国外程式语识别研究概述

李更春

（浙江广播电视大学外国语学院，浙江杭州，310030）

摘要：程式语是一种集词汇特征、语法结构、语义和/或语用功能为一体的多词单位。近些年来，程式语受到了越来越多的关注与研究。学者们从不同的研究背景（如语料库语言学、语用学、语言习得、心理语言学、认知语言学）出发，对语言使用中的程式语进行了多方位的研究。但该研究领域中的一些基本问题如术语、定义和识别标准至今还没有形成定论。尤其在如何识别程式语这一问题上，不同的研究者往往使用不同的判别标准。对国外程式语研究中主要的识别标准进行梳理和分析，指出各种标准的优点和缺点，提出要根据具体的研究对象，针对语料的大小，采取客观与主观的标准相结合的方法，制定出一套具体的识别标准，从而准确、可靠、全面地识别出语言使用中的程式语。

关键词：程式语；识别标准；惯例化；固定性；非组构性；频率

中图分类号：H319 **文献标识码**：A **文章编号**：1672-3104(2012)05−0221−07

一、程式语的定义

传统上，词汇通常被认为是一个个的单词。当然，这种观点现在已被证明是不恰当的，因为词汇还包括许多大于正字法单词的语言单位，如 give up, burn the midnight oil, one of the most 等。对大型语料库的分析表明，这些多词词汇单位(multi-word unit)至少在英语中[1]，尤其在口头话语中是无处不在的[2]。实际上，大量的研究表明，语言在很大程度上是由程式化成分构成的[3−10]。杨玉晨[11](24)甚至认为，自然话语中的 90%是由那些处于单词和固定短语之间的半固定的“板块”结构来实现的。不同的学者赋予了这种语言结构以不同的名称，如 lexical phrases[12]、lexical chunks[13]、multi-word items[1] 、 lexical bundles[7] 、 formulaic sequences[14−16] 、 multi-word units[17] 、 morpheme equivalent unit[18]等。这些术语的定义虽各不相同，但亦有重叠之处，所指称的现象也是同中有异，较易混淆。此外，学者们在对什么是程式语这一问题上意见不一，在程式语的识别标准上也是各持己见。这在一定程度上阻碍了不同学科中程式语研究的交流与借鉴，使人们难以全面地认识和领会程式性现象的本质和功能并自觉地应用于语言理解、产出与习得等过程中。本文拟对程式语的各种识别标准进行梳理和评析，以期为进一步的研究打下一定的基础。

Wray 和 Perkins[19](1)从心理表征的角度将“程式语”(formulaic sequences)定义为“一个由单词或其他意义成分组成的连续或非连续的序列，该序列是或看起来是预制的，即在使用时是从记忆中整体提取或存储的而不是通过语言的语法生成或分析的”。该定义既涵盖了那些具有高度习用性且不可改变的单词串（如 by and large）又包括了那些语义上透明、句法上灵活、含有空槽且其中可填入开放词类的单词串（如 NP be-TENSE sorry to keep-TENSE you waiting）。此后，Wray[16](9)对上述定义进行了修正，将“意义”一词从该定义中删去。在该定义的基础上，我们将程式语重新定义为：某言语社团以整体形式辨识、存储和提取，连续或非连续、具有较完整的结构与意义/功能或参与语篇构建、使用频率较高的多词单位。该定义既涵盖了传统语法所认可的程式化表达（如习语、谚语、短语/词组以及其他类别的固定表达），又涵盖了通过语料库驱动方法识别的多词单位或词束。

二、程式语的识别

程式语在构成与功能上非常多样，它们可以是填

收稿日期：2011−12−12；**修回日期**：2012−05−04
基金项目：浙江广播电视大学科研启动基金项目(KY050112121003)
作者简介：李更春(1983−)，男，安徽庐江人，博士，浙江广播电视大学外国语学院讲师，主要研究方向：应用语言学.

充词（如 sort of），简单的功能短语（如 excuse me），词语搭配（如 tell a story），习语或谚语（如 back to square one, let's make hay while the sun shines），以及较长的标准化短语（如 there is a growing body of evidence that……）。正是因为其存在形式与语言功能的多样性，所以目前很难对该现象下一个全面的定义，这种对定义的缺乏一直是该领域亟待解决的问题之一。在这种情况下，如何准确、可靠、全面地识别出语言使用中的程式语呢？不同的研究者往往采用不同的标准进行判定。我们大致可以将这些标准分为 3 类：① 基于语言学特征的识别标准[1, 10, 19−20]；② 基于使用频率的识别标准[7, 21−27]；③ 基于学习者语言产出的识别标准[28−32]。下文将分而述之。

（一）基于语言学特征的识别标准

有些学者通过对程式语语言学特征的研究，总结出判别程式语的一些标准。

Moon[1](44)认为，程式语具有以下 3 个特征，即：① 惯例化(institutionalization)，指的是某多词项目被该言语社团视作一种单位的程度；② 固定性(fixedness)，指的是某多词项目作为一种单词序列而凝固的程度；③ 非组构性(non-compositionality)，指的是某多词项目不能以单词为基础进行解读而具有一种特殊的整体意义的程度。其中，固定性与非组构性（或非透明性）是学者们普遍接受的标准，但这些标准被证实为一种连续体，这种情况增加了区分程式化与非程式化表达的难度[33](33)。Hudson[20](8−9)总结了有关固定表达(fixed expressions)研究中经常使用的 4 种标准：① 对其组成部分的意料之外的句法限制，即那些我们通常不会预料到的对句法变异性(syntactic variability)的限制，例如：数：the other day，*the other days（对比 the other boy – the other boys）；冠词：strike a light!，*strike the light!（对比 strike a match – strike the match）；词序：trials and tribulations，*tribulations and trials（对比 sorrow and pain – pain and sorrow）。② 对该表达中的单词在搭配上的意料之外的限制，即与我们通常预料的不同，该表达的某部分并不能被来自相关集合的项目所替换，例如：first of all，*second of all（对比 first in line – second in line）；above board，*below board（对比 above standard – below standard）；disaster area，*catastrophe area（对比 major disaster – major catastrophe）；how do you do，*how do they do（对比 how do you do it – how do they do it）；类似地，对此类表达的修改通常也受到了限制，例如：for good，*for very good；kick the bucket，*kick the plastic bucket。③ 反常的句法或用法，即固定表达有时候是不能就其组成部分来分析的，例如：all of a sudden（形容词 sudden 在这里被用作名词）；spick and span（spick 和 span 在当代英语中已不再使用）。④ 比喻意义。即其部分与整体之间存在着一种语义上的失配(mis-match)，如 a red herring, a hot potato, to kick ass, not much cop, at the end of the day。例如，在 a red herring 这个固定表达中，red 和 herring 的意义相加并不能得到 red herring 的意义。Hudson 指出，这些标准有时候是互相矛盾的。例如，(to) sow wild oats 根据“比喻意义”标准可以被视为固定的，但根据前两个标准则并非如此。换句话说，其组成部分的意义相加不能得到其整体的意义，但该表达容许很大程度上的变化，与此相反，像 all of a sudden 这样的表达在语义上并不是模糊的，然而根据可变性标准则是完全固定的。基于这样的考虑，Hudson 将其对固定性的定义建立在可变性标准①和②上。

Wray 和 Perkins[19](5)认为，很多程式语并不是通过语义组构的(semantically composed)，而是在句法上不规则的整体项目，如 straight from the horse's mouth 或 to pull someone's leg。这种不规则性表现在两个方面：① 对句法操控(syntactic manipulation)的限制，例如，我们不能使用 beat around the bush 的复数形式，不能使用 face the music 的被动形式，也不能说 you slept a wink 或 feeding you up 等。② 对正常约束的违反，例如，那些具有“不及物动词+直接宾语”结构的单词序列（如 come a cropper, go the whole hog）或其他违反句法规则的语言形式（如 by and large）。此外，Wray 和 Perkins[19](1)将“程式性”定义为“在一串语言项目中，每个项目与剩下部分的关系是相对固定的，且将某项目替换为同一类别的项目相对受到制约”。可见，Wray 和 Perkins 是从句法的不规则性和形式的（相对）固定性来界定程式性现象的。

Van Lancker-Sidtis 和 Rallon[10]指出，程式化表达在以下几个方面区别于新创话语：① 它们经常包含一些非字面或非标准意义的词汇项目（如 It broke the ice; just in the nick of time）。② 它们一般具有某种态度或情感上的暗指，而新创话语在情感内容上可以是中立的。例如，she has him eating out of her hand 暗含着“屈服”“依靠”与“情感依恋”等意义，类似地，have a nice day 蕴含着“愉快”的意义。相比之下，新创表达 the cat is on the chair 需要以标记性的语调说出或增加形容词才能表达特定的情感意义。③ 它们具有一种语音上的连贯性，而新创话语则不具有这种连贯性。此外，它们的措辞和语序都是固定的，语调也往往一成不变，因为对句子重音的选择是有限制的。例如，

在 I wouldn't want to be in his shoes 中，如果我们将句子重音放在 shoes 上，那么这句话听起来就不是那么地道或规范。④ 程式化表达是“似曾相识的”，因为本族语者会认出它们具有这种特殊的地位。例如，See you later!或 Let's call it a day（言语公式）；I'd like to give you a piece of my mind 或 He's at the end of his rope（习语）；以及 Look before you leap 或 He who hesitates is lost（谚语）都是“似曾相识的”，因为本族语者能够识别出这些话语、对它们进行补充（当省略部分单词时）以及知道它们的特殊意义以及所适合的语境。

（二）基于使用频率的识别标准

Butler[21](76)根据其对西班牙语语篇基于频率的研究，提出了一种结构或形式上的标准，即，“大多数较长的重复的序列是以连词、冠词、代词、介词或语篇标记词开头的”。然而，这一发现有待于进一步的考察。对一个语篇基于直觉的考察可能会使我们相信，一个以某个介词为首个固定成分的单词序列实际上是以可填入开放词类（如名词或动词）的空槽开头的。例如，NP_i be-TENSE past PRO_i-POSSESSIVE sell-by date（如 This cheese is past its sell-by date, Dad is past his sell-by date）可能被表征为 past PRO_i-POSSESSIVE sell-by date，但由于该句中充当主语的名词短语必须与代词同指(co-indexed)，所有它对整个序列来说是必不可少的。但由于该空槽中的内容是可以变化的，所以单靠语料库搜索无法将其识别为某个复现序列的一部分。所以，Butler 的观察只是告诉我们，一个复现序列的首个固定的单词往往是功能词或语篇标记词，而不是说它一定是以这样的单词开头的。

Biber 和他的同事们主要通过语料库驱动方法来识别话语和/或语篇中的程式语[7, 22–27]。他们将出现频率高于某临界值且出现在一定数量语篇中的复现单词组合判断为某种程式语或词串。但不同的研究者往往根据不同的研究目的和语料库的大小而设置不同的标准来识别语料中的词串。例如，Biber 等人[7]将出现频率不低于 10 次/mw 且至少出现在 5 个不同语篇中的单词序列视为某种词串；Cortes[24]将该频率截断点提高至 20 次/mw；Biber 等人[23]更是将之提高到 40 次/mw。因此，不同的研究得出的结果往往存在着较大的差异。例如，Altenberg[5]统计了 London-Lund 语料库中至少出现两次的单词组合，发现这些复现组合超过了该语料库词容的 80%；Biber 等人[7]发现，3-4 词的词串占他们研究的会话的 28%，学术散文的 20%；据 Erman 和 Warren[8]计算，各类程式语要占到其分析的英语口头语篇的 58.6%，书面语篇的 52.3%。可见，语言在很大程度上是程式化的，但 Butler[21]和 Moon[34]的研究结果与此结论相差甚远。Butler 发现，在他研究的西班牙语语料库(10 000 单词)的口语部分中只有 12.5%的语言是程式化的，在两个转写的访谈中（均为 14 000 词）只有 9%与 8.2%的语言是程式化的。Moon 研究了超过 1 800 万词的 Oxford Hector Pilot Corpus，发现只有 4%~5%的语言是程式化的。

（三）基于学习者语言产出的识别标准

有些学者对学习者的口头产出进行了研究，在此基础上提出了判别公式语的另一些标准。

Coulmas[28]指出，某单位必须至少具有两个词素的长度，且在语音上是连贯的才能被识别为某种公式语。他还指出，公式语在语法上可能比其他语言更加超前，相对于学习者产出的常规语言来说在句法和语音上更为复杂。

类似地，Peters[29]也提出了识别学习者语言中的公式语的一些标准，即：① 语音连贯性(phonological coherence)；② 相对其他输出来说更长且更复杂；③ 情景依赖性(situational dependence)；④ 高频性与形式的不变性(invariance in form)。

Weinert[31]在 Peters 的基础上对语言习得研究中常见的标准进行了总结。这些标准是：① 语音连贯性，即流利的、没有迟疑的编码(fluent, non-hesitant encoding)，语调升降曲线(intonation contour)没有中断；② 与学习者其他的输出相比更长且更复杂；③ 在序列中没有能产规则的使用情况；④ 在整个社团内广泛使用(community-wide use)；⑤ 特质性/不合适的用法；⑥ 情景依赖性；⑦ 高频性与形式的不变性。

此外，在语言产出研究中，学者们还使用了一些与流利性相关的时间变量(temporal variables)如语速(speech rate)、停顿(pauses)和迟疑现象(hesitation phenomenon)作为公式语的判断标准。其中后两种是程式化单位的边界标记(boundary markers)[35]（转引自 Weinert，1995：183）。

三、对以往识别标准的评析

如前所述，识别程式语的一种标准就是某语言形式的惯例化程度。诚然，某些程式语（如习语、谚语或其他固定表达）由于其深厚的历史和文化根基在某言语社团内具有很高的惯例化程度，因此，本族语者也是很有可能将之以整体形式存储和表征于心理词库中的。但语言形式的惯例化程度也是有等级之分的，Howarth[36](28)认为，不同的单词组合在惯例化程度上

构成了一种连续体：纯习语(pure idioms)的惯例化程度最高，比喻性习语(figurative idioms)次之，限制性搭配(restricted collocations)的惯例化程度中等，自由组合(free combinations)的惯例化程度最低。但是，语言形式的惯例化程度不是固定不变的，而是动态变化的，因此不同范畴之间并不存在明确的界限。此外，使用该标准难免也涉及到一些主观因素，不同的研究者在某语言形式的惯例化程度上可能有着不同的判断。

识别程式语的另一种标准是通过它们的内部构成或组构性来加以判断。的确，很多程式化表达具有的一种特质就是它们的非组构性，也就是说，它们的整体意义不能通过其词汇成分的意义推断出来。这种非组构性或不透明性根源于某言语社团语言实践的演变过程，被认为是程式语的主要特征之一[34, 37−38]。这种方法的核心思想为：一旦某单词串程式化之后，它就不受语法规则与词汇搭配的制约，即该单词串并不一定要合乎语法和/或符合语义逻辑。单词串在固化或僵化之后常常保留了那些当前并不使用的单词或语法形式（如 by dint of , If I were you）。因此，在极端情况下，为了达到对某非组构性序列的正确理解，我们只能将其加以整体记忆。然而，这种识别方法过于保守，因为它排除了那些形式规则、语义透明的程式语。

我们还可以通过固定性来识别程式语。这种识别标准建立在这样一种认识之上，即真正的习语在形式上是完全固定的。Pawley[39](107−108)通过比较 first (and only) attempt 与 first (*and only) aid 表明，这种插入成分影响了程式化表达的地道性。再如，短语 lead up the garden path 是一个比喻性的程式语，但 *lead, happily singing, up the winding garden path 即使当作比喻性的表达来理解也失去了它原来的意思。对于某些程式语来说，这种以及其他类似的固定性测试也许能起到一定的作用，但是这种作用显然是有其局限性的，正如 Wray[16](34)所言：“我们如此地习惯于进行语言游戏（无论是对程式化的还是对非程式化的语言），以至于我们可能并不清楚哪些变化将影响到以及哪些变化并不会影响某单词序列的完整性，因为在今天还是新颖的、俏皮的改写可能在明天就变成程式化的了，如果人们接受了该表达并重复使用了几次。”此外，只有少数程式语是完全固定的，语言中存在大量半固定的单词序列，在这些序列中存在一个或多个空槽，说话者可根据表达的需要填入不同的语言成分（如单词，短语或小句）使之适用于不同的使用场合。

另一个经常使用的标准就是出现频率。人们假定，如果某单词序列在语料库中出现的频率很高，这就表明至少在某种程度上它在该言语社团中是惯例化的。在语料库研究中，越来越多的研究者通过计算机对口头和/或书面语语料库进行搜索来找出其中的单词串，这些单词串因其重复出现的特点（即高频率）往往被划归为程式语。从表面上看，运用计算机搜索来识别常见的单词串并将某频率阈限(frequency threshold)作为判断某单词串是否是程式化的标准，这种做法是完全有道理的。原因是，我们越是经常使用某单词串，那么为了节省加工努力(processing efforts)，该单词串以预制的形式(prefabricated form)存储的可能性就越大，而在它以预制的形式存储后，当我们需要再次表达相同的信息时，它就会成为人们偏好的选择。这种对预制形式的偏好实际上将压制(repress)其他可能的表达方式出现的频率，因此，它们之间在频率上的差异应该是显而易见的。然而，单纯的频率标准也容易产生一系列的问题。第一，当检索工具忽略句子或主要成分(major constituent)之间的边界、说话人的变化(change of speaker)，错误的启动(false start)等因素时，我们也许要运用结构标准来删除那些在短语学上没有意义的单词组合。第二，口头语料常常含有对迟疑现象的转写（如 erm、er 等），这时研究者就必须决定这些现象是否构成单词。第三，该标准可能会纳入一些只有“十五分钟热度”(fifteen minutes of fame)的或有话题或文体偏向性的表达法[40](50)。第四，语料也许并不能反映某类程式语的真实分布情况。不容置疑的是，他们提供的信息要比我们通过直觉获取的信息好得多。然而，小型语料库的选择性(selectiveness)可能会将某些常见的但不容易搜集到的语言材料排除在外。有些语言在本族语者看来明显是程式化的，但其在语料库中的出现频率并不高，如 Long live the king!, All for one and one for all。第五，对多词表达的长度及其频率阈限的确定具有随意性，而且频率测量不能区分某单词串是程式性的还是创造性的[16, 33]。第六，研究者们对从语料库中识别的多词表达提出了多种频率与关联比(association ratio)的标准[16, 33]，这进一步增加了通过语料库研究程式语的主观性。最后，使用语料库方法识别程式语的另一个问题与这些单词序列的可变性(variability)或灵活性(flexibility)有关。Schmitt 和 Carter[41]指出，虽然由固定成分组成的程式化表达识别起来更容易，但是可变的序列难以使用目前的计算机软件来识别，即使这类表达要比某些固定表达的频率更高。在这种可变的序列中，除了某些固定成分之外，还有一些需要根据特定的语境填入特定的词（或短语、小句）的空槽。与此类似的是，还存在一些序列，其成分在语篇中的距离比较远，难以使用语料库软件对其进行识别。最后，语料库工具并不能区分出

程式语的边界，这也是我们运用本族语者直觉进行识别时所遇到的问题。

基于学习者语言产出的识别标准也存在一些问题。虽然已有证据表明，程式语在产出时具有一种完整的语调升降曲线，但这种证据经常是印象性的(impressionistic)，且只适用于对二语产出进行判断[42](434)。此外，Peters[29]指出，学习者往往通过融合(fusion)将某些经由语法规则构造的话语当做一个整体单位加以使用。这些表达虽然不为本族语者所使用，但能使他们达到交流的目的[16]。这一发现似乎与Weinert[31]总结的第三和第四个标准有所出入(即“在序列中没有能产规则的使用情况”，“在整个社团内广泛使用”)。此外，第四个标准和第五个标准似乎也是互相矛盾的。某语言形式不可能既“在整个社团内广泛使用”又是某学习者“特质性的/不合适的用法”。总之，在这些标准中，并不是每条都绝对适用于判别学习者语言产出中的程式语，它们在适用的程度上可能并不相等；此外，不同语言水平的学习者所适用的标准也是不同的，例如，语言水平较高的学习者可能适用①②③④⑤⑥⑦，而语言水平较低的学习者可能使用①②③④⑤⑥⑦。当然，到底适用什么样的标准应取决于具体的研究对象和研究方法，不能一概而论。

针对以往程式语识别研究存在的各种问题，Wray[16](43)指出，当我们对话语中的程式语进行识别时，“情况可能是，识别不能建立在单一的标准上，而需要利用一整套特征”。最近，Wray 和 Namba[43](29–32)及 Wray[18](116–125)提出了识别程式语的 11 种标准，分别是：① 该单词串在语法上存在着不同寻常的地方（即语法的不规范性）；② 该单词串的部分或整体缺乏语义透明性（即语义的不透明性）；③ 该单词串与某一具体的情景和/或语域联系在一起（即使用语域的专门性）；④ 该单词串作为一个整体除了表达其成分单词本身的意义之外，在交际或话语中还执行了某种功能（即语用功能）；⑤ 某说话者/写作者在表达该思想时最经常使用这种确切的表达方式（即个人使用频率）；⑥ 某说话者/写作者在说出或写出该单词串时伴随有一种行为、标点符号的使用或某种语音型式，赋予该词串以某种单位的特殊地位和/或重复他/她刚听到或读到的东西（即伴随的语言或非语言的标记）；⑦ 某说话者/写作者或其他人以某种方式对该单词串进行了语法或词汇上的标记，赋予了该词串以某种单位的特殊地位（即语法或词汇标记）；⑧ 基于直接的证据或我的直觉，有一种大于随机水平的可能性是，某说话者/写作者在之前与他人交流时遇到过这种确切的措辞（即似曾相识性）；⑨ 虽然该单词串是新创的，但显然它是某种程式化表达的有意或无意的派生（即派生性）；⑩ 虽然该单词串是程式化的，但它被无意地用错了地方（即使用不得体性）；⑪ 该单词串包含了某种过于复杂或过于简单的语言材料，以至于不能和说话者总体的语法和词汇能力相匹配（与使用者语言水平不符的特点）。Wray 和 Namba 建议，针对不同的研究对象需要使用不同的标准组合来识别其语言使用中的（可能的）程式语。我们认为，程式语难以识别的原因在于该现象的主体性差异。换句话说，某语言形式是否为程式语既不是语法学家规定的，也不是通过出现频率进行判定的，而要取决于语言使用者本身。实际上，每个人都有自己独特的个人语型(idiolect)，由其个人的语言库(repertoire)所构成，而作为个人语型的构成要素，程式语可能也会因人而异，随着其经验与语言接触的差异而有所不同。这种“程式语型”(formulalect)或“短语语型”(phrasalect)既包括了某言语社团大部分成员作为整体存储的程式语，又包括了言语社团成员并非作为整体存储的程式语。正如一个人的心理词库是由某些独特的单词构成的，该心理词库可能也包括一些独特的程式语。

四、余论

从以上的论述中不难看出，以往的研究多根据语言形式特征并结合主观判断来对语言使用中的程式语进行识别，这种识别方法往往费时低效，而且其识别的结果有时也存在着一定的可信度问题。语料库语言学的兴起和发展给程式语研究带来了很大的便利，人们可以借助语料库软件轻松地识别出语言使用中的程式语，但该方法也是有其局限性的。此外，对小文本自建语料中程式语的识别，目前仍没有一套完全可行并行之有效的标准。实际研究中使用较多的是通过本族语者直觉来对程式语进行识别。为了将这种方法中主观因素的影响降低到最低程度，研究者往往使用若干名本族语者（或经验丰富的语言教师）作为评判员，要求他们独立地识别出语料中（可能的）程式语，然后将他们共同识别的单词组合视为某种程式语[10][44]。但这种方法也是有其局限性的。首先，该方法只限于对少量语料的分析。其次，长时间作业中的一些人力因素（如劳累、注意力不集中）可能会导致识别标准前后不一。再次，程式语可能会出现相互嵌套(embedding)的现象，因此，即使识别标准是完全一致的，仍有可能会出现模棱两可的情况。最后，直觉并不总能全面地反映语言事实，某些常见的、非常熟悉

的程式语可能并不会引起识别者的注意。我们认为，应该将主观的本族语者和/或语言教师的直觉与客观的标准相结合，根据不同的研究对象（如本族语儿童/成人或以英语为二语/外语的儿童或成人）、语料的大小和/或具体的研究方法与目的来制定一套具体的识别标准，只有这样我们才有可能准确、可靠、全面地识别出语言使用中的程式语。

参考文献：

[1] Rosamund Moon. Vocabulary connections: Multi-word items in English [C]// Vocabulary: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press, 1997: 40−63.

[2] Michael McCarthy, Ronald Carter. Written and spoken vocabulary [C]// Vocabulary: Description, Acquisition, and Pedagogy. Cambridge: Cambridge University Press, 1997: 20−39.

[3] Helen B. Sorhus. To hear ourselves—implications for teaching english as a second language [J]. English Language Teaching Journal, 1977, 31: 211−221.

[4] Ray Jackendoff. The boundaries of the lexicon [C]// Idioms: Structural and Psychological Perspectives. Hillsdale, NJ: Lawrence Erlbaum, 1995: 133−166.

[5] Bengt Altenberg. On the phraseology of spoken English: The evidence of recurrent word-combinations [C]// Phraseology: Theory, Analysis, and Applications. Oxford: Clarendon Press, 1998: 101−122.

[6] Peter Howarth. The Phraseology of Learners' Academic Writing [A]. Phraseology: Theory, Analysis and Applications[C]. Oxford: Oxford University Press, 1998: 161−186.

[7] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan. Longman Grammar of Spoken and Written English [M]. Harlow: Longman, 1999.

[8] Britt Erman, Beatrice Warren. The idiom principle and the open-choice principle [J]. Text, 2000, 20: 29−62.

[9] Pauline Foster. Rules and Routines: A Consideration of Their Role in the Task-based Language Production of Native and Non-native Speakers [A]. Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing[C]. Harlow: Longman, 2001: 75−93.

[10] Diana Van Lancker-Sidtis, Gail Rallon. The incidence of formulaic expressions in everyday speech: methods for classification and verification [J]. Language and Communication, 2004, 24(3): 207−240.

[11] 杨玉晨. 英语词汇的“板块”性及其对英语教学的启示[J]. 外语界, 1999(3): 24−27.

[12] James R. Nattinger, Jeanette S. DeCarrico. Lexical phrases and language teaching [M]. Oxford: Oxford University Press, 1992.

[13] Michael Lewis. The lexical approach: The state of ELT and a way forward [M]. London: Language Teaching Publications, 1993.

[14] Alison Wray. Formulaic language in learners and native speakers [J]. Language Teaching, 1999, 32(4): 213−231.

[15] Alison Wray. Formulaic sequences in second language teaching: Principle and practice [J]. Applied Linguistics, 2000, 21(4): 463−489.

[16] Alison Wray. Formulaic language and the lexicon [M]. Cambridge: Cambridge University Press, 2002.

[17] Lynn Grant, Laurie Bauer. Criteria for re-defining idioms: are we barking up the wrong tree [J]. Applied Linguistics, 2004, 25(1): 38−61.

[18] Alison Wray. Formulaic language: Pushing the boundaries [M]. Oxford: Oxford University Press, 2008.

[19] Alison Wray, Michael R. Perkins. The functions of formulaic language: An integrated model [J]. Language and Communication, 2000, 20: 1−28.

[20] Jean Hudson. Perspectives on fixedness: Applied and theoretical [M]. Lund: Lund University Press, 1998.

[21] Chris S. Butler. Repeated word combinations in spoken and written text: Some implications for functional grammar [A]. A Fund of Ideas: Recent Developments in Functional Grammar [C]. Amsterdam: IFOTT, University of Amsterdam, 1997: 60−77.

[22] Douglas Biber, Susan Conrad. Lexical bundles in conversation and academic prose [A]. Out of Corpora: Studies in Honour of Stig Johansson [C]. Amsterdam and Atlanta: Rodopi, 1999: 181−190.

[23] Douglas Biber, Susan Conrad, Viviana Cortes. *If you look at…*: Lexical bundles in university teaching and textbooks [J]. Applied Linguistics, 2004, 25(3): 371−405.

[24] Viviana Cortes. Lexical bundles in published and student disciplinary writing: Examples from history and biology [J]. English for Specific Purposes, 2004, 23(3): 397−423.

[25] Douglas Biber, Federica Barbieri. Lexical bundles in university spoken and written registers [J]. English for Specific Purposes, 2007, 26: 263−286.

[26] Ken Hyland. *As can be seen*: Lexical bundles and disciplinary variation [J]. English for Specific Purposes, 2008, 27: 4−21.

[27] Douglas Biber. A corpus-driven approach to formulaic language in English [J]. International Journal of Corpus Linguistics, 2009, 14(3): 275−311.

[28] Florian Coulmas. On the sociolinguistic relevance of routine formulae [J]. Journal of Pragmatics, 1979, 3: 239−266.

[29] Ann M. Peters. The units of language acquisition [M]. Cambridge: Cambridge University Press, 1983.

[30] Tina Hickey. Identifying formulas in first language acquisition [J]. Journal of Child Language, 1993, 20: 27−41.

[31] Regina Weinert. The role of formulaic language in second language acquisition: A review [J]. Applied Linguistics, 1995, 16(2): 180−205.

[32] Florence Myles, Janet Hooper, Rosamond Mitchell. Rote or rule? Exploring the role of formulaic language in classroom foreign language learning [J]. Language Learning, 1998, 48: 323−364.

[33] John Read, Paul Nation. Measurement of formulaic sequences [C]// Formulaic Sequences: Acquisition, Processing and Use. Amsterdam: John Benjamins, 2004: 23−26.

[34] Rosamund Moon. Fixed expressions and idioms in English: A corpus-based approach [M]. Oxford: Clarendon Press, 1998.

[35] Manfred Raupach. Formulae in second language speech production [C]// Second Language Productions. Tübingen, Germany: Gunter Narr Verlag, 1984:114−137.

[36] Peter Howarth. Phraseology and second language proficiency [J]. Applied Linguistics, 1998: 24−44

[37] Chitra Fernando. Idioms and idiomaticity [M]. Oxford: Oxford University Press, 1996.

[38] Michael McCarthy. Spoken language and applied linguistics [M]. Cambridge: Cambridge University Press, 1998.

[39] Andrew Pawley. Lexicalization [A]. Language & Linguistics: The Interdependence of Theory, Data & Application[C]. Georgetown University Round Table on Languages and Linguistics, 1985: 98−120.

[40] 张霞. 基于语料库的中国高级英语学习者词块使用研究[J]. 外语界, 2010(5): 48−57.

[41] Norbert Schmitt, Ronald Carter. Formulaic Sequences in Action: An Introduction[A]. Formulaic Sequences: Acquisition, Processing and Use [C]. Amsterdam: John Benjamins, 2004: 1−22.

[42] Nan Jiang, Tatiana M. Nekrasova. The processing of formulaic sequences by second language speakers [J]. The Modern Language Journal, 2007, 91(3): 433−445.

[43] Alison Wray, Kazuhiko Namba. Formulaic language in a Japanese-English bilingual child: A practical approach to data analysis [J]. Japanese Journal for Multilingualism and Multiculturalism, 2003, 9(1): 24−51.

[44] Jie Li, Norbert Schmitt. The acquisition of lexical phrases in academic writing: A longitudinal case study [J]. Journal of Second Language Writing, 2009, 18: 85−102.

A review of the identification of formulaic sequences

LI Gengchun

(School of Foreign Languages, Zhejiang Radio and TV University, Hangzhou, Zhejiang 310030, China)

Abstract: Formulaic sequences (FSs) are a type of multi-word unit that has integrated lexical features, syntactic structures, semantic and/or pragmatic functions. In recent years, FSs have attracted ever-increasing amount of academic attention and research. Scholars from different backgrounds (e.g. Corpus Linguistics, Pragmatics, Language Acquisition, Psycholinguistics and Cognitive Linguistics) have done a lot of research on FSs from different perspectives. However, some basic issues in the research on FSs, such as terminology, definition and identification criteria, have not been solved. This is especially true of the identification of FSs. Different researchers tend to use different criteria to identify FSs. This article attempts to summarize and analyze the main identification criteria used in overseas research on FSs. It expounds the advantages and disadvantages of various criteria, and suggests that a suite of identification criteria be established, based on the subjects as well as the corpora used, so as to identify the FSs in language use correctly, reliably and comprehensively.

Key Words: formulaic sequences; identification criteria; institutionalization; fixedness; non-compositionality; frequency

[编辑：汪晓]