基于语料库的汉语语块分类研究

王凤兰[1]，于屏方[1]，许琨[2]
（1.广东外语外贸大学 留学生教育学院，广州 510420；2.上海市 长征中学，上海 200231）

摘要： 文章采取基于语料库的研究方法，从语言使用的概率性特点入手，考察现代汉语系统中相关词语序列的使用频率以及相互搭配强度，对语块进行识别和判定。在对外汉语教学视角下，从语义、语法、语用等角度对语块进行多维度的分类，并从语言学视角对汉语语块的特点进行分析。文章认为：高频率、高互信息值的搭配类语块在语言使用中的整体提取性应该得到重视，它们是语块的重要组成部分。特别是跨层类组构语块，更是语料库语言学视角下语言系统中的一个重要聚类。

关键词：汉语；语块；分类；语料库

中图分类号：H109.4　　**文献标志码：**A　　**文章编号：**1001-0823(2017)03-0016-06

1.引言

语块是在自然话语中普遍存在的一种多词汇单元结构(Altenberg 1998)。它数量众多，形式多样，兼具词汇与语法特征，常被认为以整体方式储存在心理词典中，使用时作为一个独立词条整体提取(Wray 2002)。基于使用的语言观认为：母语使用者广泛依赖语块进行交际(Langacker 2008)。二语习得研究业已证实，语块能力是二语综合能力的一个重要指标，二语学习者能否具有母语者似的选择能力和流利能力，能否如母语者似的进行流利正确和地道的表达，取决于他们心理词典中储存了多少语块以及在运用时这些语块能否整体快速提取(Cortes 2004)。近年来语块习得已然成为二语习得研究领域关注的重点问题。

语块的内部成员众多，异质性程度较高，在长度、意义的固定性、结构的凝固性、使用功能、连续性等方面都有不同。因此，学界对语块的分类标准以及结果也存在着不同程度的差异。一些学者对汉语语块分类进行了研究，取得了比较大的进展。如周健(2007)、钱旭菁(2008)、亓文香(2008)、吴勇毅等(2009)、薛小芳、施春宏(2013)、王文龙(2013)等。不过总体而言，这些研究基本上采取的是简单枚举或分类列举的方法。其不足之处在于：分类标准不一致，有时甚至相互冲突；分类结果不具有普通性、系统性，一些常见的类型无法在其分类体系中找到相应的位置；在分类中主要关注的是语义标准，对习语类语块较为关注，语法标准和语用标准较少使用。搭配类语块是汉语语块中非常重要的组成部分，在对外汉语教学中也起着非常关键的作用，以往的研究没有建立在语料库的基础上，没有明确的界定和判断标准，对搭配类语块重视不够。

本文采取基于语料库的研究方法，考察现代汉语系统中相关词语序列的使用频率以及相互搭配强度，对语块进行识别和判定。在对外汉语教学视角下，对语块进行多维度的分类。在此基础上，对汉语语块的特点进行分析。

2.语块的识别与判定

目前，学界对语块的识别主要采用三种方法：基于心理语言学方法的识别、基于语言学特征的识别以及基于语料库语言学方法(王立非、张大凤2006；段士平 2008)的识别。

基于心理语言学方法的识别是采用自定步速阅读、反应时间研究、语义启动研究、眼动仪研究以及ERP等方法，分析语块在心理词库中的表征形式、

基金项目：国家社科基金项目“外国学生汉语语块习得及教学模式研究”(14BYY092)

作者简介：王凤兰，广东外语外贸大学教授，博士，研究方向为对外汉语教学及现代汉语语法；
于屏方，广东外语外贸大学教授，博士，研究方向为词典学及计算语言学；
许琨，上海市长征中学教师，博士，研究方向为对外汉语教学及计算语言学。

提取方式、心理现实性以及加工过程等。基于语言学特征的识别在很大程度上是一种基于定性研究的识别。这种语块鉴别方法，基本上植根于语言学经验或属性，对典型语块的分析较为适用。

随着语料库语言学的发展，内省式的语言研究方法被认为无法全面、真实地反映语言使用的真实情况，基于语料库的研究方法在语言研究中被广泛使用。语块研究应建立在基于使用的语言观基础之上。而在当代语言学研究中，基于语言使用的研究必须借助于语料库及其相应的计量研究。“语料库分析法已经成为语言学研究中的一个重要的实证范式，推动了词汇和语法的深入研究”（Stenfanowitsch,Gries 2006）。

本文采取基于语料库的研究方法，从语言使用的概率性特点入手，凭借统计测量手段，考察现代汉语系统中相关词语序列的使用频率以及相互搭配强度，通过一系列提取步骤对语块进行提取。在此基础上，进行语块的分类。对语块搭配强度的计量以互信息值法为主要判断标准。学界惯用的做法是：语块的互信息值如果达到3，则被认为具有统计学上的显著性（参见Hunston 2002）。

本文首先运用Antconc软件计算多词单位的共现频数，然后参照互信息值的计算公式自编程序计算互信息值。首先从50万字（包含25万字现代汉语书面语语料及25万字现代汉语口语语料）自建语料库中提取出出现频数在3次以上、互信息值达到3的2-6词的多词单位，然后进行人工干预。从理论上说，两词或以上的多词单位都可能成为语块，但如果超过一定长度，就很难被整体记忆。英语界研究语块时多把2-6词语块作为考察范围，因此，本文也把语块的长度定为2-6词。依照一定的标准进行人工干预，如剔除无意义的言语碎片、带词缀的复合词等，另外《现代汉语词典》中标注词性的条目不纳入语块范围。

需要说明的是，我们使用语料库驱动的方法对词语序列进行语块识别，主要针对的是搭配类语块。习语类语块作为多词单位可以被提取出来，我们根据其语言特征判定为语块，不论其出现频次是否在3次以上。

3.对外汉语教学视角下汉语语块的分类

语块作为一个范畴，其内部成员性质不同，类型多样，具有非常明显的异质性特征。因此，对语块的分类，需要从不同的维度进行，以实现分类的系统性。

语料库语言学家Sinclair（1991）认为人类语言系统的运作遵循着两个原则：自由选择原则和习语原则。前者指语言使用者在语言使用中根据相关语言系统中的规则生成的各种搭配。后者指语言使用者在交际过程中使用大量的半预制短语。

从内部的凝固程度看，一些语块意义的整体性突出，比如“老弱病残”，另一些语块如大量的三字以上的惯用语以及四字以上的成语甚至形成了脱离字面意义之外的修辞义，这种语块因为意义上的规约性以及在本族语使用者心理词库中的整体提取性特点，我们将其归入“习语类”范畴。习语类范畴基本上遵循的是Sinclair提出的“习语原则”。与这一类相对的遵循的是“自由选择原则”——在自由组合的过程中，某些独立成分与其他成分的共现频率达到某一特定数值，形成较强的搭配力，表现为较大的互信息值。在对外汉语教学中，这些高频率、高互信息值的组合单位也应被认为是语块的一个组成部分。

简言之，语块内部的异质性决定了对其进行划分的标准不可能具有唯一性。习语类侧重的是语义标准，搭配类侧重的则是语法标准。我们对汉语语块的分类如下：

表1　汉语语块分类

语义标准				语法标准										语用标准	
				搭配						框式结构					
				同层结构				跨层结构							
惯用语	成语（含固定短语）	俗语	歇后语	名词性结构	动词性结构	主谓结构	其他结构	基本词汇化结构	未词汇化结构	双项双框式	单项双框式	双项单框式	单项单框式	话语标记语	社交客套语

3.1　汉语语块分类的语义标准

在语块的上述分类中，语义标准是对外汉语界长期以来秉承的一个标准，也是学界接受度比较高的一个语块判定标准。这应该与习语类语块在相同频率下比非习语类语块更具有心理突显性有关。语义标准强调的是语块意义的凝固性、规约性和不可类推性。惯用语、成语、俗语以及歇后语等都属于此类。此类语块，在传统语言学中多被归入“语”的范畴，与“词”形成对照。从语义的整体性程度分析，惯

用语、成语、俗语和歇后语都具有意义的完整性和提取的整体性。还有一些四字格固定短语，如“衣食住行”“有问必答”“无人不知”等，虽未收录在成语词典中，但在母语者看来已经固化为一个整体，显然也属于语义型语块的范畴。

3.2 汉语语块分类的语法标准

3.2.1 搭配类语块

语块分类的语法标准，必然与结构密切相关。从这一维度关注语块的分类，相关符号的共现频率以及搭配强度必然成为重点考虑的因素。语言系统中符号组合的线性规则使得语言符号总是一个接一个依次出现。有些语块的内部成分在一个层面，有些则不然，因此会有同层语块与跨层语块的区分。

(1)同层结构语块

同层结构语块指的是语言单位中相邻共现的成分在语法结构上处于同一层次。比如“这个时候”“长时间”为定中关系，而“找对象”则为动宾关系。因为语法关系清楚明晰，语言共同体对同层语块的整体接受度比较高。在我们所提取的语块中，大部分都属于这一类型。

同层结构语块包括名词性结构语块、动词性结构语块、主谓结构语块、其他结构语块等。其中，名词性结构语块和动词性结构语块下边又分成若干类。具体如下表：

表2 同层结构语块分类

大类	小类	举例
名词性结构语块	定中结构语块	经济条件、发展方向、生活方式、自然风光、大嗓门、快节奏
	指量名结构语块	这个时候、这辈子、这个岁数、那种感觉、那段时间、那个过程
	数量名结构语块	一缕阳光、一座山、一顿饭、一种可能、一首歌、一束花、两码事
	量化结构语块	大多数、大多数、若干年、每次、所有人、全世界、有些情况
动词性结构语块	动宾结构语块	办签证、打电话、谈恋爱、做生意、刷卡、找工作、掉眼泪、钓鱼
	状中结构语块	随身携带、高速发展、互相帮助、热烈欢迎、衷心祝愿、深入了解
	连动结构语块	外出就餐、出去玩儿、看图说话、倒杯茶喝、买本书看
	动补结构语块	避开、变成、意识到、摔倒、坐落在、爱上、比不上、露出、受到
主谓结构语块		记忆深刻、交通便利、精力旺盛、阳光明媚、经验丰富、气候宜人、经济发达
其他结构语块	虚义方位结构语块	表面上、基本上、课堂上、社会上、一路上、事实上、某种程度上
	“动词+于”结构语块	归结于、莫过于、相当于、取决于、有利于、有助于、来源于
	“动词+着”结构语块	意味着、洋溢着、伫立着、标志着、充满着、照耀着、怀着、藏着、试着、冒着

定中结构、指量名结构、数量名结构、动宾结构、状中结构、主谓结构类语块属于词语搭配类语块。在外国学生出现的偏误中，词语搭配错误非常多。究其原因主要是学生习惯于孤立地记忆生词，不了解词语的组合。

应该让学生养成以词语搭配语块为单位记忆与运用的习惯。一些动补结构，如“意识到”“摔倒”“受到”等类推性差，几乎总是一起出现，宜作为整体进行教授。其他结构语块类，如虚义方位结构、“动词+于”结构、“动词+着”结构等，其中的“上”“于”“着”依附性非常强，几乎所有场合都与前边的词语一起出现，甚至有凝固成词的趋势。

(2)跨层结构语块

董秀芳(2011)将跨层结构定义为“不在同一句法层次上而只是在表层形式的线性语序上相邻近的两个成分的组合”。跨层结构在任一语言中都可以出现。比如英语中的“It's argued that”“I'm afraid that”等都属于此类。

汉语中的一些跨层结构有时会伴随着词汇化。董秀芳(1997)认为一些跨层结构是由于语音、语法上的原因而形成。汉语的韵律节奏模式最常见的是两个音节为一个音步。这样，原本处在不同的语法层次上的词，如果恰好处在同一音步之内，就有可能由于韵律的需要而连读在一起，久而久之，就会粘合成为一个结构体，从而形成跨层结构。语法上的原因主要是处在不同语法层次上的两个语言单位在句法排列中经常作为相邻成分连用，经过历时的发展，在进行重新分析之后，其相关组成部分进行了融合，语义逐渐凝固，内部语法结构不可分析，从而彻底实现了词汇化。

如“否则”、“既然”、“非常”等。还有一些双音节的跨层结构，共现频率非常高，但还是可以分析其内部结构，如“不太”、“还没”“从未”等，尚未彻底词汇化，但可以说基本实现了词汇化。这类属于语块的研究范围内。

更多的跨层结构并未实现词汇化。它们在线性排列中位置相邻，并且不属于同一语法结构层次，也不会作为一个词使用，比如“真的是”“简直是”“主要是”“最好还是”“是为了”等。从语感上看，“真的是”等很难归入词汇单位的范畴，但是这些语段的互信息值都超过了阈值，应划为语块的范围之内。

表3　跨层结构语块分类

类别	举例	
双音节基本词汇化结构语块	不太、还没、不再、总会、只能、才能、从未、就像、总要	
未词汇化结构语块	两个词语义相近	可能会、一定要、必须得、全部都、然后再、但是却
	两个词语义不同	并没有、从来不、是为了、是因为、决不能、倒不如、最好还是
	“~是”结构	或许是、尤其是、确实是、实在是、就算是、关键是、几乎是、完全是、真的是
	“~就”结构	动不动就、根本就、然后就、从小就、本来就、后来就

跨层结构语块在传统的语言学研究和语言教学里并没有得到相应的重视，主要原因在于：第一、在传统语言学研究模式下，跨层非短语结构难以凸显并呈现出来。语料库技术介入语言学研究后，跨层非短语结构才得以凸显；第二、跨层非短语结构并不具有语义上的心理现实性。在传统的以语义为词汇单位主要判断标准的情况下，跨层非短语结构自然会被排除在外。第三、除了已经基本词汇化的跨层语块之外，其他跨层结构语块在语义上不具有自足性，需要进行进一步的填充，比如“最好还是”，必须在后面出现附加成分。

3.2.2　框式结构语块

邵敬敏(2016)指出：典型的框式结构指前后有两个不连贯的词语相互照应，相互依存，形成一个框架式结构，具有特殊的语法意义和特定的语用功能。本文按照邵敬敏(2016)的分法把框式结构语块分为以下四个类型：

(1)双项双框式。即有两个前后可变项，也有前后两个不变项，这是最典型的框式结构，结构紧凑。如：连A带B、一A不B；与其A不如B；不仅A，而且B等。

(2)单项双框式。即一个由非连续的前项后项构成的框架内只插入一个可变项。如：拿A来说、非A不可、像A似的等。

(3)双项单框式。框架只有一项，而可变项则为同形的两项，分别在框架的前后。如A就是、A什么A等。

(4)单项单框式。框架只有一项，可变项也只有一项，可在框架项之前或之后。如：都是A、到底是A等。邵敬敏(2016)指出此类属于非典型框式结构。

3.3　汉语语块分类的语用标准

语块的语用标准关注的是语块实现的功能，包括组织话语信息、构建连贯的文本、实现元话语功能、表达人际意义等。从语用功能标准来看，语块分为话语标记和社交客套语两种类型。

3.3.1　话语标记

话语标记语在语法层次上包括词、短语和句子，其中以小句性单位居多。本文的研究对象为语块，因此主要指的是小句类的话语标记语，比如“这样看来”“如此说来”“你比如说”“说老实话”“不瞒你说”“在我看来”“还有就是”等。

根据我们对语料库的语块提取结果，话语标记型语块的数量不算很少，包括“怎么说”“怎么说呢”“怎么样”“只不过”“总的来说”“你要知道”“说老实话”“坦白讲”“完了以后”“我觉得”“想想看”“要不然”“也就是说”“一般来说”“一句话”“意思就是”“应该说”等，涵盖了语篇的构建功能、人际关系的调整功能以及指向功能等各个方面。

3.3.2　社交客套语

在常用语块中，除了根据常规的语言规则生成的话语之外，还有一些不经规则生成的、习用性的社交套语，以适应不同的交际意图或交际环境。社交客套语对语境有较高程度的依赖性，在同一类型语境中会被本族语使用者反复使用，从而获得一种规约性的语用意义，比如“不好意思”“请留步”“生日快乐”“旅途愉快”“恭敬不如从命”等。另一种类型的客套语是重复型客套语。比如“恭喜恭喜”“哪里哪里”“久仰久仰”“幸会幸会”“岂敢岂敢”“过奖过奖”等。

客套语语块的运用，强调的不是语法的合法性，而是语用的得体性以及场景的适切性，二语学习者往往会因为按照目标语的语法规则进行常规语言解码而产生语言的误用。

4.语块的特点分析

4.1　语块的语义特点

语块的语义特点表现在其语义透明性程度的差异上。语义的透明性指的是一些多词单位，其整体意义是相关成分各自意义按照通常的语法规则进行意义加合的结果。语块的语义透明性具有梯度性特点。一些语块在语义上完全透明，比如“心理危机”“听音乐”“随身携带”等。一些语块在语义上具有部分透明性特点。其中的一部分是因为出现了字面意义与引申意义的叠合，比如“找钱”“晒幸福”“圈子里”等，其实际使用的意义是其字面意义的引

申。有一些语块的语义半透明性体现在语块的一个组成部分凸显的是常见的基本义，另一个组成成分所凸显的则是引申义或修辞义，比如“上厕所”中的“厕所”呈现的是本义，而“上”体现的则是虚化的泛指义。还有一些语块的语义透明性程度非常低，基本上可以认为属于语义隐晦性特点，比如“怎么着”“以至于”“也没什么”等。随着语义透明性程度的降低，其搭配的语法可分析性程度也在不断地降低。在语法结构基本不可分析之后，语块的凝固性程度也在不断加大。

4.2 语块的语法特点

通过语料库提取到的语块与语言共同体的语言直觉可能并不完全吻合。比如说，对于“感兴趣”“就是说”“环境保护”等多词单位，大部分汉语母语使用者可能会觉得符合“语块”的判定标准，因此这些单位具有意义上的整体性；而从语料库中提取出来的“这样一个”“动不动就”“真的是”等高频共现成分，在语言社团认知的接受度上会有较大程度的降低。但是，这些语块恰恰是语料计量分析之后显示出的高频组构类语块。组构性语块在传统的语言学研究中通常被忽略（Nekrasova 2009），其原因在于：在使用频率相当的情况下，意义相对自足的短语语块比非短语语块在心理上更具有认知凸显性（Nesi & Basturkemen 2006）。许莹莹、王同顺（2015）认为，“结构完整性影响中低水平学习者的语块加工和表征，即使是高频出现的非短语语块也可能不具有与传统语法上的短语一样的地位”。

4.3 语块的功能性特征

不同的语块在文本或话语中担负不同的功能。一部分语块侧重表达的是语义内容，一部分语块侧重表达的是句法功能，比如“从来不”“从来没”等，揭示的是其高频组合成分以及否定性的语言环境。一些框式结构，比如“何必……呢”之类，作为待填充式结构，其句法填充功能更为明显。另一部分语块表达的是语用功能，所实现的是语篇的衔接、话题的延续或转换以及人际意义的实现等，比如“对我来说”“而且我觉得”等，既引入了新的话题，同时也表示了自己的态度。

语块的上述功能并不是互斥的，一些语块会同时兼具多种功能。

5.结语

本文采用语料库语言学的方法，着眼于对外汉语教学，从语义、语法、语用等维度对汉语语块进行了分类。在汉语教学中，习语类语块作为语块的核心成员，一直被关注，也自然而然被视为整体性单位进行学习。而搭配类语块在传统语言学研究中因为技术手段的限制，在课堂教学中经常被忽视。我们认为：与习语单位一样，高频率、高互信息值的搭配类语块在语言使用中的整体提取性应该得到重视，它们是语块的重要组成部分。特别是跨层类组构语块，更是语料库语言学视角下语言系统中的一个重要聚类。从功能角度看，跨层类组构语块是小句生成以及语篇构建的框架式构件。“学习者对结构不完整语块的习得和加工最初可能受制于此类语块的高灵活性以及低凸显性，具有一定困难”（许莹莹、王同顺 2015）如果从“高频优先”的二语教学策略看，高频共现、较难习得的组构类语言单位恰恰是教学中应该受到重视的。它们在对外汉语教学中应拥有自己的一席之地。

参考文献：

董秀芳. 1997. 跨层结构的形成与语言系统的调整[J]. 河北师范大学学报：社会科学版，(3).

董秀芳. 2011. 词汇化：汉语双音词的衍生与发展[M]. 北京：商务印书馆.

亓文香. 2008. 语块理论在对外汉语教学中的应用[J]. 语言教学与研究，(4).

钱旭菁. 2008. 汉语语块研究初探[J]. 北京大学学报：哲学社会科学版，(5).

邵敬敏. 2016. 现代汉语通论（第三版）[M]. 上海：上海教育出版社.

王立非，张大凤. 2006. 国外二语预制语块习得研究的方法进展与启示[J]. 外语与外语教学，(5).

王文龙. 2013. 对外汉语初级阶段语块构建研究[D]. 北京大学博士学位论文.

吴勇毅，何所思，吴卸耀. 2009. 汉语语块的分类——语块化程度及其教学思考[A]. 第九届世界华语文教学研讨会论文集第二册：语言分析2[C]. 台湾：世界华文出版社.

许莹莹，王同顺. 2015. 语块频率、结构类型及英语水平对中国英语学习者语块加工的影响[J]. 外语教学与研究，(3).

薛小芳，施春宏. 2013. 语块的性质及汉语语块系统的层级关系[J]. 当代修辞学，(3).

周健. 2007. 语块在对外汉语教学中的价值与作用[J]. 暨南学报：哲学社会科学版，(1).

Altenberg, B. 1998. On the phraseology of spoken English. The evidence of recurrent word-combinations [A]. In Cowie, A. P. (Ed.), Phraseology. Theory,

Analysis, and Applications[C]. Oxford: Oxford University Press.

Cortes V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology[J]. English for Specific Purposes, (23).

Hunston, s. 2002. Corpora in Applied Linguistics [M]. London: Cambridge University Press.

Langacker R. 2008. Cognitive grammar as a basis for languageinstruction[A]. InRobinsonP & Ellis N (eds.). Handbook of Cognitive Linguistics [C]. NewYorkandLondon: Routledge.

Nekrasova. T. 2009. English L1 and L2 speakers' knowledge of lexical bundles[J]. Language learning, (59).

Nesi, H. & H. Basturkmen. 2006. Lexical bundles and discourse signaling in academic lectures. International Journal of Corpus Linguistics[J], (11).

Sinclair, J. 1991. Corpus, Concordance and Collocation. [M]. Oxford: Oxford University Press.

Stenfanowitsch, A. & S. Gries. 2006. Corpus-based approaches to metaphor and metonymy[M.] Berlin and New York: Mouton de Gruyter.

Wray. A. 2002. Formulaic Language and the Lexicon [M]. Cambridge: Cambridge University Press.

A Corpus-based Study on Chinese Language Chunks Classification

WANG Feng-lan[1], YU Pingfang[1], XU Kun[2]

(1.Institute for International Students Education, Guangdong University of Foreign Studies ,Guangzhou 510420,

2Shanghai Industrial Technical School, Shanghai 200231, China)

Absract: Based on the Corpus research method and starting with the probabilistic characteristics of language use, this paper explores the usage frequency of relevant words sequence and mutual collocation strength under modern Chinese system. From the perspective of teaching Chinese as a foreign language, the author classifies the lexical chunks in terms of semantics, syntax and pragmatics and other multi-dimensional angles. Based on which, this paper conducts the analysis on the characteristics of lexical chunks from the angles of linguistics and psychology respectively. This paper concludes that the overall extracting of collocation chunks featuring high frequency and high mutual-information value, as an important composition of chunks, should be attached great importance in language use. Cross-layer chunks, in particular, are more important clusters in the language system of Corpus linguistics perspective.

Key words: Chinese; chunks; classification; Corpus

【责任编辑:王巧玲】