

流行语特征与流行模式^{*}

——基于大型网络流行语数据的相关性分析及其在辞书编纂中的应用

彭馨葭 吴婷婷 严文静

摘 要 随着互联网的发展,网络空间涌现了大量的新词新语,成为了现代汉语新词汇的重要来源。然而,不是每一个流行语都能经得住时间的考验,长期留存在汉语词汇当中,具有被辞书收录的价值。文章探寻不同类型的网络流行语是否与不同的流行和使用模式存在相关性。通过对2010—2019年的300余个网络流行语在网络平台上的历时使用变化进行分析,我们发现,网络流行语的某些特征可能对其使用频次、存续时间有影响。文章对辞书收录新词有一定的参考意义。

关键词 网络流行语 流行模式 新词收录

DOI:10.16134/j.cnki.cn31-1997/g2.2025.03.009

一、引 言

随着互联网空间成为中国国民生活的一个重要场域,网络语言生活成为中国民众语言生活的重要空间,网络空间自然而然地成为了现代汉语新词萌生、传播的一个重要场域。关注网络流行语言的使用对于把握现代汉语的发展具有积极意义。

《现代汉语词典》(以下简称《现汉》)从第3版起,每一版的修订都积极地收录一些新词汇,一方面反映了汉语词汇的新发展,同时也有助于促进语言使用的规范化。《现汉》第6版收录的相当一部分新词便来源于网络。除了备受语言学家关注的“被”(王灿龙 2009;沈家煊 2010)、和互联网密切相关的词汇如“博客、团购”、反映社会现象和流行文化的“闪婚、潜规则、达人、月光族”等词外,许多新收录词都是经由互联网助推、形成一定的使用规模而被考虑收录的。新词的收录与否须极为审慎。《现汉》第5版、第6版所参考的标准,除了学理上的合理性,最重要的两项是:是否具有普遍性(广泛使用),是否具有生命力(一定时间内稳定使用)。(韩敬体 2006;江蓝生 2012)使用的普遍性可以用词语在一段时间内的使用频率佐证,而是否具有生命力涉及对未来情况的预测,根据以往和当下的使用数据无法直接回答。

既往的研究显示,大部分流行语稍纵即逝。(张志毅,张庆云 1997)王炎龙、夏赛

^{*} 本研究受国家语言文字工作委员会科研项目“基于发展模式与特征的网络语言分类标准研究”(项目编号 YB136-166)及中国社会科学院语言学重点实验室(项目编号 2024SYZH001)资助。

楠(2017)计算出网络流行语持续流行的平均时间为 16.1 个月。但流行语之间的流行模式存在很大差异——有些迅速蹿红后逐渐淡出了人们的视野,有些则平稳增长并长期流行。前人研究已表明,历史上语言形式的盛衰与更替和语言本体、社会文化、认知等多方面的因素有关。(Boulanger 2002; Metcalf 2004; Jing-Schmidt, Hsieh 2019)具体到网络流行语,有研究观察到,流行语不同的走红模式、持续流行模式与词语的内容指向和来源有一定的关系。比如,王昊(2014)留意到“突发型”网络流行语通常与负面事件相关;季乃礼等(2021)发现此类流行语“大多聚焦于重大社会事件、热点新闻、公众心理及具备现实意义的相关影视作品”,而发展相对平稳的流行语大多“源自公众人物或娱乐作品”并“呈现出自主群体划分的倾向或反思现实问题的生活态度”。

延续前人的研究思路,本研究假设网络流行语的一些自身特征可能影响其流行模式。有多大的使用量,有多长的持续流行时间,是辞书在收录新词时重点考虑的使用普遍性和生命力两个维度。以上提及的研究并未对此进行系统性考察,多是基于概率的直觉观察,并且涉及的特征也较为有限。本研究旨在全方位地考察流行语自身特征与其流行模式之间的关联,希望能为辞书收录流行语类新词提供一定的参考。

二、研究设计

(一) 研究目标

本研究旨在通过量化手段系统分析流行语自身特征与其流行模式之间的关联。具体而言,本研究将通过相关性统计分析确定这两类变量是否存在关联。一旦确定关联存在,便可进一步利用回归模型进行建模。在进行统计分析之前,需要先确定进行分析的具体流行语的范围,而后准备两方面的数据:一是其流行模式的量化呈现,二是流行语自身特征的确定和标记。以下将逐一介绍。

(二) 研究材料

本研究选取 2010—2019 年的网络流行语及其在新浪微博^[1]的流行和使用情况为例进行研究。

在网络流行语的选择上,我们以穷尽式的思路,集合了几个官方和非官方的 2010—2019 年每个年度网络流行语的评选榜单,^[2]整理后确定 324 条流行语作为考察对象。^[3]

在语料方面,我们自建了一个十年微博语料库作为语料来源。语料库的抽样方法参考 Hino 和 Fahey(2019),着重考虑容纳大样本用户和语料延续性之间的平衡,最终抽样了共计 12 万用户人次的语料,形成了库容为 440 万余条的微博语料库。本研究中,我们以月份为单位进行变化趋势的描写。可以发现,用户月平均发布 36 条微博,活跃度良好;语料字数月均达 150 万字,也就是说,每个月份的语料规模基本上都能达到百万字的级别,因此可以把网络流行语的使用情况与汉语常用字汇的使用情况做一个横向的对比。

对于被调查的 324 条流行语,我们逐一从自建微博语料库中提取语料,形成包含发

布时间标签的索引行,并以月为单位绘制了反映使用情况历时变化的图表——下文称“使用走势图”。

(三) 变量设置与初步发现

1. 描述使用情况的变量: 两个热度变量

本研究关注流行语使用的普遍性和生命力。如果采用量化手段描述,这两个方面体现为:横向的,即时间维度的走势;纵向的,即单月使用量的走势。本研究将这两个维度分别称为“持续热度”和“顶峰热度”。

持续热度:网络流行语持续流行的时长(以年计);

顶峰热度:根据网络流行语最流行月份的词频转换的热度等级。

(1) 持续热度

描述网络流行语使用的时间维度的变量是“持续热度”,即持续的、具有“热度”的使用的时长。为了保证判定的准确性,我们采取人工方式进行判定,即观察词条的使用走势图,以流行语的流行年度为准线,检视两侧曲线在X轴上持续变化所覆盖的年份;当曲线接近X轴,即词频在年度范围内基本处于0的水平且无变化时,视其在该年度不流行。图1是“厉害了我的×”的使用走势图,我们视其流行年份为2016年中后期至2019年末,化整后计为4年。以这种方法进行统计,本研究样本内的网络流行语的持续热度平均为4.38年,远高于此前流传的“生命周期只有几个月”的说法和“16.1个月”的研究发现。

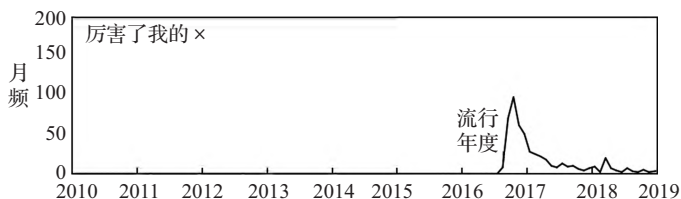


图1 流行语“厉害了我的×”在2010—2019年的使用走势图

网络流行语的“持续热度”整体呈正态分布(详见图2),但也有为数不少的一部分流行语在十年之内持续流行。表1列出了不同持续热度的流行语的例子,其中持续流行十年的流行语有不同的情况。比如,“吐槽”是2011年的网络流行语,2010年左右从日语舶来之后,作为一个形式新颖、表义特殊的新词汇,融入了汉语,成为一个日常用语;“呵呵”大概是随着网络聊天的诞生而存在的一个表达,但是在2017年才因为偶然被评为流行语,其后也持续流行。当然必须指出,因为本研究使用的是抽样数据,这个结果也有一定的偶然性。

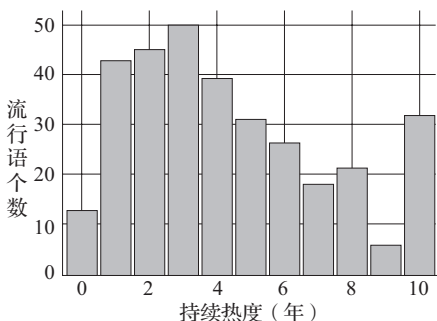


图2 不同持续热度的流行语个数统计

表 1 不同持续热度的流行语

持续热度(年)	例子
10	××style、呵呵、吐槽、仙女、鸡汤、朋克
8	奇葩、剁手、躺枪、中国好××、怎么破、高富帅、给跪了
3	吸猫、猴赛雷、良心不会痛吗、累觉不爱、友谊的小船
0	霸凌主义、不知知网、说闹觉余、这届 90 后

(2) 顶峰热度

描述网络流行语使用频率维度的变量是“顶峰热度”,即指以月份为单位进行使用频次统计中最高的频次。我们对每个词的月份词频进行标准化处理,^[4]以相对词频进行统计。我们参考了汉语词汇词频,根据词频排名的范围划定了七个词频等级(详见表 2),^[5]将定量数据转换为定序数据,方便后面的相关性分析。

表 2 顶峰热度两种数值的转换

顶峰热度(等级)	顶峰热度(词频)	参考汉语词频排名	区间词汇
7	>250	前 500	以及、东西、存在
6	130—250	500—1000	身上、首先、无法
5	63—129	1000—2000	女人、唯一、错
4	40—62	2000—3000	生气、天空、二百
3	18—39	3000—6000	牙齿、表面上、全面发展
2	9—17	6000—10000	活儿、常年、肃清
1	1—8	10000 以后	成交、撰写、公约

分析结果表明,流行语整体的顶峰热度呈现冥律分布(详见图 3):顶峰热度最高的网络流行语极少——顶峰热度相当于汉语词频前 1000 名的网络流行语只有 10 个,只占整体网络流行语的 3%;大部分网络流行语的使用频率接近于汉语专业领域的词汇——词频排名在 1 万往后的流行语占了整体的 40% 以上。顶峰热度的平均值为 1.94,接近于 2 的热度,即词频排名 1 万左右。

我们也注意到,顶峰热度和持续热度存在正相关的关系——顶峰热度越高,持续流行时间也越长(详见图 4)。

2. 描述流行语特征的变量:五组特征变量

此前对网络流行语的研究着重根据形式特征对网络流行语进行分类。“谐音”“缩写”“四字格”等特征由于其“不规范性”而备受关注,本研究也将特别考察具有这些形式特征的网络流行语是否在使用热度上具有一定的共性。除此以外,我们也假设网络流行语的功能特征(类似于名词、动词、形容词的指称功能,这里我们指的是网络流行语在

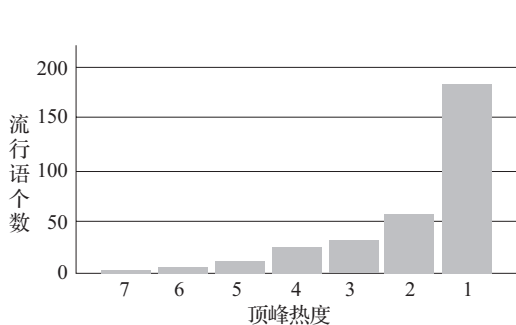


图3 不同顶峰热度的流行语个数统计

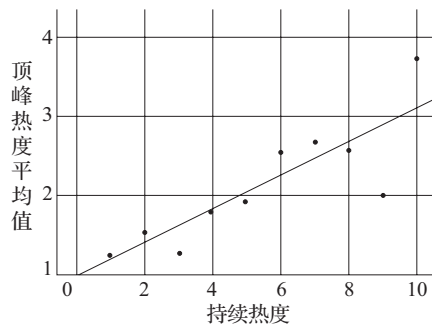


图4 顶峰热度和持续热度的相关性

互联网会话中的某种语用功能)、流行语的来源和一些其他的明显特征可能影响其使用热度。受到罗曼·雅柯布森(2012)语言六功能(指称功能、表情功能、呼吁功能、寒暄功能、元语功能、诗歌功能)的启发,结合流行语实际的一些突出特征,我们形成了以下五大类共23个特征。特征的具体内涵我们会在第三节详述。我们对324个网络流行语是否具有这些特征一一进行了标注,标注时遵循了大类之间不互斥的原则,即,允许一个网络流行语同时具有多个大类特征。表3显示的是该研究中具有这些不同特征的网络流行语占比(不具备这些特征的未予统计)。图5显示的是网络流行语长度的分布情况。

表3 具有不同形式和功能特征的网络流行语占比

指称功能		交际功能		语言特征		来源		其他	
特征	21%	抒发情感	27.5%	新短语	13.65%	娱乐节目	34.9%	旧词新义	13.9%
行为	16%	发表评价	18.2%	谐音词	11.1%	社会时事	28.4%	框式结构	9.6%
具体范畴	14.8%	开启对话	16.4%	缩略词	9.6%	官方媒体	6.2%	粗鄙字眼	4.6%
抽象范畴	14.8%	阐发哲理	7.4%	字母词	6.5%			长度	
				外来词	5.6%				
				新词	5.2%				
				四字格	2.8%				
				方言	1.9%				

结果显示,指称功能和交际功能这两个大类之间具有一定的互斥性(图6-a),并且在统计学上具有显著意义。指称功能和语言特征这两个大类之间具有统计学意义上的显著吸引力(图6-b),也就是说,具有指称功能的网络流行语更多地具有可辨识的、鲜明的语言特征。交际功能和语言特征之间具有一定的互斥性(图6-c),但不是那么明显,这也可以理解为,具有交际功能的网络流行语更倾向于使用语言社群使用者相对不那么陌生的语言符号。

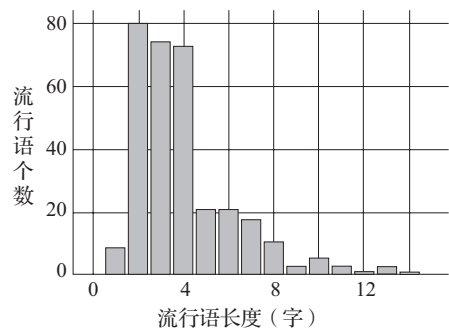


图5 网络流行语平均长度

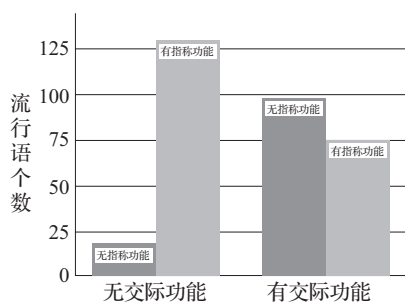


图 6-a 指称功能和交际功能的共现情况。 $\chi^2(1, n=324)=69.04$,
 $p=0.00<0.001$

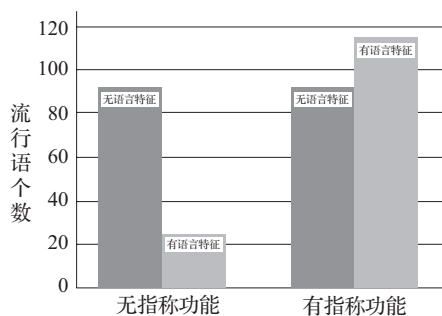


图 6-b 指称功能和语言特征的共现情况。 $\chi^2(1, n=324)=35.6$,
 $p=0.00<0.001$

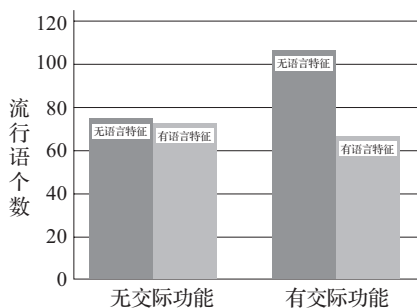


图 6-c 交际功能和语言特征的共现情况。 $\chi^2(1, n=324)=3.76$,
 $p=0.052>0.05$

三、研究结果

本研究假设,网络流行语的热度可能受自身形式和功能特征的影响。基于此,我们先考察热度变量与特征变量的相关性。在确认存在相关性之后,再尝试进行建模。

(一) 相关性分析

我们把持续热度、顶峰热度作为因变量,把五组特征变量作为自变量,检验这些特征变量作为独立的自变量对于热度是否有影响。特征变量中的大部分是定性变量,对这一部分我们使用了单因素方差分析(one-way ANOVA);^[6]特征变量中只有“长度”是定量变量,我们使用了斯皮尔曼相关性分析(Spearman correlation analysis)。^[7]分析结果详见表4。我们把特征变量对于趋势变量有显著差异的情况着重标出,其中,颜色深浅代表差异的显著性程度,无边框的着色单元格代表因素间的正向相关关系、虚线边框的着色单元格代表反向相关关系。从辞书新词收录的要求来看,应该重点关注的是与顶峰热度和持续热度都具有正相关关系(即无边框着色单元格)的特征类别。

表 4 网络流行语中特征变量对热度变量的差异效应规模

自变量	因变量		自变量	因变量	
	顶峰热度	持续热度		顶峰热度	持续热度
指称功能			语言特征		
特征	0.001	0.064	新短语	0.868	0.389
行为	0.271	0.072	谐音词	0.155	0.151
具体范畴	0.043	0.008	缩略词	0.263	0.618
抽象范畴	0.718	0.260	字母词	0.045	0.751
交际功能			外来词	0.003	0.035
抒发情感	0.638	0.017	新词	0.000	0.000
发表评价	0.086	0.969	四字格	0.050	0.193
开启对话	0.017	0.001	方言	0.429	0.546
阐发哲理	0.007	0.007			
来源					
娱乐节目	0.349	0.381			
社会时事	0.816	0.110			
官方媒体	0.637	0.103			
其他					
粗鄙字眼	0.876	0.931			
旧词新义	0.008	0.023*			
框式结构	0.285	0.260			
长度	0.000	0.000			

注：* 表示该值与“长度”有交互作用，未予置信。

表 4 显示,在我们所考察的 23 个特征中,有 10 个特征表现出对于热度的差异,也就是说,的确存在一些与网络流行语使用热度相关的关键因变量。另外,各个特征变量对热度变量的相关性方向趋同,这可归因于两个热度变量本身存在正向相关性(见图 4)。本文中,我们要特别留意的是与高持续热度和高顶峰热度有关的变量。

1. 流行语长度

“长度”作为所有特征变量中唯一的一个定量变量,对于流行语的使用趋势有极为显著的影响,表现为:流行语长度越短,持续流行时间越长,顶峰热度越高(见图 7)。齐普

夫定律 (Zipf's Law) 能很好地解释长度与两个热度变量的关联: 词汇长度越长, 其使用频率也越低, 而低频的使用也直接导致了使用上难以持久。

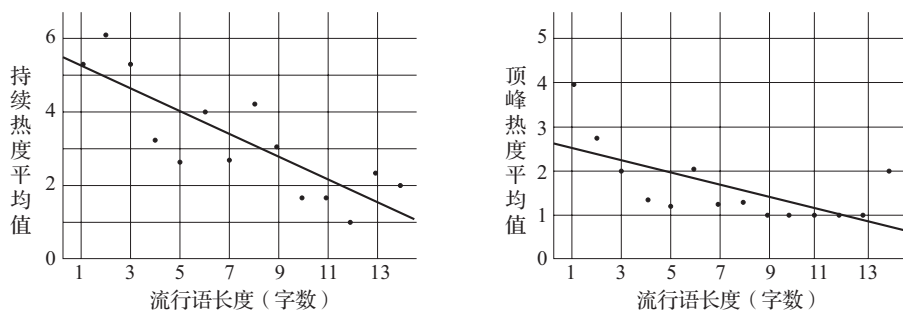


图 7 流行语长度与持续热度 (左)、顶峰热度 (右) 的相关性

2. 指称功能

指称功能具有四个小类。具体范畴指称如“杠精”, 指称某个人群; 抽象范畴指称如“996”, 指称某种社会现象; 行为指称如“吐槽”; 特征指称如“丧”。四个小类中, 特征指称对于顶峰热度具有极为显著的效应, 而具体范畴指称对于持续热度有极为显著的效应。

观察持续热度高的具体范畴指称类流行语, 相比其他指称类流行语, 它们更具有恒常性, 是社会中长期存在的现象。因此, 一旦当下的经历触发了相关概念, 通过使用该流行语, 使用者能够与网络社区的其他用户形成一种共鸣、呼应。在我们的样本中, 持续热度较高的具体范畴指称类流行语大都是人群标签, 比如“达人”“二代”“直男”“屌丝”“女汉子”等。

再观察顶峰热度高的特征指称类网络流行语, 比如“给力”(顶峰热度 7/7)、“萌萌哒”(顶峰热度 6/7)、“最美 × ×”(顶峰热度 5/7)、“奇葩”(顶峰热度 4/7), 我们发现, 它们的内涵都非常宽泛, 宽泛到几乎可以用于一切事物和现象, 其语义很难用某一个具体属性来概括, 但在语用上都是突出对谈论对象的一种肯定或否定。网络流行语的使用本来就是一种跟风行为, 使用网络流行语是为了显得自己“前卫”“潮流”。当流行语被推向公众视野时, 相比其他指称类别的流行语, 特征指称类能够随意地用于谈论各种事物现象, 适用范围更广, 因此也更容易达到较高的使用频率。

3. 交际功能

本研究所定义的交际功能大致相当于雅柯布森的“寒暄功能”(phatic), 其用意并不在于传递具体的信息, 而是与沟通渠道有关, 用于“建立交流、延长交流、终止交流; 为了检查交流的渠道是否通畅; [……] 为了吸引对方的注意力”(雅柯布森 2012)。交际功能具有四个小类。抒发情感如“蓝瘦香菇”, 发表评价如“辣眼睛”“涨姿势”, 阐发哲理如“且行且珍惜”, 开启对话如“× × 了解一下”“冲鸭”。

整体而言, 流行语具有交际功能与持续热度小 (流行时间短) 相关, 这可能与同类表达数量多造成的频繁更迭有关——在我们的样本中, 抒发情感类和开启对话类流行语都

有较大的数量(两类分别是数量上排名第1和第3的特征类别)。

除了以上两类,开启对话和阐发哲理类流行语也与较低的持续热度和顶峰热度关联。如果进一步挖掘这个现象的成因,或许与微博平台这个特定的媒介有关。先讨论为什么具有开启对话功能的流行语的持续热度和顶峰热度低。区别于其他非面对面的沟通媒介,微博是一个非即时沟通的平台,因此它不像电话等媒介,需要随时随地确认沟通渠道的顺畅,也不需要沟通的开始和结束有一套共识性的仪式——微博的沟通并非始于用户发帖的那一刻,而是始于其他用户对帖子的阅读;相应地,微博沟通的结束也并非终于用户发帖的那一瞬,它并不具有严格意义的终结点,只要帖子不删,只要任何用户还有任何渠道可以阅读这个帖子,它的沟通使命就未完成也未结束。换言之,微博平台并没有明显的对“寒暄/开启对话”的需求,因此,那些呈现出明显的“寒暄/开启对话”功能的流行语受到沟通媒介的限制,并未受到用户的青睐。而阐发哲理类流行语的反热度效应更强。发表评价、抒发情感、阐发哲理类流行语在功能上都属于“表情功能”,都属于用户定义微博平台的主要功能。但是相比其他类别,阐发哲理类流行语[比如“世界这么大,我想去看看”“生活不止(眼前的苟且)”]的广泛使用,一方面需要具体事件的触发,同时更需要其他用户的认同和呼应才有可能得到进一步传播,其使用场景相比其他交际功能的流行语更加局限,加之阐发哲理类流行语通常字数更多,由此更不容易出现高热度的使用。

4. 语言特征

此前的研究发现,标新立异的语言形式可能与新词语的流行有关。本研究所标注的几类“语言特征”变量都是通过形式上的标记性(或称“陌生化”)引起语言使用者的注意,这是网络语言的一个突出特征。这个大类包括谐音词(笔芯—比心)、缩略词(高大上—高端大气上档次)、字母词(xswl—笑死我了)、外来词(控—complex)、四字格式(不明觉厉)、方言词(猴赛雷—粤语“好犀利”),另外,新短语指符合汉语造词规律、用既有词汇造出的组合词,如“颜值”“点赞”“女汉子”,而新词往往不能从汉语的造词规律生成,如“达人”“悲催”。

在我们的样本中,有43.2%使用了这些标记性特征中的至少一种形式。过去十年,语言文字专家对网络语言的各种标记性形式不乏关注,尤其担忧字母词、四字格等形式的网络语言破坏了汉语的“纯洁性”。从统计结果来看,字母词容易出现高顶峰热度(比如“hold住”“duang”“打call”),但对于持续热度的表现没有显著差异;四字格形式对于两个热度变量的表现均没有显著的差异。相比之下,新词(比如“悲催”“土豪”“给力”)和外来词(比如“吐槽”“控”“切克闹”)与高热度的关联非常显著。其中,新词类的网络流行语尽管数量极少($n=17$),但新词与非新词在两个热度上的数值差距明显——新词的持续热度平均值接近8年,远高于非新词的4.18年;新词的顶峰热度接近4(相当于汉语词频排名前3000的水平),而非新词的顶峰热度仅为1.85。外来词的数量也并不多($n=18$),包含了多种外来语本土化形式,比如完全音译的“控”(来自英语单词complex)、“狗带”(来自中式英语go die),完全意译的“区块链”“硬核”,以及夹有英文单词的“hold住”。大部分外来词在形式上符合汉语的造词法,只有极少数使用了陌生化

的形式(比如来自英文说唱中常用的 check it out 的汉语音译“切克闹”)。但外来词与非外来词在顶峰热度的表现上差距明显——外来词的平均值为 2.89,非外来词为 1.89,正好处于顶峰热度平均值的两边。

从以上的事实我们可以推测,在这几种标记性形式中,新词和部分外来词由于具有规范性、符合汉语词汇语法的系统规则,所以最能被使用者接受,也最有可能保留在语言系统里;其他的标记性形式因为呈现出新奇感容易达到相对高的使用频率(比如字母词),却因为表义不透明(比如谐音词、缩略词、四字格)阻碍了它们在语言使用社区的传播,因此没有持续流行的优势。

5. 其他特征

在其他特征中,我们考察了旧词新义、框式结构(“中国式××”“舌尖上的××”)、粗鄙字眼几种情况,其中只有旧词新义显示出与顶峰热度的相关性。

旧词新义是网络语言创新的一种重要手段(在本研究的样本中占 13.9%),比如“割韭菜”“盘他”“佛系”“辣眼睛”都属于旧词赋予新义的情况。它的“新”不如形式标记手段那么明显,但也并不那么隐蔽——当读者试图通过上下文解读遭遇失败,不得不推翻原有义项之后,一个新义项存在的可能性便浮现出来。因此,旧词新义作为一个变量,和新词、外来词一样,对于顶峰热度有显著的差异。在使用上,使用者并不会故意(甚至会避免)解释新义项的内涵,因为这种让人“不知所云”的类似于“黑话”的用法恰恰赋予了使用者一种圈层私密感和优越感。

框式结构是语言学本体研究中关注较多的一种现象,比如“中国式××”“舌尖上的××”“被××毁掉的××”。但这个形式的流行语在本研究中并未表现出对持续热度和顶峰热度的敏感性。

6. 来源

未发现网络流行语的来源对于热度变量呈现显著差异。

(二) 统计模型

通过相关性分析,我们确认存在与网络流行语使用热度具有相关性的几个变量。以下,我们使用多重线性回归分析,一方面,进一步确认以上变量中哪些具有影响关系、其影响方向和程度如何,另一方面,探索通过模型进行热度预测的可能性。

经检验,数据集符合多重线性回归分析的几个假设,可以使用进行建模。我们使用 SPSS 的线性回归分析模块,基于上一部分的结果输入可能影响模型的多个自变量,软件会比较采用不同自变量建模的准确度,并提示各个因素对提高模型准确度的效应(即“Sig. F Change”一项的数据)。我们在比较了不同的模型后,选择了仅包含对提高模型准确度有显著效应的因素。(1)、(2)是多重回归分析所得出的持续热度和顶峰热度的模型公式。

$$(1) \text{持续热度} = 5.847 - 0.403 * \text{长度}_{(\text{字})} + 3.077 * \text{新词}_{\text{是}=1}$$

$$(2) \text{顶峰热度} = 2.307 - 0.151 * \text{长度}_{(\text{字})} + 0.394 * \text{特征指称}_{\text{是}=1} + 0.512 * \text{外来词}_{\text{是}=1} + 1.437 * \text{新词}_{\text{是}=1} + 0.441 * \text{旧词新义}_{\text{是}=1}$$

两个公式中,持续热度公式的拟合度为 0.18,说明持续热度变异的 18% 可由流行语

长度和是否为新词来解释;顶峰热度公式的拟合度为 0.20,说明顶峰热度变异的 20% 可由长度,是否为特征指称,是否为外来词、新词、旧词新义来解释。

从拟合度来看,两个公式作为预测模型尚不够理想。如果我们把结果仅分为两类考察,其准确性有显著提升。从现有的数据集来看,如果把持续热度以 4.38 划分为高/低持续热度,准确度可达 66.97%;顶峰热度以均值 1.94 划分为高/低顶峰热度,准确度可达 70.06%(见图 8)。

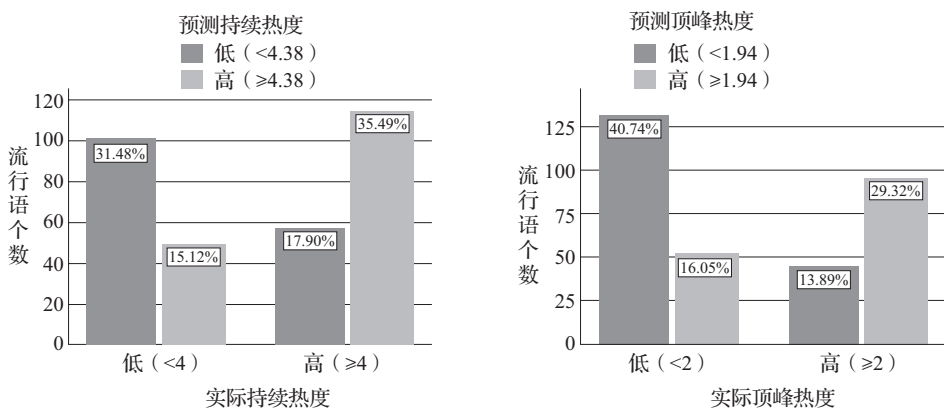


图 8 二分法调整后模型的准确度:持续热度(左)、顶峰热度(右)

为了进一步检验回归模型对其他数据集的分析能力,我们又抓取了 2020—2023 年共计 4 年的微博数据(与 2010—2019 年的数据结构一致、体量相当),并收集了 119 条这四年间的流行语,运用(1)、(2)的模型公式进行计算。结果显示,采用分类法,对于持续热度整体的预测准确率虽然仅有 10.9%,但对于实际达到高持续热度的流行语($n=10$),准确率达到 80%——实际上,由于许多流行语在 2023 年还未停止流行,实际的准确率可能会更高;对于顶峰热度的预测能达到 53.8%的准确率——由于预测顶峰热度相对实际整体偏高,模型对实际达到高顶峰热度的流行语的预测准确率更高,达到 72%。

四、结 论

本研究的初衷是,通过发现网络流行语自身特征与其使用热度的规律,为辞书的新词收录提供参考。本研究考察了 2010—2019 年十年间的 324 个网络流行语,离析出了 10 个对使用热度有显著效应的因素,并建立了有一定预测能力的模型。

根据模型我们发现,在所考察的特征中,指称具体范畴的网络流行语及来源为外来词和新词的流行语最有可能有较为广泛的使用和较强的生命力,这是辞书收录新词时最应重点考虑的类别。而意义相对虚无缥缈的、实现交际功能的流行语,以及形式上为字母词、四字格、谐音词、缩略词的流行语并不同时满足使用普遍和生命力强的收录要求。总体而言,长度短的流行语使用热度更高。

本研究预测模型的准确性需要进一步加强,但在辞书收录新词时可作为参考指标。使用频次(使用普遍性)毋庸置疑是一个重要因素,但康健等(2019)的研究显示,

《现汉》第 7 版新收录的词中虽然高频词居多,但低频词也未必就被排除在外。本研究模型的价值尤其体现在对新词生命力维度的预测,但是否收录的具体阈值还需要专家做出评估后设定。辞书的新词收录是一项高难度的系统工程,还有多重因素需要权衡考虑。(晁继周等 2000; 周一民 2003)本研究仅在“普遍性”和“生命力”两方面做讨论。

附 注

[1] “新浪微博”于 2014 年 3 月 27 日更名为“微博”,下文统称“微博”。

[2] 官方榜单包括:“汉语盘点”的“年度十大网络用语”,《咬文嚼字》编辑部的年度“十大流行语”,以及新华网、人民网的评选。非官方的榜单来源于萌娘百科等网站论坛。

[3] 一些流行语因为无法通过设置关键词检索到准确的语料而被剔除,包括:(1)旧词新义(如“油腻”“流量”“安排”等);(2)关键词过于简短(如“微×”“说×就×”);(3)开放框式结构(如“休假式治疗”)。

[4] 为了更为准确地进行比较,我们对词频进行标准化(normalization)处理,按照当月的语料字数计算其在每百万字语料中出现的频次。

[5] 汉语词汇词频的排名来源为“语料库在线网站”(www.cncorpus.org)。

[6] 由于样本量较大,在单因素方差分析(one-way ANOVA)和独立样本 t 检验之间我们选择了前者。

[7] 由于样本未通过齐性检验,相关性分析我们使用了斯皮尔曼而非皮尔森(Pearson)。

参考文献

1. 晁继周,黄华.关于规范型词典收词问题的再思考.//中国辞书学会学术委员会编.中国辞书论集(1999).上海:上海辞书出版社,2000.
2. 韩敬体.增新删旧,调整平衡——谈《现代汉语词典》第 5 版的收词.中国语文,2006(2).
3. 季乃礼,焉超越,陈问天.网络热词的生命周期与传播机制研究——基于 2012—2020 年 73 个网络热词的聚类分析.山东科技大学学报,2021(6).
4. 江蓝生.守正拓新 与时俱进——《现代汉语词典》第 6 版修订述略.秘书工作,2012(10).
5. 康健,刘柏君.《现代汉语词典》第 7 版新增词语的收录原则及特点.辞书研究,2019(3).
6. 罗曼·雅柯布森.雅柯布森文集.钱军译.北京:商务印书馆,2012:60.
7. 沈家煊.世说新语三则评说——被自杀·细小工作·有好酒.当代修辞学,2010(4).
8. 王灿龙.“被”字的另类用法——从“被自杀”谈起.语文建设,2009(4).
9. 王昊.基于百度指数的网络热词关注度分析:以互动百科 2010—2012 年年度十大热词为例.新闻传播,2014(5).
10. 王炎龙,夏赛楠.网络语言强势模因的传播变异研究——基于 2012—2016 年度网络用语的分析.新闻界,2017(8).
11. 张志毅,张庆云.新时期新词语的趋势与选择.语文建设,1997(3).
12. 中国社会科学院语言研究所词典编辑室编.现代汉语词典(第 3—7 版).北京:商务印书馆,1996、2002、2005、2012、2016.
13. 周一民.新词新义的收录与规范.语言文字应用.2003(4).

14. Boulanger V M. *What Makes a Coinage Successful? The Factors Influencing the Adoption of English New Words*. Ann Arbor, MI: UMI Dissertation Services, 2002.
 15. Hino A, Fahey R A. Representing the Twittersphere: Archiving a Representative Sample of Twitter Data under Resource Constraints. *International Journal of Information Management*, 2019 (48): 175-184.
 16. Jing-Schmidt Z, Hsieh S. Chinese Neologisms. //Latham K. *The Routledge Handbook of Chinese Applied Linguistics*. London: Routledge, 2019.
 17. Metcalf A A. *Predicting New Words: The Secrets of Their Success*. Boston: Houghton Mifflin Company, 2004.
 18. Peduzzi P, Concato P, Kemper E, *et al.* A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, 1996 (12): 1373-1379.
- (彭馨霞 中国社会科学院语言研究所语料库暨计算语言学研究
中心 / 中国社会科学院语言学重点实验室 北京 100732;
吴婷婷 青岛市崂山区金家岭学校 山东 266000;
严文静 北京百度网讯科技有限公司 北京 100094)
(责任编辑 郎晶晶)

(上接第 32 页)

61. 章宜华. 语义·认知·释义. 上海: 上海外语教育出版社, 2009: 78-83.
 62. 章宜华. 基于模因仿造的欧美外来词汉化机制研究. 语言文字应用, 2024 (2).
 63. 赵明. 明清汉语外来词史研究. 厦门: 厦门大学出版社, 2016: 5.
 64. 钟吉娅. 汉语外源词: 基于语料的研究. 华东师范大学博士学位论文, 2003: 90-95.
 65. 钟焜茂. 现代汉语对外来词的接受与汉化. 龙岩师专学报, 1993 (2).
 66. Bussmann H. *Routledge Dictionary of Language and Linguistics*. Beijing: Foreign Language Teaching and Research Press; 2000.
 67. Dawkins R. *The Selfish Gene (2nd Ed.)*. Oxford: Oxford University Press, 2006.
 68. Fauconnier G, Turner M. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books, 2002.
 69. Zhang Yihua. Memetic Calque of Chinese in English for Cross-cultural Communication: A Corpus-based Study of Language Contact. *Current Research in Chinese Linguistics*, 2020 (2).
- (黄婷婷 广东外语外贸大学词典学研究中心 广州 510420 /
江西科技师范大学文学院 南昌 330038;
章宜华 广东外语外贸大学外国语言学及应用语言学
中心 / 词典学研究中心 广州 510420)
(责任编辑 马 沙)