

英语词法构式内嵌能产性计量研究

——基于齐普夫定律的拟合

杭州师范大学 詹宏伟

提要:判断一个假拟的构式能否确立,需要有一定的依据。在以往的研究中,依据有的来自内省式的语义辨析,有的来自语料和量化数据。这些依据的共同逻辑是:一个确立的构式具有自身的句法语义独特性。本文探讨如何把内嵌能产性作为词法构式确立的依据,通过分析某构词过程在假拟的词法构式内部和外部的能产性,对比两者有无差别;如果有差别,就可以确立该词法构式。本文以英语中假拟的词法构式 $[N_1 + N_2\text{-ed}]$ 为例,从 COHA 语料库中提取每个词族的词频分布表,运用 R 语言的工具包 {zipfR} 建构齐普夫分布模型,用外推法测算复合词“ $N_1 + N_2$ ”的能产性指标,通过比较发现“ $N_1 + N_2$ ”的能产性在 $[N_1 + N_2\text{-ed}]$ 构式内与构式外有显著差别,从而断定 $[N_1 + N_2\text{-ed}]$ 词法构式的确立。本研究的方法和结论对词法能产性和构式形态学研究有一定的启示。

关键词:构式形态学、词法构式、内嵌能产性、齐普夫定律、复合词

[中图分类号] H314.3 [文献标识码] A [文章编号] 1000-0429 (2025) 03-0363-13

DOI:10.19923/j.cnki.fltr.2025.03.004

1. 引言

在构式语法视角下,传统的构词单位和构词方式都可以被看作构式网络中不同层级的构型或构例,词法构式就是词法层面的构式。构式(包括词法构式)是认知语法理论中的重要概念,也是语法系统的基本单位。构式的确立是构式语法研究面临的根本问题之一。

确立一个假拟的构式需要有一定的依据,依据有的来自内省式的语义辨析,有的来自语料和量化数据。词法构式确立的依据主要来源于单词的音系结构、句法语义结构和词素共现。本文探讨如何把内嵌能产性(embedded productivity)作为词法构式确立的依据。“内嵌能产性”指当某种构词过程内嵌于特定的形态或句法环境时,它的能产性受到制约或提升(Booij 2010; Booij &

Audring 2017: 295), 比如在“ $N_1 + N_2$ ”构成复合词之后, 如果被内嵌到了另一个 $[N_1 + N_2\text{-ed}]$ 构式中, “ $N_1 + N_2$ ”的能产性可能会发生变化。内嵌能产性是一种局部的语法特征, 蕴含了构式内部与构式外部的差异。构式形态学的常规研究范式是: 分析某个构词过程在假拟的词法构式内部和外部的能产性, 对比两者是否有差别; 如果有差别, 就可以确立该词法构式。

本文探讨英语中假拟词法构式 $[N_1 + N_2\text{-ed}]$ 能否确立, 具体方法如下: 首先从美国历史英语语料库 (Corpus of Historical American English, 简称 COHA) 中把与 N_2 相同的词项归总成若干词族 (word family), 提取每个词族的词频分布表, 然后运用 R 语言的工具包 {zipfR} 建构齐普夫分布模型 (参见 Evert & Baroni 2007), 最后采用外推法测算“ $N_1 + N_2$ ”的能产性指标, 通过比较 $[N_1 + N_2\text{-ed}]$ 构式内部与外部的能产性差异, 判断该词法构式能否确立。

2. (词法) 构式确立的依据

在认知语言学理论中, 构式是基于使用的语言范畴化的产物, 是人对语言系统的心理表征。构式涵盖不同层级的语言单位, 可大可小, 大到句式, 小到词语乃至词素。构式形态学探究词语的构成, 把词语看作构式, 把多个语素的复杂词 (complex word) 看作含有槽位的图式。随着构式词法在汉语词汇研究中的应用 (宋作艳 2023; 杜凤娇 2024; 司卫国、文旭 2024), 构词也可以进行跨语言的比较研究, 如 Masini, Mattioli & Pepper (2023) 从跨语言角度分析了复合名词的独特性。字词的结构可以揭示英、汉语的时空性差异: 英语构词依靠语素的线性接续, 而汉语构字则基于部件的空间并置或叠置 (刘庚、王文斌 2021)。

关于构式的确立, 尽管缺乏专题研究 (Taylor 2002), 但相关讨论散见于各种具体的研究话题中。在早期文献中, 构式确立的依据来自内省和句法操作, 例如, 研究者通过对存在构式 There BE (Lakoff 1987)、have a V 构式 (Wierzbicka 1988)、双宾构式 (Langacker 1987; Goldberg 1995) 的分析, 发现了这些构式所具有的特殊结构和意义; 后续研究则引入基于语料和量化数据的方法 (Stefanowitsch & Gries 2003; Hampe 2011; Hartmann 2014; Hilpert 2015) 以及心理实验的方法 (Bencini & Goldberg 2000)。这些依据所遵循的共同逻辑是: 一个已经确立的构式拥有其独特的句法和语义属性。这种独特性 (idiosyncrasies) 通常表现为以下几个方面: 首先, 它与近似的构式之间存在不可互换性, 即不能简单地用其他类似构式来替代; 其次, 它在心理层面具有较高的凸显度, 更容易被人们感知和识别; 此外, 它还表现出特定的搭配偏好, 即倾向于与某些词语或短语共现; 最后, 它在语用和语体方面也存在偏好, 更适用于某些特定的语境或文体。

词法构式的确立也有依据,主要来源于单词的音系及韵律特征、形态句法语义结构、词素共现偏好等。Booij (2010)最早提出将内嵌能产性作为词法构式确立的依据之一。例如,荷兰语中复合词构式 $[NV]_V$ 本身的能产性不高,但在 $[[NV]_V -er]_N$ 构式内的能产性较高,说明能产性依赖于特定的构式环境(同上)。然而,目前关于内嵌能产性的讨论多基于主观判断和少量例证,未提供量化的数据支持。

后续的构式形态学研究提供了词法构式的数据,只是未明确使用“内嵌能产性”这一术语。例如, Schönefeld (2013)通过构式搭配强度分析了 *go un-V-en* 构式(用于描述主语处于某种持续的状态,如 *go undone*)的分布,发现该构式多用于学术语体,与口语语体中的 *go Adj* 构式(如 *go mad*)形成互补,表明构式能产性受到语体限制。Hilpert (2015)分析了英语过去分词(*ppl*)在“名词-过去分词”(N-*ppl*)复合词中的内嵌能产性。过去分词是被动句(*be ppl*)和 N-*ppl* 的共有成分,而且意义相同,在传统语法中两者被认为是转换或互换关系,或者说同一种形式的两个变体。事实上,被动句和名词-过去分词复合词在搭配偏好上存在很大差异(同上)。Hilpert (同上:131)进而提出了后续研究问题:“名词-过去分词”复合词的新词创造受什么条件制约?

通过对这些内嵌能产性现象的分析,我们可以发现语言系统各部门间存在互相协同联动,而不是相互独立的模块,同时这些分析也揭示了构式确立的条件:每种构式都具有独特的形态、句法、语音、语义等语言学特征。

3. 将词法能产性看作构式槽位的词素丰富度

本文在构式形态学的框架下探讨动态视角的词法能产性(morphological productivity)的原理和算法。构式语法认为,语言是由网络化的构式单位组成,构式是一种“形式-意义配对体”(form-meaning pairings)(Goldberg 1995),既具有独立于组成成分的整体意义,同时又制约其组成成分的意义。传统的语言单位(如单句、复句、词组、习语、词语等)均被看作构式。

3.1 构式形态学的基本观点

构式形态学通过分析词语内部的音系结构和语义结构以及词语零部件之间的关联,发现它们之间的关系主要包括组构关系(部分-整体)和“父子关系”(图式-样例)。传统的形态规则被重新界定为词法图式:派生过程是以词缀为轴心的图式,而词根则是槽位上的填充项。例如,当 X-able 图式中的 X 槽位被实现为 *believe* 时,就构成 *believable*。图式中槽位的可变性和自由度即体现为词法能产性。

构式具有不同的复杂度和图式性,这使得心理词库呈现层级性。在分层心理词库中,抽象图式统辖其具体样例(Booij 2015: 190)。图式与其样例共存,新图式的出现并不会消除记忆中构成图式基础的原有样例。以英语中的 N-ppl 复合词为例,其心理词库包含四个层级(见图 1),图式化程度从下到上逐层提高。相邻两个层级之间构成图式-样例的关系,例如: N-oriented 既是动源词族 N-deverbal_ppl 的一个样例,同时又是 goal-oriented、consumer-oriented、market-oriented 的图式。图式统辖样例,而样例继承图式特征,因此样例的意义可根据上层图式推导得出。图式的层级类似于概念的层级,层级越低则图式越具体,反之则越抽象。

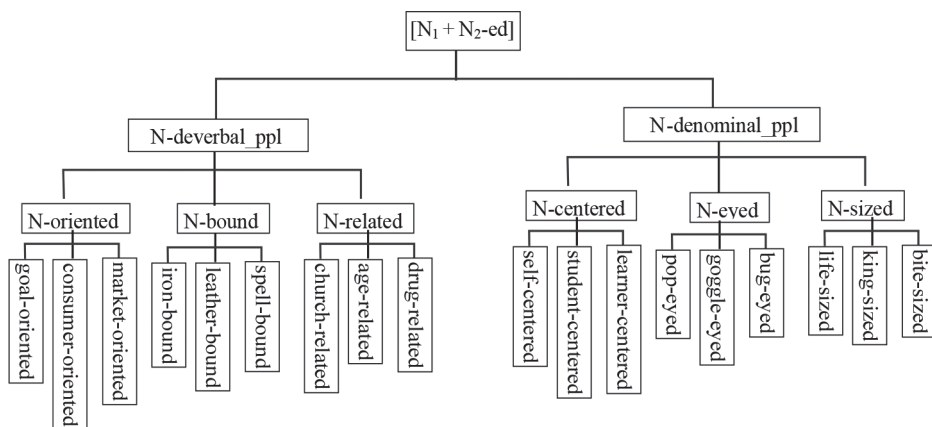


图 1. 词法图式的层级性架构

3.2 词法能产性作为词法构式槽位的词素丰富度

在构式形态学视角下,形态复杂的词语(如派生词、合成词)也被看作构式,准确地说是半填充构式。因此,构式能产性的算法同样适用于词法能产性。构式能产性的一种经典算法是基于词汇丰富度,其基本思路是:把构式某个槽位上所有可能的填充项视为一个小型文本,从而套用文本的词汇丰富度算法来衡量该构式的能产性(Baayen 2009)。

词法能产性的算法同理。需要注意的是,短语构式槽位上的填充项是单词,而词法构式槽位上的填充项是词素。因此,词法能产性可被视为词法构式槽位中词素的丰富度。由于统计词素的频数在技术方面较为困难,故需要将其转化为词族中的词种(type)数和词例(token)数。每个词族的内部成员称作“词种”。例如,若要统计 [X-able] 图式中词素 x_1 - 的频数,需在语料库中检索 x_1 -able 这个词(如 believable)的出现次数(即词例数)。每个词素的频数对应一个词种的出现次数(如图 2 所示)。

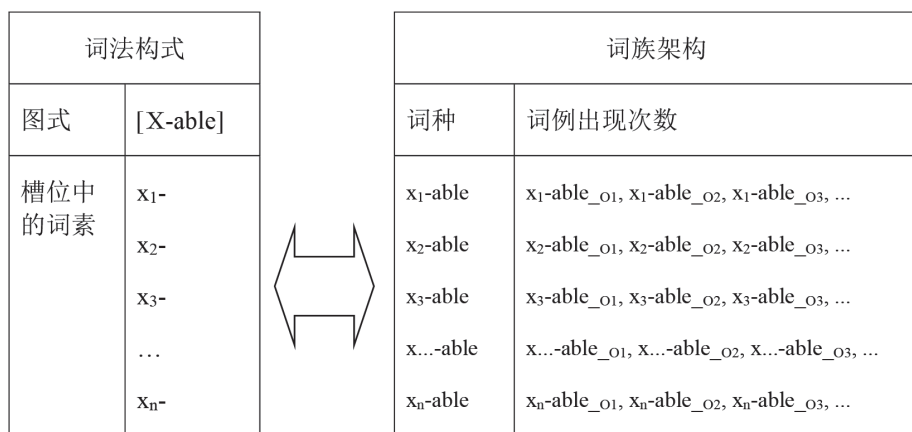


图 2. 词法图式与词族架构的对应关系(后缀“_O+n”表示第 n 次出现的词例)

在构式形态学的视角下,我们在词法能产性与词族的词汇丰富度之间建立起关联。具体而言,词法能产性被视为词法构式中槽位的词素丰富度,而后者又被操作化为词族的词汇丰富度。

3.3 动态视角的词法能产性

能产性是语言的基本属性,指创造新的语言单位的可能性,而且是有限手段的无限运用。能产性是一个抽象概念,通常被操作化为某些量化指标。

在构式形态学视角下,词法能产性是指对于某个形态过程 M,每出现一个词时,这个词是新造词的概率(Bauer 2001: 162)。R 语言的工具包 {zipfR} 提供了 16 种能产性算法(Evert & Baroni 2007),这里主要介绍 Baayen (2009) 总结的三种算法: 1) 现实能产性(realized productivity),即某个形态过程 M 中业已存在的词种数(V)¹与词例数(N)的比率(type-token ratio, TTR)。由于这种算法统计的是已经存在的词的数量,因此不能说明某形态过程 M 是否仍有产生新词的能力。2) 潜在能产性(potential productivity),即某个形态过程 M 中的一次词的词种数(V1)除以词例数,用于衡量某形态过程 M 的词汇增长率。这种算法的值对应于词汇增长曲线的斜率(Evert 2004: 130)。Baayen & Lieber (1991: 809) 认为这种算法最符合语言学家对于能产性的直觉。3) 扩展能产性(expanding productivity),即某个形态过程 M 中的一次词的词种数除以整个语料库中一次词的种数,用于衡量某形态过程 M 对语料库整体词汇增长率的贡献。类似的算法还有一次词/词种数比率(hapax/type ratio, HTR),即某个形态

1 由于词种数和词例数的英语术语首字母都是 T,为避免混淆,本文沿用文献中的符号,用 V 表示词种数,用 N 表示词例数, V1 表示一次词(hapax)的种数, V2 表示二次词的种数。

过程 M 中的一次词的词种数除以该词族的词种数,相对于潜在能产性,可以更直接地衡量某形态过程 M 的词汇增长率。Evert (2004: 130) 认为 HTR 的值对应于齐普夫模型的参数 a 。

计算能产性的基本指标是词例数 N 、词种数 V 和一次词 $V1$ 。过去的算法是对词族特征的宏观计量,如果深入观察词族的内部,就可发现各种词的词例数分布存在差异。词族内部的词频分布服从齐普夫分布,可以用齐普夫模型拟合。本文尝试将齐普夫模型应用于词族内部的词频分布,采用动态视角研究词汇丰富性。如果把文本切分成若干个片段,从第一个片段开始统计词汇量,依次累加,就可以追踪词汇量随着文本长度的增加而变化的过程,因而研究具有动态性。动态视角的能产性比静态视角的更精准。静态的能产性研究是在特定样本量中测量的“孤立”统计数据,而动态研究能更深入地了解某个构式槽位的能产性。在当前的能产性研究中,这种动态的观点还没有得到充分利用 (Zeldes 2012: 120)。

动态视角的词汇增幅率 (vocabulary growth) 用于动态衡量文本 (或构式、词族) 中词汇量 (包括词种数、一次词) 的增长情况,是能产性的主要指标。冯志伟 (2012) 提出了生词增幅递减律: 在书面文本阅读过程中,随着文本长度的增加,文本中生词的增幅呈现出递减的趋势。Fan (2006) 发现,不同长度的语料样本,其词汇增长模式符合不同的数学模型: 6 万词次以下的文本符合三次函数分布,6 万词次以上符合幂律分布。当语料达到三百万词例 (词次) 及以上时, HTR 趋于稳定值,接近 1 (Fan 2010)。

4. 研究方法

4.1 语料及数据基本形态

本研究借鉴 Hilpert (2015) 分享的英语复合词数据库,该数据库基于从 COHA 中提取的“名词-过去分词复合词”,共 31,659 条记录。我们将所有复合词按照第二成分 (过去分词) 归并成 3186 个词族,以 N -ppl 命名,如 N -based、 N -shaped 等。每个词族的统计特征包括词种数 V 、词例数 N 以及两者的比例 TTR。一些规模较小的词族 (例如词族 N -conceited 和 N -paged 分别仅包含 self-conceited 和 front-paged 一个词种) 因不满足建模条件而被剔除。

我们使用 R 语言工具包 {zipfR} 统计所有词种的频序 (rank)、频次 (frequency) 等信息,并制成词频表 (type frequency list)。为了获取“ $N_1 + N_2$ ”复合词的语料,我们将 N_2 整理成自定义搜索词表,用于批量检索 COHA 中的相关复合词,检索结果通过 R 语言的 {zipfR} 包转换为“频数分布表” (frequency spectrum),用于后续模型拟合和能产性计算。

我们认为,词族的词频分布可能符合某种概率分布规律(如齐普夫分布)。通过模型拟合,可以测算词族的能产性等特征,并且可以大规模地进行词族之间的比较。

4.2 建构词族的词频分布齐普夫模型

齐普夫定律(Zipf's law)用于描写自然语言中单词的出现频次与其频率之间的关系。根据该定律,在一个较大的文本中,单词的出现频次(f)与其频率(r)成反比关系,即频率越大的单词出现的频次越少,频率越小的单词出现的频次越多。该定律在很多语言中已得到验证,包括英语、法语、汉语等。齐普夫定律经历了几次修正,从单参数模型到双参数模型和三参数模型,也称作齐普夫-曼德博(Zipf-Mandelbrot)模型(冯志伟 1983)。

为了对每个词族的能产性进行拟合,我们运用 R 语言 {zipfR} 包中的 lnre() 函数建构每个词族的齐普夫模型,具体步骤如下。1) 筛选语料:根据 Bauer (2001: 151) 提出的样本标准,保留词种数为 30 及以上的词族,剔除规模过小的词族,最终保留 191 个词族。2) 建构模型:基于 lnre() 函数,并结合自建函数 modeler2(), 根据每个词族的词频分布表建构齐普夫模型,同时生成模型的参数、拟合度等信息。模型评估标准参考 Baayen (2008:233):卡方值(χ^2)越小越优,且 $p > 0.05$ 。部分结果见表 1。

表 1. 部分词族的模型参数及拟合指数

词族	alpha	B	A	χ^2	df	p
N-based	0.71	0.04	1.52E-05	10.58	9.00	0.31
N-shaped	0.62	0.06	2.8E-06	8.67	7.00	0.28
N-colored	0.56	0.07	5.9E-06	10.09	7.00	0.18

为了更直观地检测每个词族的齐普夫模型拟合的效果,我们利用 {zipfR} 包中的 lnre.spc() 函数推测这些词族的词种数期望值,并与观测值进行比较。由于词频分布情况不同,每个词族的模型拟合情况也不同。总体而言,大部分词族的拟合度 p 值大于 0.05,模型效果较为理想,但也有少数例外,需要调试模型参数,重新拟合。

如前文所述,词法能产性体现为多种指标,其衡量依赖于词例数(N),而不能直接比较词种数(V),也不能比较两者的比值 TTR。即使两个词族的词种数(V)和词例数(N)相同,也不意味着它们的能产性相同。通过模拟词族的增长曲线,可以更准确地推断词法能产性。我们采用 lnre.vgc() 函数,在词族模型的基础上模拟出词族的动态增长过程:以 50 为词例数增量,计算相应词种

数的期望值(含一次词)。设定总词例数尺度范围为 0—2500,参照 Baayen(同上: 235)的方法,根据词族的词例数观测值,对模型推测方法进行区分,内推法(interpolation)是回顾性质的,对照已有数据进行拟合;外推法(extrapolation)是展望性质的,没有现成的真实数据对照。若某个词族的词例数观测值小于 2500 则采用外推法进行分析,反之则采用内推法。由于本研究中大部分词族的观测词例数小于 2500,所以主要采用外推法进行分析。

5. 结果:内嵌能产性与词法构式的确立

本文讨论 $[N_1 + N_2\text{-ed}]$ 构式。前期研究(Hilpert 2015; Zhan, Huang & Sun 2023)根据语料和理论文献论证了这个构式的特殊性,但还只是假拟的构式,仍需提供该构式确立的依据。

- (1) $\langle [N_1\text{-}N_2\text{-ed}]_A \leftrightarrow [\text{having SEM}_2 \text{ which has relation R with SEM}_1] \rangle$
(Zhan, Huang & Sun 2023: 565)

如(1)所示,该构式内嵌了复合词“ $N_1 + N_2$ ”,通过比较“ $N_1 + N_2$ ”在构式内部和外部的能产性是否有差异,就可为该构式的确立提供依据。一般而言,名-名复合词“ $N_1 + N_2$ ”表达的意义较为复杂,Jackendoff(2016)归纳了 13 种语义关系,如部分整体关系(cigarette butt)、原料关系(apple juice)、象似关系(piggy bank),等等。从中观图式来看,某些词族具有独特的语用意义(Mattiello 2022)。我们假设,通过选定某个 N_2 为词族的中心词,分别收集“ $N_1 + N_2$ ”在构式内和构式外的能产性,如果两者有差异,即可说明该构式具有特殊性。

为了计算词族的能产性,本研究以 N_2 为词族的中心词,计算每个词族中 N_1 这个槽位的词汇丰富度。根据 TTR 公式,当各词族的词例尺度保持一致时,其词种数(期待值)就可以作为衡量词族能产性的指标。这样也就方便对构式内部与外部的能产性进行比较,只需统计某个词例尺度上这些词族的词种数,并计算其均值即可完成比较。表 2 显示,当词例尺度为 1000 时,从能产性指标(V、V1 和 HTR)看来(参见第 3.3 节的论述),“ $N_1 + N_2$ ”在构式外的能产性高于构式内。这一结果表明,在 $[N_1 + N_2\text{-ed}]$ 构式中,“ $N_1 + N_2$ ”复合词的能产性有别于在构式外的环境。

表 2. “ $N_1 + N_2$ ”复合词在构式内、外的能产性指标均值

“ $N_1 + N_2$ ”的构式环境	V	V1	HTR
$[N_1 + N_2\text{-ed}]$ 构式内	270	185	0.667
$[N_1 + N_2\text{-ed}]$ 构式外	363	258	0.714

构式是特定的局部环境,对内部成分的能产性形成压制或制约。以“N + size”复合词为例,有些在构式内和构式外是通用的,如 life size 或 life-sized,有些更倾向于构式内,如 pocket-sized、bite-sized、king-sized;还有些更倾向于构式外,如 sample size、class size、effect size。这意味着 size 被嵌入 $[N_1 + N_2\text{-ed}]$ 构式后,其语义和用法发生了某种程度的变化。

根据所建构的模型,我们能够推断出每个词族的词种数的动态增长过程(如图3中实线所示)。词种数是比较直观的能产性指标,新增词(V1)则是更为关键的核心指标,因为它直接反映了词法模式生成新词的能力。新增词增幅率有两种算法,拓展能产性和 HTR 都是以一次词 V1 为计算基础。在特定的词例尺度上,一次词 V1 的词种数能够体现某个词族的能产性(如图3中虚线所示)。

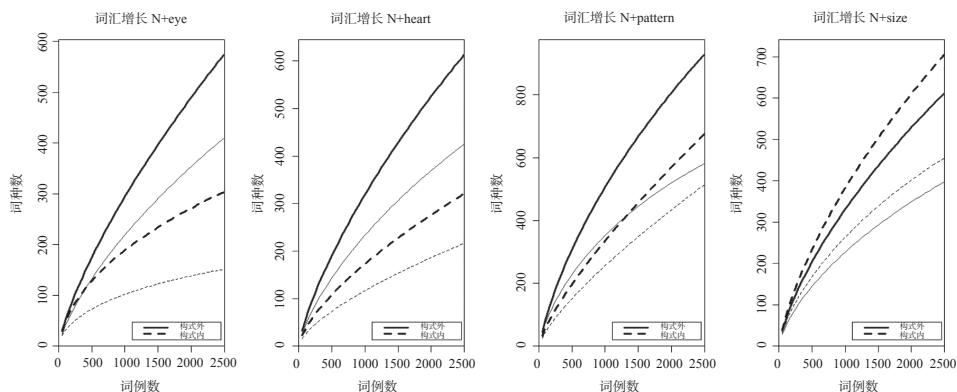


图3. 构式内外的词种数对比

从数据看,新增词的增长与词种数的增长呈现同步性,这充分验证了模型的稳定性和一致性。随着词例数的不断增加,新增词的词种数也在增加。图中线条的倾斜角度直观地反映了增长率的差异:构式外的词种数高于构式内(“N + size”除外)。需要指出的是,这里的数据是词族大类的整体平均值,对于个体的词族,仍存在一定的差异。因为 HTR 能够反映新增词 V1 对词族的词种总数的贡献程度,所以它更能精准地体现词族的能产性。图4展示了构式内外新词的增长曲线,呈现了每个词族的新词增长趋势。在图3中,前三个词族的增长曲线非常相似,但在图4中三者的区别就很明显。这表明在分析词族能产性时,要综合考虑构式的总体和个体特征。根据复合词内部的论元关系(Zhan, Huang & Sun 2023),可以解释词族能产性的差异。比如在 N-mounted 图式中,名词槽位的论元角色比较灵活,可以是方位(如 a ceiling-mounted fan)、材料/工具(如 gold-mounted spectacles)等;而 N-wrapped 图式的名词槽位论元角色比较单一,仅限于包装材料。

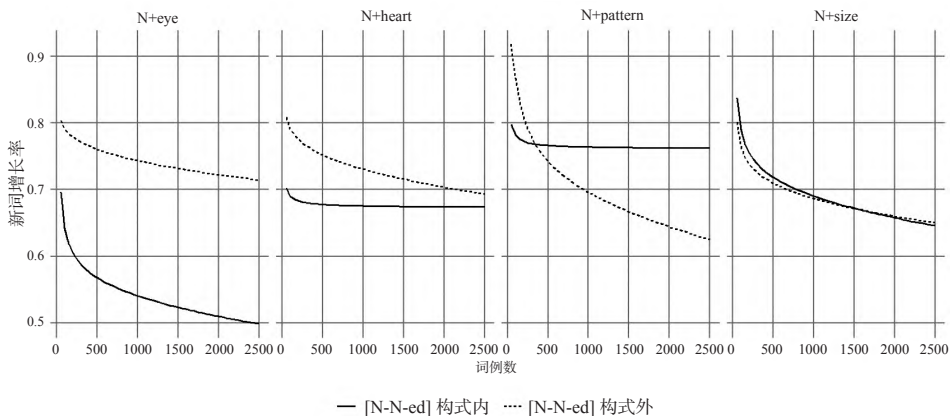


图 4. 构式内外的新词增长率对比

不同词族展现出各自独特的新词增长模式,且构式内外的增长模式存在差异。总体而言,新词增长有两种模式:增长率保持平稳或增长率递减。其中,增长率平稳的模式反映出较高的能产性。以“N + pattern”为例,尽管开始时构式外的增长率高于构式内,但随着词例数的增加,构式外的增长率急剧下降,最终低于构式内。再如,“N + heart”在构式内的增长率呈现平稳模式,而在构式外则为递减模式。据此可以合理推测,随着词例数的进一步增加,构式外的增长率将逐渐低于构式内。

内嵌能产性表明构式作为一个相对独立的环境,其内部与外部存在显著差异,这种差异可能引发词义的变化,或者语义漂移(semantic drift)。再以“N + pattern”为例,它在 $[N_1 + N_2\text{-ed}]$ 构式内、外的意思有共同之处。pattern 一词有多种意义,在名-名复合词中都有使用,可以表示模式(如 behavior pattern、thought pattern),也可以表示图案(如 star pattern、crisscross pattern)。当 pattern 表示图案时,还可能出现歧义,因为 pattern 是象征性名词(symbolic noun),具有象征性名词的歧义特征(Taylor 2002)。如 wave pattern 有两种解读:一种是“波浪所呈现的图案”,如例(2);另一种是“类似波浪的样式”,如例(3)。

(2) That Ocean Wave pattern's the purtiest I've ever seen. (COHA, 1955, FIC)

(3) She chose hair with a larger wave pattern, which a... (COHA, 2002, MAG)

然而,在构式内部, $[X\text{-pattern-ed}]$ 的语义相对固定,只能表示“X 是图案呈现的内容”,如 star-patterned。这种用法的能产性很高,说明图案所呈现的内容比较宽泛,可以自由创造新词。这种意义变化的产生可能受到多种因素的影响:一是上下文语境,它为词义的生成提供了具体的语用背景;二是 $[X\text{-pattern-ed}]$ 中 X 槽位上词汇的词义聚合效应,不同词汇进入该槽位时可能引发语义的重新

整合；三是多种因素共同作用。由于在实际语言使用中很难区分也不必区分各个成分的贡献，所以构式语法更倾向于从整体上解读构式的意义。

6. 结论

本文在构式形态学的框架下，论述了内嵌能产性作为词法构式确立的依据。以英语中假拟的词法构式 $[N_1 + N_2\text{-ed}]$ 为例，我们从语料库中提取了每个词族的词频分布表，并运用 R 语言的专业工具包建构了齐普夫分布模型。通过外推法测算“ $N_1 + N_2$ ”的能产性指标，我们发现“ $N_1 + N_2$ ”在 $[N_1 + N_2\text{-ed}]$ 构式内部与构式外部的能产性有显著差异。这一发现从计量角度为确立 $[N_1 + N_2\text{-ed}]$ 词法构式提供了数据支撑。词频分布作为构式性特征之一，不仅具有一定的预测能力，也是衡量语言使用特征的重要维度之一。本研究延续了过去相关研究（如 Hilpert 2015；Zhan, Huang & Sun 2023）的思路，还对其进行了补充和拓展。

本文论证了词法能产性与词族的词汇丰富度之间的关联。我们把词法能产性视为词法构式中槽位的词素丰富度，并将其操作化为词族的词汇丰富度。本文采用动态视角论述了词汇增幅率，即文本（或构式、词族）中词汇量（包括词种数 V 、一次词 V_1 、TTR、HTR）的增长情况，是能产性的主要指标。本文提出的词频拟合算法有效解决了原始语料文本长度差异所带来的问题，通过将不同的词例数内推或外推至同一尺度，实现了对词频数据的标准化处理。这一算法可用于语体分析、作者个人风格、作者识别等研究，也可用于外语教学中的“词频概貌”研究，借助该算法可以精准分析教材、阅读文本、产出文本中的词汇丰富度以及词汇覆盖度，从而为教学内容的优化、学习者语言能力的评估以及教学方法的改进提供科学依据。

参考文献

- Baayen, R. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* [M]. Cambridge: CUP.
- Baayen, R. 2009. Corpus linguistics in morphology: Morphological productivity [A]. In A. Lüdeling & M. Kytö (eds.). *Corpus Linguistics: An International Handbook* (Vol. 2) [C]. Berlin: Mouton de Gruyter. 899–919.
- Baayen, H. & R. Lieber. 1991. Productivity and English derivation: A corpus-based study [J]. *Linguistics* 29(5): 801–844.
- Bauer, L. 2001. *Morphological Productivity* [M]. Cambridge: CUP.
- Bencini, G. & A. Goldberg. 2000. The contribution of argument structure constructions to sentence meaning [J]. *Journal of Memory and Language* 43(4): 640–651.
- Booij, G. 2010. *Construction Morphology* [M]. Oxford: OUP.
- Booij, G. 2015. Word-formation in construction grammar [A]. In P. Müller *et al.* (eds.). *Word-Formation* (Vol.1) [C]. Berlin: De Gruyter Mouton. 188–202.

- Booij, G. & J. Audring. 2017. Construction morphology and the parallel architecture of grammar [J]. *Cognitive Science* 41(S2): 277–302.
- Du, Fengjiao [杜凤娇]. 2024. The application and the value of Construction Morphology in the study of Chinese lexical research [J]. *Language Teaching and Linguistic Studies* (5): 90–100. [构式词法在汉语词汇研究中的应用及价值探析, 《语言教学与研究》5]
- Evert, S. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations [D]. Ph.D. Dissertation. University of Stuttgart.
- Evert, S. & M. Baroni. 2007. *zipfR*: Word frequency distributions in R [A]. In S. Ananiadou (ed.). *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* [C]. Prague: Association for Computational Linguistics. 29–32.
- Fan, Fengxiang. 2006. A corpus-based empirical study on inter-textual vocabulary growth [J]. *Journal of Quantitative Linguistics* 13(1): 111–127.
- Fan, Fengxiang. 2010. Squibs: An asymptotic model for the English hapax/vocabulary ratio [J]. *Computational Linguistics* 36(4): 631–637.
- Feng, Zhiwei [冯志伟]. 1983. A historiographic study on Zipf's law [J]. *Information Science* (2): 37–42. [齐普夫定律的来龙去脉, 《情报科学》2]
- Feng, Zhiwei [冯志伟]. 2012. Studying language by quantitative methods [J]. *Foreign Language Teaching and Research* (2): 256–269. [用计量方法研究语言, 《外语教学与研究》2]
- Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure* [M]. Chicago: The University of Chicago Press.
- Hampe, B. 2011. Discovering constructions by means of collocation analysis: The English denominative construction [J]. *Cognitive Linguistics* 22(2): 211–245.
- Hartmann, S. 2014. Constructing a schema: Word-class changing morphology in a usage-based perspective [J]. *Yearbook of the German Cognitive Linguistics Association* 2(1): 235–252.
- Hilpert, M. 2015. From *hand-carved* to *computer-based*: Noun-participle compounding and the upward strengthening hypothesis [J]. *Cognitive Linguistics* 26(1): 113–147.
- Jackendoff, R. 2016. English noun-noun compounds in Conceptual Semantics [A]. In P. ten Hacken (ed.). *The Semantics of Compounding* [C]. Cambridge: CUP. 15–37.
- Lakoff, G. 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind* [M]. Chicago: The University of Chicago Press.
- Langacker, R. 1987. *Foundations of Cognitive Grammar* (Vol. 1): *Theoretical Prerequisites* [M]. Stanford, CA: Stanford University Press.
- Liu, Geng & Wenbin Wang [刘庚、王文斌]. 2021. The temporality of English and the spatiality of Chinese in the etymology of word/character formation [J]. *Foreign Language Research* (1): 108–115. [从构词词源看英汉时空性差异, 《外语学刊》1]
- Masini, F., S. Mattioli & S. Pepper. 2023. Exploring complex lexemes cross-linguistically [A]. In S. Pepper, F. Masini & S. Mattioli (eds.). *Binominal Lexemes in Cross-linguistic Perspective* [C]. Berlin: De Gruyter Mouton. 1–20.
- Mattiello, E. 2022. Slang and verbal aggression: A morphopragmatic analysis of the compound families X-ass, X-brain, X-face, and X-head [J]. *Journal of Language Aggression and Conflict* 11(1): 101–120.
- Schönefeld, D. 2013. *It is ... quite common for theoretical predictions to go untested* (BNC_CMH).

- A register-specific analysis of the English *go un-V-en* construction [J]. *Journal of Pragmatics* 52: 17–33.
- Si, Weiguo & Xu Wen [司卫国、文旭]. 2024. The motivations of the modern Chinese verb-copying construction [J]. *Journal of Foreign Languages* (3): 53–62. [现代汉语动词拷贝构式形成的理据考辨, 《外国语》3]
- Song, Zuoyan [宋作艳]. 2023. A constructionist account of N_1VN_2 compounds in Mandarin Chinese [J]. *Studies of the Chinese Language* (5): 531–547. [汉语 N_1VN_2 型复合词的构式解读, 《中国语文》5]
- Stefanowitsch, A. & S. Gries. 2003. Collocations: Investigating the interaction of words and constructions [J]. *International Journal of Corpus Linguistics* 8(2): 209–243.
- Taylor, J. 2002. *Cognitive Grammar* [M]. Oxford: OUP.
- Wierzbicka, A. 1988. *The Semantics of Grammar* [M]. Amsterdam: John Benjamins.
- Zeldes, A. 2012. *Productivity in Argument Selection: From Morphology to Syntax* [M]. Berlin: De Gruyter Mouton.
- Zhan, Hongwei, Sihong Huang & Lei Sun. 2023. Paradigms as second-order schemas in English noun-participle compounding [J]. *Review of Cognitive Linguistics* 21(2): 542–575.

Quantitative Approach to the Embedded Productivity of English Morphological Constructions: Model Fitting Based on Zipf's Law

ZHAN Hongwei

(School of International Studies, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: Establishing a proposed construction requires supporting evidence, which in previous studies has typically derived from introspective semantic judgments or corpus-based quantitative analysis. The rationale is that a valid construction demonstrates distinctive syntactic and semantic idiosyncrasies. This article explores how to use embedded productivity as a criterion for establishing morphological constructions by analyzing and comparing the productivity of the word formation processes within and outside a posited construction — where significant differences indicate the establishment of the construction. Taking the English morphological construction $[N_1 + N_2\text{-ed}]$ as an example, this article extracts the word frequency distribution of each word family from the Corpus of Historical American English (COHA), constructs a Zipf distribution model using the {zipfR} package in the R language, and calculates the productivity of “ $N_1 + N_2$ ” using the extrapolation method. Through comparison, it is found that the productivity of “ $N_1 + N_2$ ” is significantly different within and outside the $[N_1 + N_2\text{-ed}]$ construction, thus confirming the establishment of the $[N_1 + N_2\text{-ed}]$ morphological construction. The methods and conclusions of this study have implications for studies on morphological productivity and construction morphology.

Keywords: construction morphology, morphological construction, embedded productivity, Zipf's law, compound words

收稿日期: 2024-10-28; 修改稿 2024-12-25; 本刊修订 2025-03-19

通讯地址: 311121 浙江省杭州市 杭州师范大学外国语学院

电子邮箱: hwzhan78@163.com