文词做了如下假设：内部凝固、外部自由、使用频繁——这是可以量化的特征。我们也可以考虑以下还有哪些可以量化的方法和特征。

从语料中发现构式要困难一些。现有的分词模型的基本假设是“词是凝固的（substantive）”，不会考虑说还有一类这样的半开放式的词语。我们可以验证这一点。我们目前打算用的方法基于的假设是，未登录的半习语性构式实例用分词软件分词的结果应该是四个单字词，例如：在500万条微博语料中，“千车万人”、“百转千回”分别出现了39次和22次，它们都没有被张华平博士的ICTCLAS分词软件发现：





我们可以考虑一下能否从这些单字词序列中找出我们需要的构式。比如：对500万条微博语料分词、并提取最长单字词串的结果排名前五的如下：



其中“有感而发”非常像我们要找的实例，它可能实例化于“有X而Y”构式。我们可以研究一些策略发现这些实例，例如借助冯胜利（1997）的假设，我们判断间隔的两个单字词是否能成词；也可以加入一些参数量化，例如“假设词串中出现了非成词语素词（以g）结尾的词性标注，则该词串可能包含实例的几率增加”；再比如，能否排除一些肯定不是实例的单字词串组合？

2. 能产性分析

能产性分析中我们主要考虑以下几点：

a. 频率分析；

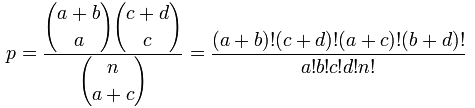
b. 词库扩展分析；

c. 语料差异分析。

频率分析包括简单频率分析和相关性分析。简单频率的意义是很直观的，也是Bybee（2010）一书中致力于阐述的理念。假如语料是充分的情况下，一个构式实例的频率越高，证明它越凝固，习语感越强；一个变项在一个构式槽中出现的越频繁，证明该构式与该变项的联系越紧密。此外，Casenhiser and Goldberg（2005）做过一个实验，目的是检验高频词是否有助于习得构式，实验对象是小孩，实验方法是构造一些不存在的词和一个不存在的构式（实验里是SOV语序），检验小孩能否避免新词的干扰快速习得构式义。实验结果是，假如某一个新词在某一个构式中出现的频率特别高，对于该构式的习得是有帮助的。实验受到另一个实验的启发，Bates *et al.* (1988)中分析了一个母子对话的自然语料，发现对于三类构式：S V Loc，S V O1 Oblique[[15]](#footnote-15)和S V O1 O2，动词是go、put和give的概率都很高，因此认为高频动词的意义会独立地与构式义联系。

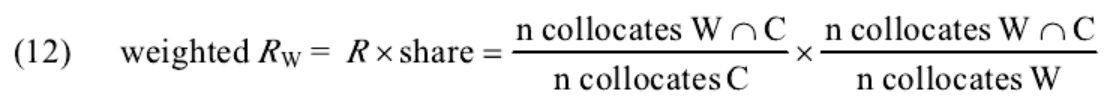
相关性分析主要基于如两篇文献：

1. Gries（2005）利用一个被为“费舍尔精确性检验”的方法对构式（他称为collostruct，下简记为C）与变项（他称为collexeme，下简记为L）间的联系强度作了衡量，对于每一对C与L，需要记录4个参数，分别是L在C中出现的次数（记为a）、L在C外出现的次数（记为b）、其它词在C中出现的次数（记为c）及其他词在C外出现的次数（记为d），然后套用如下公式：



这种计算方法较简单频率的方法而言还考虑了变项对构式的选择性，它主要用于描述变项义与构式义间的相关程度（作者称变项相对于构式的典型程度）。

2. Wulff（2013）利用了一个相似的相关性检验方法（她声明该方法推广自Berry-Rogghe(1974)的weighted R方法）。她所利用的公式为：



实际上我们可以发现，该公式中的参数与Fisher Exact Test中的参数是部分等价的，该公式可以看作上个公式的弱化版本。但作者用了另一种解释方法：她认为，那些不典型的词填入到构式槽中所形成的实例恰恰是构式感比较高的实例。（试想一下，give me a ball和kick me a ball哪个构式感更强？）

我们会对这两个公式再考察，并给出一个比较满意的统计学解释。

词库扩展分析是基于如下事实：收集到的十余个不同的词库中，其四字词的规模主要有以下几个量级：五千，一万三，一万七，三万，四万二，五万。一般的成语词典中收录的成语大概在五千条左右。我们希望考察一下这些不同量级的词库究竟多出来哪些词，是否随着量级的提升新增的词语也从典型实例向非典型实例演变？

语料差异分析主要考虑规范语料库（CCL）和网络语料库（微博）间的构式实例差别，我们将根据构式发现的结果分别在两个语料库中进行实例的统计，以观察“半习语性构式在发展的过程中有哪些变化”。

此外，我们还考虑引入一些实验分析来辅助检查我们得出的结论，包括调查问卷方法和国外已有的一些构式实验方法，看看我们分析出的能产性高低符不符合人们的认知，以及看看人们意识中是否存在半习语性构式的概念。主要参考文献是Gries, Wulff（2005），其主旨是通过实验验证*Do foreign language learners also have constructions?* 该文献又主要引用两篇文献中设计的实验进行重现，包括Pickering, Branigan（1998）[[16]](#footnote-16)对于句法启动效应的研究和Bencini, Goldberg（2000）[[17]](#footnote-17)中设计的排序实验。

除了构式发现和能产性研究外，我们还打算进行一些深化的讨论，包括类似构式，如李宇明提出的“词语模”[[18]](#footnote-18)概念，或前人研究的离合词现象。我们希望将对半习语性构式的研究推向更多的汉语构式研究，因为半习语性构式是最简单的带变项构式，是其他构式的“形式上的浓缩和简化版本”。

六、小结

本文最终的目的是去描述半习语性构式的现状，为此，我们得首先知道这类构式“长什么样”。因此，我们首先通过一些统计指标去发现构式，然后再通过另一些统计指标和实验去评判这些构式，评判的结果又可以反过来修正我们发现构式的方法。

半习语性构式的边界是模糊的，因此，我们一开始只能从一些形式标准，例如“四音节特定格式字符串”出发，因为我们还不知道哪些是“半习语性构式”。但当我们有了足够多的数据后，我们就可以为他下一些断言，例如对于“X可餐”、“X巴巴”这类疑似构式，如果他们的实例分布非常的不均匀，“秀色可餐”和“可怜巴巴”分别占到了90%以上的比例，那我们就比较倾向于将“秀色可餐”和“可怜巴巴”当作两个词收入词典，而不把“X可餐”、“可怜巴巴”当作两个构式收入词典了；又比如，我们发现“X可餐”在CCL中几乎没有实例，而在微博中有了实例，并且有不同的实例，那我们又可以考虑把它收入词典了，因为我们可以判断它是一个新产生的构式，正在逐渐发展。

这些工作不仅可以评价半习语性构式的能产性现状，还可以推演到其他的图式化构式的分析上，从而将计量语言学方法和构式方法更好地介绍给汉语学界，使其理论在应用中得到完善。

**附录**

孟祥英（2010）收录格式：

挨…受… 爱…不… 百…百… 百…不… 百…千… 半…半… 半…不… …背…腰 本…本… 避…就… 边…边… 兵…将… …兵…将 冰…雪… 不…不… 不…而… 不…其… 不…一… 不…之… 不…自… 残…剩… …长…短 …肠…肚 车…马… 称…道… 成…成… …仇…恨 出…入… 春…秋… 此…彼… 粗…大… 大…不… 大…大… 大…其… 大…特… 大…小… 大…一… 待…不… 得…且… 东…西… …东…西 独…独… 断…残… 多…多… 多…善… 多…少… …而…之 …儿…女 耳…目… 翻…倒… 反…为… 飞…走… 非…非… 非…即… 风…浪… …风…浪 风…日… 风…雨… …风…雨 改…换… 敢…敢… 高…大… …歌…舞 各…各… 各…其… …根…底 …公…私 孤…寡… …古…今 鬼…神… 国…民… …国…民 含…带… 好…好… 红…绿… 忽…忽… 胡…乱… 互…互… 化…为… …魂…魄 或…或… 鸡…狗… 既…又… …家…户 街…巷… …街…巷 …金…银 金…玉… 紧…慢… 精…细… 旧…重… …旧…新 可…可… …口…舌 …来…去 …来…往 连…带… 龙…凤… …龙…凤 龙…虎… …龙…虎 乱…乱… 屡…不… 屡…屡… …眉…目 眉…眼… …眉…眼 没…没… …门…户 明…暗… …模…样 南…北… …南…北 男…女… …男…女 内…外… 能…能… 能…善… …年…月 …朋…友 披…戴… …皮…脸 七…八… …七…八 奇…异… 千…百… 千…万… 前…后… …前…后 …腔…调 且…且… …情…理 …情…意 …情…义 …缺…少 人…马… 日…月… 如…如… 如…似… …入…出 若…若… 三…两… 三…六… 三…四… …三…四 三…五… 山…海… …山…海 山…水… …山…水 上…下… …上…下 神…鬼… …生…死 …声…气 时…时… …始…终 …是…非 是…是… …手…脚 说…道… 说…就… …思…想 死…活… …死…活 四…八… 似…不… 似…非… 似…似… 随…随… 天…地… …天…地 …天…日 …头…脸 …头…面 头…脑… …头…脑 …头…尾 万…千… 惟…是… 文…武… 无…不… 无…可… 无…无… 无…之… 五…六… 先…后… 现…现… 相…相… 小…小… …心…胆 …心…肺 心…气… 心…神… 心…手… 心…意… …心…意 …牙…齿 …言…语 …衣…食 一…百… 一…半… 一…不… 一…而… 一…二… 一…就… 一…千… 一…三… 一…无… 一…一… 一…再… 一…之… 以…代… 以…为… 以…相… 由…而… 有…必… 有…可… 有…无… 有…有… 又…又… 愈…愈… 越…越… 云…雾… …云…雾 …灾…祸 招…惹… 朝…暮… 真…实… …针…线 争…夺… …枝…叶 转…为… 装…作… …子…孙 自…自… …嘴…舌 左…右…

**参考文献（部分）**

周　荐．　双字组合与词典收条［J］．　中国语文，1999，（4）.

周　荐．　《现代汉语词典》中的待嵌格式［J］．　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］．　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

张卫国．　现代汉语实用语型［M］．　北京：中国人民大学出版社，1992

袁　宾．　近代汉语概论［M］．　上海：上海教育出版社，1993

齐沪扬．　有关类固定短语的问题［J］．　修辞学习，2001，（1）

曹　炜．　现代汉语词汇研究［M］．　北京：北京大学出版社，2003

王　寅．　构式语法研究（上卷）：理论思索［M］．　上海：上海外语教育出版社，2011

Stefan Th. Gries．　Data in Construction Grammar［C］．　The Oxford Handbook of Construction Grammar, 2013: 93 - 108．Oxford: Oxford University Press

Adele E. Goldberg．　Constructions: A Construction Grammar Approach to Argument Structure［M］．　Chicago and London: The University of Chicago Press

[Laura A. Michaelis](http://www.researchgate.net/profile/Laura_Michaelis)．　Making the Case for Construction Grammar［J］．　Sign-Based Construction Grammar, 2012: 29 – 60．　CSLI Publications

Stefanie Wulff．　Words and Idioms［C］．　The Oxford Handbook of Construction Grammar, 2013．Oxford: Oxford University Press

1. cf. Gries（2013），见2.1节。 [↑](#footnote-ref-1)
2. 《现代汉语词典（第六版）》P1031 [↑](#footnote-ref-2)
3. 如孟祥英（2010）言：“‘不卑不亢’‘不伦不类’是成语，‘不大不小’‘不高不低’则无法被看作成语，可以把它们归作类固定短语……历史角度看，成语最初都是类固定短语，其中有些可能会随着使用频率提高而词汇化为成语。这与汉语词汇的模糊性特点有关。” [↑](#footnote-ref-3)
4. 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-4)
5. 包括张卫国（1989）（后并入张卫国（1992）），袁宾（1993），齐沪扬（2001），曹炜（2003）等。 [↑](#footnote-ref-5)
6. 比如“帮云亦云”，如果它还能扩展出“师云亦云”、“腕儿云亦云”，那么应该将“X云亦云”收入半习语性构式，而把这三个表述都看作该构式的实例。 [↑](#footnote-ref-6)
7. 见王寅（2011）：463 [↑](#footnote-ref-7)
8. 作者的意思大概是，允准模型对于什么是“错误的结构”非常的谨慎，它承认所有存在的表述都是合法的，而不像抑制模型那样将合法与不合法严格区分开来。因此，作者认为前者是积极的，而后者是消极的。 [↑](#footnote-ref-8)
9. Croft等人认为构式语法从人们的语言知识中为习语找到了存在的空间；而Chafe更直言习语是“乔姆斯基范式中的异常”。 [↑](#footnote-ref-9)
10. 举例说明：如果你在北京遇到一个黑人，他更可能来自美洲还是非洲？如果仅关注“什么是对的”，可能会根据“非洲人口中黑人比例高”这一事实得出该黑人来自非洲的结论，但事实上，由于非洲人口较为贫困，不太可能出现在北京，因此北京出现的黑人更可能来自美洲。统计学中，前者被称为极大似然估计，而后者称为贝叶斯估计。后者较前者的优势在于，它没有忽略来自参数本身（在这里是该人来自美洲和非洲的可能性）的影响。因此，对一个尤其是已经形成固定认识的概念佐以数据分析是必要的。 [↑](#footnote-ref-10)
11. 可能人们存在这样一种认识，即时间词前置的表达比时间词插入主语之后的表达更规范，但在口语中，至少笔者认为，后者出现的频率要高于前者。 [↑](#footnote-ref-11)
12. 对于英语而言，则是主谓构式的前后。 [↑](#footnote-ref-12)
13. Stefan Th. Gries，语料库语言学家。根据最新的wikipedia资料显示他 born in Hamburg, Germany。不过，除德语、英语之外，他应该还精通丹麦语，并学习过4年俄语。现主要从事corpus linguistics研究工作，力图将统计学方法引入语言学研究中。 [↑](#footnote-ref-13)
14. Gries 从2003年始提出Collostruction的计算方法，并广泛应用。其2014年再发一篇文章***Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics***重申其意义与宗旨，旨在回应Bybee在她2010年出版的*Language, Usage, and Cognition*书中对Collostruction的质疑。 [↑](#footnote-ref-14)
15. “即使现在有一个偶造的动词mooped，在句子*Pat mooped the feather onto the table.*中，我们也能感受到句子的意思是feather in question has somehow found its way to the table with the help of Pat.” [↑](#footnote-ref-15)
16. The notion of syntactic priming refers to the fact that speakers tend to repeat syntactic structures they have just encountered (**produced** or **comprehended**) before. [↑](#footnote-ref-16)
17. Bencini & Goldberg (2000) 通过句子排序证明了约有半数的被试者在面对语义上相似的句子时更倾向于基于构式的排序而不是基于动词的排序。 [↑](#footnote-ref-17)
18. 主要是一些词缀性质的词，如“X党”“X帝”。 [↑](#footnote-ref-18)