文章编号：1003-0077(2021)01-0017-08

中文词汇增长研究

王 珊[1,2]，王会珍[3]

（1. 澳门大学 人文学院，澳门；
2. 澳门大学 协同创新研究院，澳门；
3. 东北大学 计算机科学与工程学院，辽宁 沈阳 110167）

摘 要：词汇增长研究能够分析文本的 TTR 在不同时期的变化，该文选取 1954—2018 年的中国政府工作报告为语料，分析文本中词例与词种的曲线变化，挖掘政府工作报告中的词汇丰富度与政策的相互关系。该文首先对语料进行了分词，然后根据曲线拟合效果选择拟合更好的 Heaps 模型进行预测。以中国的“五年计划”作为基础时间周期，对各周期模型预测值与现实观测值的差值进行分析，并与随机打乱后的文本计算结果进行对比，进一步验证了实验的结果。研究发现随着时间变化，词汇增长呈现出一定的倾向性：在深化改革、新政策出台等时期，一般需要更多的词语来描述，此时观测值高于预测值，而在政策相对稳定的时期，对原有词汇的使用较多，此时观测值低于预测值。该文以中文语料作为研究对象，分析其历时变化，能够为中文词汇增长研究提供借鉴。

关键词：中文；词汇增长；词汇丰富度；TTR；Heaps 模型
中图分类号：TP391 **文献标识码**：A

A Study of Chinese Vocabulary Growth

WANG Shan[1,2], WANG Huizhen[3]

(1. Faculty of Arts and Humanities, University of Macau, Macau, China;
2. Institute of Collaborative Innovation, University of Macau, Macau, China;
3. School of Computer Science and Engineering, Northeastern University, Shenyang, Liaoning 110167, China)

Abstract: Vocabulary growth research is based on the type-token-ratio (TTR) changes of the texts in different periods. This article selects *Reports on the Work of the Chinese Government* from 1954 to 2018, analyzes the curves of tokens and types in the texts, and explores the interaction between vocabulary richness of the reports and the policies. It first conducts Chinese word segmentation on the corpus and then selects the Heaps model for prediction according to different curve fitting effects. Taking China's Five-Year Plan as the basic time cycle, the difference between the predicted value and the observed value of each cycle is compared with that of the random texts. The study reveals that vocabulary growth with time changes shows a certain tendency: in the period of deepening reforms and launching new policies, more words are needed to describe the phenomenon and the observed value is higher than the predicted value. With the analysis of the diachronic changes of Chinese texts, this paper provides references for the study of Chinese vocabulary growth.

Keywords: Chinese; vocabulary growth; vocabulary richness; TTR; Heaps model

0 引言

在计量语言学领域，文本中词例(tokens)与词种(types)的关系研究是重要的研究方向，二者数量的比值 type-token-ratio(TTR)是衡量文本词汇丰富程度的有效指标。大量关于 TTR 的研究用于分析文本的词汇丰富度特点，进而研究不同作者、语言、内容、表达方式等方面的特点。不同的文本词汇丰富度有差别，TTR 值也有差异，如统计不同文章的 TTR 值，有助于判断文本作者[1]。在已知作者身份的情况下，研究其文章内容的 TTR 值，可以分

收稿日期：2019-09-09 **定稿日期**：2020-08-17
基金项目：澳门大学多年研究基金(MYRG2019-00013-FAH)和启动基金(SRG2018-00126-FAH)

析作者的语言风格[2]。相同的文本,在不同区域其TTR值也有变化,体现出不同区域文本内容的特点。分析文本时,由于词汇数量会影响TTR值,所以不可以直接计算TTR值进行比较,而往往采取以下两种方法:其一是使用移动平均TTR,确定固定长度的窗口,再统计窗口内出现的词种数[3];其二是使用拟合效果良好的TTR模型,如Heaps模型[4]等,预测出TTR值在不同词数下的值,通过统计结果与预测结果的差值,分析局部区域词汇丰富度的特点。

当研究对象是按照时间顺序组合的文本时,不同时期文本的TTR值体现了相应时代特征,如利用美国总统两百年来的演讲语料,分析TTR增长与各时期社会特点、总统政策之间的关系[5]。然而,这类研究的对象多为有自然词语划分特点的语言,目前还缺少利用TTR预测模型对中文进行的分析。在中文等亚洲语言中,文本是连续而不具有自然划分性质的,这给TTR的研究带来困难。为弥补TTR研究在中文词汇增长领域的空缺,本文借鉴Savoy[5]的研究方法,选取1954—2018年的中国政府工作报告作为研究对象,阐述了中文TTR分析的可行性。在比较了Heaps模型与Hubert模型的拟合效果后,本文采用Heaps模型作为研究的预测模型。通过不同阶段统计词种数与预测词种数的比较,分析了不同阶段该差值与政策之间的联系,并使用随机乱序的文本进行了模型效果的验证。

1 相关工作

TTR是句子中词种数量与词例总量的比值,如在句子"农业贷款的增加和农村信用合作的发展"中,共有10个词语(农业、贷款、的、增加、和、农村、信用、合作、的、发展),9类词语(农业、贷款、的、增加、和、农村、信用、合作、发展),其TTR即为9/10。TTR体现了单位长度的文本中出现的词种数量,可以用于衡量词汇的丰富程度。

不同类型的文本,作者、语言不同等,TTR值也存在差异,对文本数据的分析与预测有重要意义。在已知作者身份的文本中,TTR值可以用于分析不同作者的表达特点,如对特朗普与希拉里在2016年竞选期间的辩论与演讲等语料,利用TTR分析了其语言风格与修辞特点[2]。根据实验结果,特朗普的TTR要小于希拉里,说明其语言更为简单直接,往往避免复杂的语法,少用修辞,而多用短句,希拉里更善用修辞和复杂的表达方式,这样的分析结果也与使用词汇密度分析的结果相同。在未知作者的文本中,TTR值可以用于判断作者的身份,如有的研究对12位作者的词种数量进行统计分析,证明其词汇丰富度与作者身份的强关联性[6]。除此之外,不同语言的TTR值也有不同,如对于21种语言的词汇复杂度进行了统计分析[7],发现语言的TTR、MATTR值与使用语言熵衡量词汇复杂度的方法,结果具有一致性[8]。

同一种文本,不同部分词汇的丰富度也有不同,可以使用移动平均TTR来分析。固定词的数量,对于文本的不同位置统计出现的词种数。这样计算的TTR值被称为MATTR(移动平均TTR)。Covington等人[3]提出了一种基于窗口的、快速计算MATTR特征的算法,采用此方法分析*The Adventures of Sherlock Holmes*,发现文章内容与MATTR的关系:MATTR值在每个故事的开始会上升,而在冗长的对话中呈下降趋势,这说明MATTR值对于分析文本内部的风格同样具有帮助。

词例与词种数量的增长关系可以用数学函数刻画。在词种较少时,词种数量与文本长度几乎保持1∶1的增长关系。随着语料库的增长,其梯度逐渐下降。对此,许多学者对于增长过程进行了建模分析。为建立这样的关系,指数类型的预测模型被提出[4],如式(1)所示。在这个模型中,文本词种数量V'被看作是以词例数量为自变量n的函数。对于等式进行以自然常数e为底的对数变换,得到等价的线性关系,如式(2)所示。

$$V' = an^C \tag{1}$$

$$\ln(V') = \ln(a) + C\ln(n),\ 0 < C < 1 \tag{2}$$

这样的模型较好地拟合了观测的TTR增长曲线,但也存在一些弊端。对于TTR计算中的常数a和C等,并不是常量,其变化表现出随机性[9],也难以进行解释。针对此现象,提出更加复杂的模型。文本中的词汇可以分为常用词与不常用词两类。不常用词,如时间、数量及一些专用名词等,往往在文本中不会重复出现,这导致它们的数量关系与类别数量关系表现为梯度为1的线性函数。而对于那些常用的词汇,如一些助词、介词等,其词语的数量要远远大于类别的数量。假设前一类词语占总词语比例为p,一种基于常用词与不常用词比例的模型得以提出,如式(3)所示。[1,10]。

$$V'(u) = puV + (1-p)\left[V - \sum_{i=1}^{k} V_i\ (1-u)^i\right] \tag{3}$$

其中，i 指词语出现的次数，V_i 指出现 i 次的词语的数量。p 指语料中只出现过一次的词所占的比例，$(1-p)$指出现多次的词占的比例。p 作为模型中唯一的参数，反映了语料中常用词与不常用词的比例。u 指用于预测语料占总语料的比例，当 $u=1$ 时，式(3)即为全部语料的预测结果，如式(4)所示[5]。

$$V'(1.0)=pV+(1-p)V \tag{4}$$

该模型考虑了词语出现的概率分布，与 Heaps 模型相比，只需要一个参数 p，即可预测 TTR 的增长关系。通过对两百年来美国总统的演讲语料，计算得到 $p=0.453$[5]，两位著名法国作家 P Corneille 和 Racine 的文章使用该模型计算的结果分别为 $p=0.02, 0.33$[1]。可见在不同的文本环境下，p 有较大变化，这与不同语言的词汇复杂度也有很大关系。随着以上模型的提出，一些专门用于 TTR 计算的软件也开发了出来[11]。

以往对文本特点的分析往往聚焦于文本使用的词汇本身，缺少历时分析，忽视了更为普遍的规律；常侧重于对文本个体的分析，忽略了不同文本之间的相互关系。本文选取中国从 1954 年到 2018 年的政府工作报告作为文本材料，使用 Heaps 模型，对 TTR 值进行建模分析。

2　研究方法

2.1　语料选取

政治性的演讲、发言往往反映了时代关注的热点，对于社会发展变化有着很强的预测性。此类文本具有权威性、公开性，又蕴藏着珍贵的社会价值，因而被广泛用作定量语言学研究的文本材料。如使用 2007—2008 年 Barack Obama 和 John McCain 等人的发言，分析各自的语言特点[12]；采用法国大选电视辩论语料作为素材，分析不同情感倾向词汇的分布[13]；使用中国政府官员发言，利用 TTR 与语言信息熵分析发言人词汇丰富度与社会、教育信息的关系[14]。

本研究选择国务院政府工作报告作为实验的文本材料。政府工作报告是中国政府的一种公文形式，是中国政府对国家建设发展的年度总结，各级政府必须在地方人民代表大会和政治协商会议的年度会议上，向大会主席团、人大代表和政协委员发布。报告的内容主要为国家发展的阶段性总结与未来规划，反映了各时期中国社会面临的主要任务与时代特征，其内容具有客观性、概括性。此外，它们始终保持着固定的文风，这对降低实验的误差有重要意义。中国从 1953 年开始制定第一个五年计划，1954 年第一次发布政府工作报告，截至 2018 年，除1961—1963、1965—1974、1976—1977 期间受社会其他因素的影响，政府工作报告有缺失现象外，其他年份均每年发布一次，在时间上具有连贯性，很大程度上提高了本文实验的置信度。

2.2　语料预处理

中文文本中词语的切分是一个复杂的问题。一些基于词典的分词方式，如“正向最大匹配算法”“最少词数匹配算法”等先后被提出，分词效果得到逐步改善。近年来，统计机器学习的方法，如隐马尔科夫模型、条件随机场模型、神经网络算法也被用于分词。本研究采用了 NLPIR- ICTCLAS 分词系统①，它由张华平博士开发维护，在 2002 年“中国 973 评测”、2003 年“国际 SIGHAN 分词大赛”中获得综合第一名的成绩，2010 年获得“钱伟长中文信息处理科学技术奖”一等奖，是当今汉语分词最可靠的系统之一。该系统拥有“新词发现”功能，在较长的文本内容中，可基于信息交叉熵自动发现新词语，适合对词种的研究。在使用该系统自动分词后，人工审核并修正分词结果。

2.3　使用的模型

本文对比了 Heaps 模型与 Hubert 模型，对政府工作报告中 TTR 进行了建模预测。采取深度学习框架 Pytorch，采用随机梯度下降的方法，以均方差(the mean squared error，MSE)[5]为损失函数，拟合了这两种模型。其中，Heaps 模型得到的参数为 $a=e\hat{}2.857$，$C=e\hat{}0.5137$，Hubert 模型中比例参数 $p=0.0711$，如式(5)所示。

$$\mathrm{MSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}[V'(i)-V(i)]^2} \tag{5}$$

两种模型预测值的总体偏差不同。Hubert 模型认为在文本长度与总长度比例为 $u:1$ 的文本中，整个文本中只出现一次的词语，出现的几率是 u。在整个文本中出现了 i 次的词语不出现的几率是 $(1-u)^i$。这一结论是由词语在文本中的出现几率只与文本长度有关的假设推导的。

对于更为一般的情况，若假设词语 w 在文本 C 中的分布满足概率函数 $F_w(X)$，则整个文本中只出

① http://ictclas.nlpir.org/

现一次的词语，在长度比例为 u 的文本中，出现的几率为 $F_w(X=u)$。而在整个文本中出现了 i 次的词语，在长度比例为 u 的文本中不出现的几率为 $(1-F_w(X=u))^i$。因而一般化的预测模型如式(6)所示。

$$V'(u)=pF_w(X=u)+(1-p)\left[V-\sum[1-V_iF_w(X=u)^i]\right] \quad (6)$$

当 $F_w(X=u)=u$ 时，式(6)即为 Hubert 模型。文本中出现的词汇可分为两类：第一类词汇在文本中的分布满足均匀分布，则其满足 Hubert 模型的假设；第二类词汇在文本中的分布不满足均匀分布，则该类词在式(6)中计算得到的值与 Hubert 模型不同。政府工作报告与国家发展阶段息息相关，且具有明显时代特点，不同词汇在不同阶段的分布是大不相同的，即存在一定第二类的词汇，对 Hubert 模型的预测结果造成了影响。这解释了实验中 Hubert 模型与现实值偏离更大的原因。而 Heaps 模型不考虑词语在文本中的频率分布，因而受政府工作报告中局部特征差异明显的影响较小。Heaps 模型主要依靠函数增长的数学特征进行预测，符合现实观测值曲线的增长趋势，得到了更好的拟合效果。由于 Heaps 模型预测值的总体偏差更小，因而我们选其拟合后的曲线作为实验的预测模型。

3 词汇增长模型

3.1 政府工作报告的总体情况

1954—2018 年政府工作报告词例数量为589 000，实验中每 590 个点选取一个作为采样点，用于 Heaps 模型与 Hubert 模型的拟合。对于拟合得到的曲线，以及这 1 000 个采样点，绘制得到图 1 与图 2。

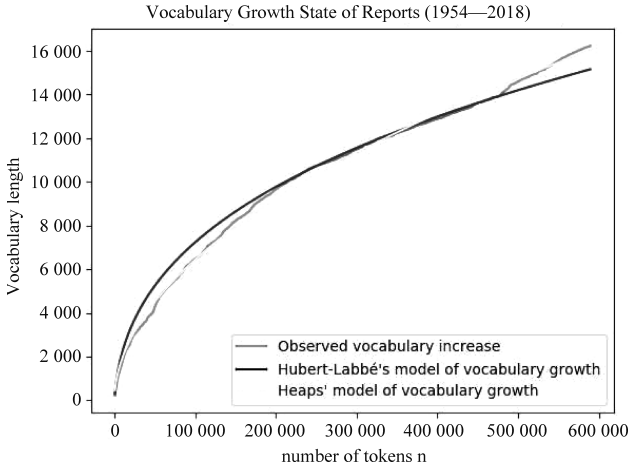

图 1 两种模型预测曲线

图 1 显示了两种预测曲线与现实曲线，展示了现实中与两种模型中词种与词例数量的增长关系。在词例数量较小时，词种数量随其迅速增长，而当词例数量较多时，其增长的速度会降低。

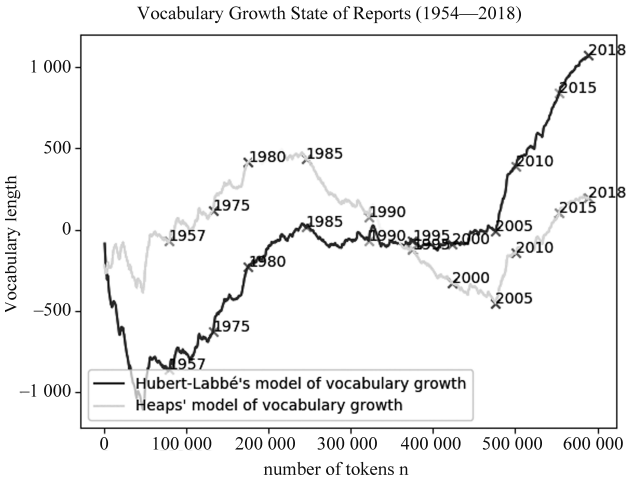

图 2 两种模型预测值与现实观测值之差

图 2 体现了 1954 年到 2018 年间，两种模型与观测值的差值，并对五年计划(详见 3.2)的结束年份进行了标注。虽然 Heaps 模型在许多位置上仍然不能与现实值达到完全匹配，但这是由政府工作报告自身的词语分布特点决定的。整体上，Heaps 模型的拟合效果要好于 Hubert 模型，因此在下文的词汇增长分析中我们采用了 Heaps 模型。

3.2 词汇增长

中国的五年计划，是每五年中国政府对国家重大建设项目、生产力分配分布和国民经济重要比例关系的规划。在每个五年计划开始的年份里，政府都会对于旧的五年计划做总结，而对新的五年计划进行部署。若使用每一年作为分析的周期，容易受到该年份随机时间的影响，其结果具有偶然性。五年计划作为政府工作的一个阶段的建设方案，具有整体性与稳定性，表现了一个较长时期中国的发展状态，以其作为分析周期，可以避免部分年份的突变，具有说服力。

从 1954 年至今，一共有十三次五年计划，选其作为最小的时间周期，结合曲线的增长特征，对于现实观测值与 Heaps 预测值进行分段分析，结果如表 1 所示。

表 1 对政府工作报告在不同阶段的词例数量、词种数量、Heaps 预测值、新词语数量等信息进行了详细展示。在中国建国初期，受当时社会因素影响，

表 1　每阶段现实观测值与 Heaps 预测值

时间段	词例	词种	Heaps 预测值	观测值-预测值	观测值 TTR	预测值 TTR	新词语数量	该阶段政府工作报告数量
1953—1957	78 987	5 710	5 641	69	0.072 29	0.071 42	\	4
1958—1964	129 096	7 349	7 435	−86	0.056 93	0.057 59	(在上一个阶段没出现的词语)1 639	4
1966—1975	132 162	7 439	7 554	−115	0.056 29	0.057 16	(在前两个阶段没出现的词语)100	1
1976—1980	174 941	8 591	9 002	−411	0.049 11	0.051 46	1 152	5
1981—1985	245 450	10 224	10 662	−438	0.041 65	0.043 44	1 633	5
1986—1990	321 936	11 752	11 829	−77	0.036 50	0.036 74	1 528	5
1991—2000	424 164	13 541	13 207	334	0.031 92	0.031 14	1 789	10
2001—2005	476 151	14 369	13 915	454	0.030 18	0.029 22	828	5
2006—2010	500 679	14 745	14 604	141	0.029 45	0.029 17	376	5
2011—2015	553 933	15 633	15 531	−102	0.028 22	0.028 04	888	5
2016—2018	589 990	16 042	16 236	−194	0.027 19	0.027 52	668	3

工作报告在 1961—1963，1965—1974，1976—1977 年出现了缺失现象。对于这一阶段的研究，我们选取其附近年份中有代表性的报告来体现这一阶段的特征，如 1975 年工作报告内容主要为对之前数年工作的总结，因而使用 1975 年政府报告，补充 1965—1977 年整体的缺失。

在中国第一个五年计划中（1953—1957），Heaps 模型预测结果要略小于现实观测值，截止到 1957 年，现实观测值为 5 710，模型预测值为 5 641，前者较后者多 69 个词种，说明此时有更多的新词语，这与当时中国所处的历史背景是有关联的。1953 年到 1957 年，中国进行了第一次工业化建设。所产生社会变化需要更多的词汇去描述，这些新词语主要包括“油菜籽”“烧碱”“公私合营”“合作小组”“改造”等。该时期的政府工作报告还出现了许多的数字性的发展指标，以及“农业生产合作社”等新兴事物。

在中国第二个五年计划（1958—1962）及 1963—1964 年的国民经济恢复时期，预测结果大于现实观测值，截至 1964 年，现实观测值为 7 349 个词，模型预测值为 7 435 个词，前者较后者少 86 个词，说明该阶段中国的政策相对稳定。该时期中国进行了人民公社化等运动，国家以快速工业化建设为主要奋斗目标，政府工作报告聚焦于工业、农业的建设。在 1962 年之后，中国进入了国民经济恢复时期，此时政府统筹兼顾农业、制造业与工业的发展关系，提出了全面的发展策略。这与 1964 年的增长趋势是匹配的，此阶段新增词汇 1 639 种，如“虫害”“轴承”“学兵”等。

在中国第三到第四个五年计划时期（1966—1975），仅在 1975 年有政府工作报告，主要内容为对十年期间的世界局势做出总结，这篇报告的预测结果要大于现实观测值，截止 1975 年，现实观测值为 7 439 个词，模型预测值为 7 554 个词，前者较后者少 115 个词。随后的第五个五年计划（1976—1980）中，中国政府将工作重心回归到了经济发展上来，政府工作相对稳定。该阶段出现的新词共 1 152 种。

在第六个五年计划时期（1981—1985），政府工作报告中现实观测值与模型预测值差值的扩大趋势明显减弱，该差值由上一个阶段的 411 变化到 438（预测值 10 662，观测值 10 224），可以认为是差值变化的正常浮动。这是因为这段时期中国的体制改革与对外开放逐步加深，市场经济得以承认，一些经济特区先后开放，人口、教育、外交、能源、交通方面的内容也有增加，出现的新词共 1 633 种，多属于社会活动和具体事物，如“二胎”“多子多福”“晚婚”“精神文明”“合资企业”“一国两制”等。

第七个五年计划（1986—1990）中，中国的科技、教育、经济等各个领域都得到了进一步的发展，出现新词共 1 528 种，多为各领域的具体事物，如“义务教育”“养老保险”“展览馆”“信贷”“基金”等。

第八个与第九个五年计划（1991—2000），是改革开放推进的快速时期，与上一阶段比，现实观测值保持上升趋势，观测值高于预测值。这段时期社会

主义市场经济的目标、总体开放的格局得到了实现。中国在这一阶段进行了企业制度、教育制度、住房制度的改革等。随着改革开放的深入，中国社会生活发生了广泛而深刻的变化，社会经济成分、分配制度、就业方式等进一步发展。这一阶段词种数量继续保持稳步上升趋势，共增加新词语 1 789 种，如“通讯卫星”“租赁制”“股份制度”“保险”“再就业”体现了各个领域的迅速发展。

在第十个五年计划时期(2001—2005)，市场经济地位得到进一步发展，现实观测值高于预测值，前者保持增长趋势。这一时期中国加入了世界贸易组织，中央提出科学发展观的战略思想，对城乡发展、区域发展、人与自然和谐发展、可持续发展做出了进一步阐释。在此时期的报告中，“西部大开发”“东北工业基地振兴”等政策，带来了“青藏铁路”“西气东输”“反垄断法”“信息化”“数字化”等新词语共 828 种，极大地扩展了词种的数量。此阶段词种的模型预测值小于观测值，且后者仍保持着较高的增长速率，与现实中对新政策的阐述需要更多词汇的需求相一致。

在第十一个五年计划中(2006—2010)期间，预测值保持着上升的趋势，观测值高于预测值，新增词语 376 种。这段时期提出提高发展质量，反思开发中的自然环境保护问题，注重可持续发展。

随后的第十二个五年计划与十三个五年计划前段[①](2011—2018)共计新增词语 1 556 种。十一五将反腐败与深化改革、环保问题等经济结构初步转型列为重点目标，十二五以全面建成小康社会为中心目标，提出了创新、协调、绿色、开放和共享五项发展理念。新增词语 1 556 种，如“供给侧结构”“一带一路”“PM 2.5”“两学一做”等，体现了该阶段政府出台的新政策所产生的影响。

4 验证程序

上文以五年计划作为分析的基本时间单位，分析了 Heaps 模型下词种的预测值与现实观测值的差距。两者的差距是由政府工作报告的特点决定的还是由模型拟合的误差造成的呢？本节使用随机化的方法进行验证。

验证的方法如下：以词语为单位，随机化地打乱顺序，生成新的文本。新生成的随机文本的词语总量、词种总数与原文本是完全相同的，但完全打乱次序后的文本不再保持原文本的语义信息，以及原文本的词频分布特征。若可以证明拥有语义信息的原政府工作报告文本的观测值与模型预测值存在较大差距，而随机化处理后的，失去语义信息的文本中此二者差距较小，即可证明政府工作报告的内容是导致这一差值的重要原因；反之，则是由模型拟合的误差造成的。

考虑到每次生成的随机文本可能有偶然性，本文随机生成 1 000 个随机文本，采用与前文实验中相同的采样方式，每个随机文本获得 1 000 个采样点，并根据采样点拟合计算参数 a 与 C，绘制现实观测值与 Heaps 预测值曲线，如图 3 所示。

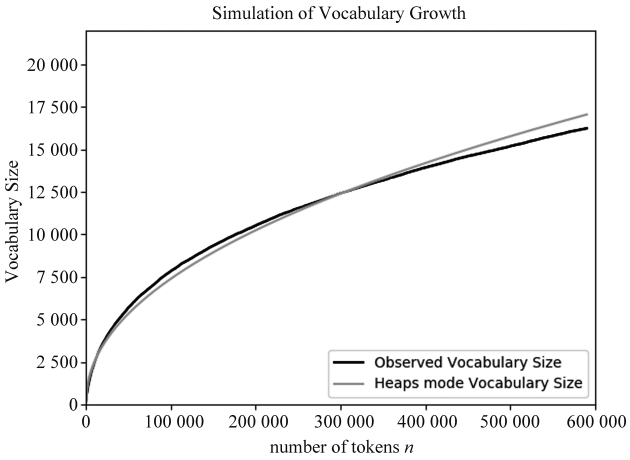

图 3 随机文本预测结果曲线

图 3 展示了随机乱序后的文本，其 Heaps 模型的预测值与现实观测值的关系。经过计算，随机文本拟合得到的参数 $a=e^{\wedge}3.48$，$C=e^{\wedge}0.471\ 4$。通过图 3 可知，此时 Heaps 模型已较好地拟合了观测值绘制的曲线。为了更好地观测二者差值，使用拟合的结果计算预测值与观测值差的标准分数(Z-Score)V''，作为衡量拟合程度的指标[5]，如式(7)所示。

$$V''(n)=\frac{V(n)-V'(n)}{\sigma_{V'(n)}} \tag{7}$$

图 4 反映了乱序后的文本，其标准分数(Z-Score)随词语数量增长的关系。其观测值与预测值之差的标准评分始终在(−2,2)的范围内，因此可以认为该变量符合正态分布(99%以上的数据均在 $-3*\sigma$ 到 $3*\sigma$ 的范围内)。

图 5 表现了 1954—2018 年政府工作报告文本 V'' 的增长关系。尽管图 4 中 Heaps 模型预测的结果并非与现实观测值完全相等，但这一差值(−2～2)远小于乱序前政府工作报告中的差值(−4～5)

① 开展本研究时，政府工作报告发布至 2018 年

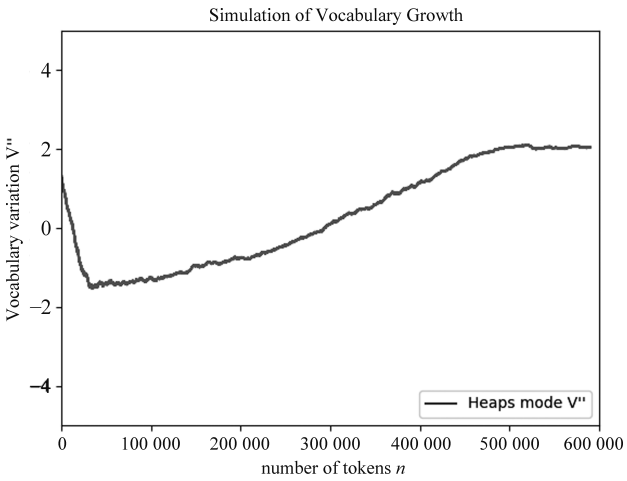

图 4　随机文本预测变化

(图 5)。因此可以说明，政府工作报告中预测值与观测值的差距很大程度上是受报告的内容影响的。

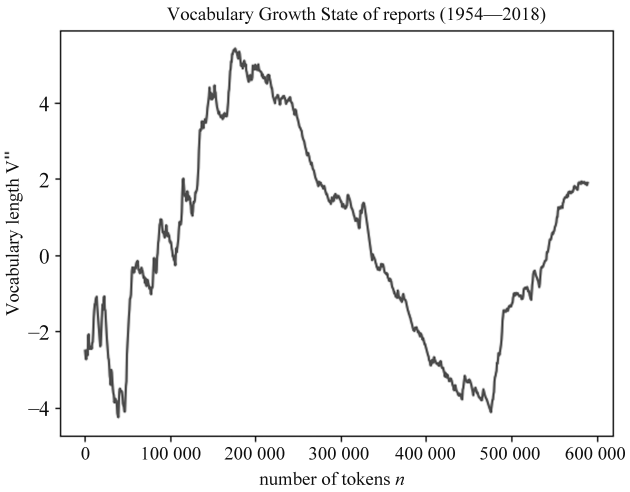

图 5　现实文本预测变化

以上对于中国政府工作报告(1954—2018)的随机化模拟实验证明，词汇增长模型的变化受到工作报告中词汇特点的影响，而非模型误差导致的随机事件。现实观测值与模型预测值之差的变化，是由国家发展重心、所处时代特点等决定的。

5　结论

本文选取 1954 年至 2018 年的中国政府工作报告为语料，根据 Heaps 模型预测值与现实观测值之差，分析了政府工作报告与此差值的关系以及两者之间的联系：在深化改革、新政策推出等时期，需要更多的词汇去描述，此时现实值高于 Heaps 模型的预测值，而在政策相对稳定的阶段，对原有词的复用频率更高，此时现实观测值要低于 Heaps 的预测值。这一结论，也说明了使用 TTR 预测曲线与观测值之差，分析文本词汇丰富度变化的可行性。此外，我们将文本的顺序随机打乱，并使用 Heaps 模型进行了拟合。根据原文本与乱序文本的标准分数(Z-Score)的比较，说明了实验结果的可靠性。此外，以往 TTR 领域的研究，大多采用英语、法语等印欧语系语言，它们本身具有词划分的特性。而中文文本无词汇划分，其 TTR 的统计需要先进行分词处理。本文采用中文语料作为研究对象，分析其历时变化，丰富了对中文词汇增长的研究。

致谢：感谢东北大学张林峰同学为本文数据处理提供的帮助。

参考文献

[1] Labbé C, Labbé D, Hubert P. Automatic segmentation of texts and corpora[J]. Journal of Quantitative Linguistics, 2004,11(3): 193-213.

[2] Savoy J. Trump's and Clinton's style and rhetoric during the 2016 presidential election[J]. Journal of Quantitative Linguistics, 2017,25(2): 168-189.

[3] Covington M A, McFall J D. Cutting the Gordian knot: The moving-average type-token ratio (MATTR)[J].Journal of Quantitative Linguistics, 2010,17(2): 94-100.

[4] Heaps H S. Information retrieval, computational and theoretical aspects[M]. Academic Press, Cambridge, 1978.

[5] Savoy J. Vocabulary growth study: An example with the state of the Union addresses[J]. Journal of Quantitative Linguistics, 2015,22(4): 289-310.

[6] Hoover D L. Another perspective on vocabulary richness[J].Computers and the Humanities, 2003,37(2): 151-178.

[7] Kettunen K. Can type-token ratio be used to show morphological complexity of languages? [J]. Journal of Quantitative Linguistics, 2014,21(3): 223-245.

[8] Juola P. Assessing linguistic complexity[J].Language Complexity: Typology, Contact, Change, 2008: 89-108.

[9] Tweedie F J, Baayen R H. How variable may a constant be? Measures of lexical richness in perspective [J].Computers and the Humanities, 1998,32(5): 323-352.

[10] Hubert P, Labbé D. A model of vocabulary partition [J].Literary and Linguistic Computing,1988, 3(4): 223-225.

[11] McKee G. Measuring vocabulary diversity using dedicated software[J]. Literary and Linguistic Computing, 2000,15(3): 323-338.
[12] Savoy J. Lexical analysis of US political speeches[J]. Journal of Quantitative Linguistics, 2010, 17(2): 123-141.
[13] Arnold E, Labbe D. Vote for me. Don't vote for the other one[J]. Journal of World Languages, 2015, 2(1): 32-49.
[14] Zhang Y. A corpus based analysis of lexical richness of Beijing mandarin speakers: Variable identification and model construction[J]. Language Sciences, 2014, 44: 60-69.

王珊(1982—),博士,助理教授,主要研究领域为语言学、计量语言学、应用语言学。
E-mail: shanwang@um.edu.mo

王会珍(1980—),博士,讲师,主要研究领域为自然语言处理、信息抽取、自动问答。
E-mail: wanghuizhen@mail.neu.edu.cn

第二十届中国计算语言学大会(CCL2021)征稿启事

"第二十届中国计算语言学大会"CCL 2021(The Twentieth China National Conference on Computational Linguistics)将于 2021 年 8 月 13—15 日在呼和浩特市举行,会议由内蒙古大学承办。中国计算语言学大会创办于 1991 年,由中国中文信息学会计算语言学专业委员会负责组织。经过近 30 年的发展,中国计算语言学大会已成为国内自然语言处理领域权威性最高、规模和影响最大的学术会议。作为中国中文信息学会(国内一级学会)的旗舰会议,CCL 聚焦于中国境内各类语言的智能计算和信息处理,为研讨和传播计算语言学最新学术和技术成果提供了最广泛的高层次交流平台。

会议网站:http://cips-cl.org/CCL2021

时间表

- 论文投稿截止日期:2021 年 4 月 1 日
- 录用通知发出日期:2021 年 5 月 15 日
- 论文终版提交日期:2021 年 6 月 1 日

论文投稿

CCL 2021 同时接受中文和英文投稿。组委会将决定被录用的稿件是由口头报告或者海报的形式进行展示。被录用的中文稿件将被推荐至《中文信息学报》、《清华大学学报》(自然科学版)、《中国科学》及其他计算机类中国科技核心期刊(中国科学技术信息研究所制订)。作者必须根据会议和期刊的审稿意见进行相应修改,《中文信息学报》对未完成修改的稿件保留不予发表的权利。《清华大学学报》(自然科学版)、《中国科学》和其他计算机类中国科技核心期刊可能会要求对推荐论文进行再审,通过后方能发表。被录用的英文稿件将由 Springer Lecture Notes in Artificial Intelligence (LNAI)出版。此外,CCL2021 会议论文集将被 CCL Anthology 和 ACL Anthology 收录。