中文词汇增长研究

王珊[1]　王会珍[2]

（1.澳门大学人文学院，澳门；2.东北大学计算机科学与工程学院，沈阳）

摘要：本文选取中国政府工作报告为语料，研究中文的词汇增长。政府工作报告是中国政府对国家建设发展的年度总结，本文选取 1954-2018 年中国政府工作报告为语料，分析文本中词语数量与词语类别的变化曲线，挖掘了政府工作报告的词汇丰富度与国家发展状态的相互关系。对于中国政府工作报告，首先进行了中文分词及对于分词结果的人工审查。根据曲线拟合效果，选择拟合最好的词汇增长模型进行预测。以中国的五年计划作为基础时间周期，对于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，在中国建设全面展开、制度改革、新政策频出的时期，现实值高于预测值，需要更多的新词去描述；而在政策相对稳定的阶段，现实观测值低于预测值，对于原有词的复用较多。据此结果，中国自 1954 年至 2018 年可以分为四个阶段。TTR 领域的研究往往使用英语等具有自然分词特性的语言文本作为研究对象。本文采用了中文语料作为研究对象，补充了 TTR 研究在这一方向的缺失，为中文词汇增长研究提供借鉴。

关键词：词汇增长；词汇丰富度；TTR

A Study of Chinese Vocabulary Growth

Shan Wang[1], and Huizhen Wang[2]

(1. Faculty of Arts and Humanities, University of Macau, Macao, China; 2. School of Computer Science and Engineering, Northeastern University, Shenyang City, Liaoning Province, 110819, China)

Abstract: This paper analyses Chinese vocabulary growth based on *Reports on the Work of the Chinese Government*. The Report is the annual summary of the Chinese government's national construction and development. This paper selects the reports of 1954-2018, analyses the change curve of the number and types of words in the reports, and excavates the relationship between the lexical richness of the reports and the state of national development. For the *Reports on the Work of the Chinese Government*, this study carries out automatic Chinese word segmentation and manual examination of the segmentation results. According to the effect of curve fitting, the best model of vocabulary growth is selected to predict. Taking China's five-year plan as the basic time cycle, the difference between the predicted value of each cycle model and the observed value is analyzed, and the results are compared with those of the randomly scrambled text, which further verifies the experimental results. We find that in the period of China's overall construction, system reform and frequent new policies, the actual value is higher than the predicted value, and thus more new words are needed to describe it. In the period of relatively stable policies, the actual observation value is lower than the predicted value, and thus the original words are more reused. According to this observation, China can be divided into four stages from 1954 to 2018. For the research on TTR, language texts, such as English, which have the characteristics of natural word segmentation, are often used as research objects. This paper takes Chinese data as the research object, supplements the lack of TTR research in this direction, and provides reference for the study of Chinese vocabulary growth.

Key words: Vocabulary growth; Vocabulary richness; TTR

0 引言

在计量语言学领域，文本中词语数量（tokens）与词语种类（types）的关系研究是重要的研究方

向，二者的比值 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 的研究被先后提出，用于分析不同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。

不同的文本其词汇丰富度有差别，TTR 值也有差异。统计不同文章的 TTR 值，可以分析不同文本在词汇丰富度上的区别，有助于判断文本作者[8]。在已知作者身份的情况下，研究其文章内容的 TTR 值，又可以分析作者的语言风格[12]。相同的文本在不同的区域其 TTR 值也有变化，这体现了不同区域文本内容的特点。分析统一文本时，由于词汇数量会影响 TTR 值，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语种类数[2]。其二是使用拟合效果优秀的 TTR 模型如 Heaps 模型[3]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。

当研究的对象是按照时间顺序组合的文本时，不同区域文本的 TTR 值即体现了时代的特征。如有研究选择美国总统两百年来的演讲语料进行研究，分析 TTR 增长与各时期社会特点、总统政策之间的关系[11]。然而，这类研究的对象多为有自然词语划分特点的拉丁语言，目前还缺少利用 TTR 预测模型对中文进行分析的研究。在以中文等部分亚洲语言中，文本是连续而不具有自然划分性质的，这给 TTR 的研究带来了困难。

中国政府工作报告是中国政府对国家建设发展的年度总结，第一次发布时间是 1954 年。为弥补 TTR 研究在中文领域研究的空缺，本文选取了 1954 年至 2018 年的语料作为研究对象，意在证明中文 TTR 分析的可行性。在比较了 Heaps 模型与 Hubert 模型的拟合效果后，采用 Heaps 模型作为研究的预测模型。在不同阶段统计词语种类数与预测词语种类数的比较中，分析了不同阶段该差值与中国政府工作之间的联系，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 是句子中词语种类数量与词语总量的比值，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类别的数量，可以用于衡量词汇的丰富程度。

不同类型的文本，作者不同、语言不同，TTR 值也存在差异，对于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。如有研究对特朗普与希拉里在 2016 年竞选期间的辩论与演讲等语料进行分析，利用 TTR 分析了其语言的风格与修辞特点[12]。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。有的研究对十二位作者五万个词的词类别的数量进行统计分析，证明了词汇丰富度与作者身份的强关联性[4]。除此之外，不同语言的 TTR 值也有不同，有的研究对于 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[6]。

同一种文本，不同部分的词汇丰富度也有不同，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，统计出现的词的类别数。这样计算的 TTR 值被称为 MATTR（移动平均 TTR）。一种基于窗口的、快速计算 MATTR 特征的算法的得到提出[2]，采用此方法分析 *The Adventures of Sherlock Holmes*，发现了文章内容与 MATTR 的关系：MATTR 值在每个故事的开始会上升，而在冗长的对话中表现下降趋势。这说明了 MATTR 值对于分析文本内部的风格同样具有帮助。

词汇数量与词语种类的增长关系可以用数学函数刻画。在词语种类较少时，词语类别数量与文本的长度几乎保持着 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。为了建立这样的关系，指数类型的预测模型被提出[3]（见式 1）。在这个模型中，文本词语类别数量 V 被看作是以词语数量为自变量 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（见式 2）。

$$V' = an^C \quad (1)$$

$$\ln(V') = \ln(a) + C\ln(n) \ with\ 0 < C < 1 \quad (2)$$

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[13]，也难以通过语言学进行解释。对于这一现象，更

加复杂的模型被提出[5]。

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[8]，一种基于常用词与非常用词比例的模型得以提出（见式 3）。

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

在式（3）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)指出现多次的词占用的比例。p 作为模型中唯一的参数，反应了语料中常用词与非常用词的比例。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。

$$V'(1.0) = pV + (1-p)V \quad (4)$$

该模型考虑了词出现的概率分布，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[11]，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[8]。可见在不同的文本环境下，p 有较大的变化，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，一些专用于 TTR 计算的软件也被开发了出来[9]。

这些对于文本特点的研究，往往聚焦于文本使用的词汇类型，最高使用频率的词语，而忽视了更为普遍的规律。有的研究着重于对文本个体的分析，但是忽略了不同文本之间的相互关系。除此之外，部分研究中分析文本的方法过于简单，而不具有说服力。本文选取中国从 1949 年到 2018 年的政府工作报告作为文本材料，使用 Heaps 模型，对于 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、发言往往反应了时代关注的热点，对于政治有着很强的预兆性。此类文本具有权威性、公开性，又蕴藏着珍贵的政治价值，因而常常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，分析了各自的语言特点[10]。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[1]。使用中国政府官员发言，利用 TTR 与语言信息熵分析了发言人词汇丰富度与社会、教育信息的关系[14]。

本研究选择中国政府工作报告作为统计实验的文本。政府工作报告是中国政府的一种公文形式，在每年的人民代表大会会议和政治协商会议中公开发布。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。

自 1954 年第一次发布政府工作报至 2018 年，除 1961-1963，1965-1974，1976-1977 期间受中国文化大革命等因素的影响，政府工作报告有缺失现象，其他年份每年发布一次，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

中文文本中词语的划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”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 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”中获得第一名的成绩，是当今汉语分词最可靠的系统之一。NLPIR 系统拥有“新词发现”功能，从较长的文本内容中，基于信息交叉熵自动发现新词汇，非常适合对词汇类别数量的研究。在使用 NLPIR 系统自动分词后，人工审核并修改了分词结果。

2.3 使用的模型

本文对比了 Heaps 模型与 Hubert 模型，对中国政府工作报告中 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，以均方误差（见式 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^2.857，C=e^0.5137，Hubert 模型中比例参数 p=0.0711。

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

两种模型预测值的总体偏差有差别。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。

[1] http://ictclas.nlpir.org/

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即为：

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

当 $Fw(X=u)=u$ 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。中国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出现的分布是大不相同的，即存在一定第二类的词汇，对于 Hubert 模型的预测结果造成了影响。这解释了实验中，Hubert 模型对于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 年共有政府工作报告 50 篇，其中词语数量为 589000，词语种类为 19230，对应产生了 58990 个二维点。实验中，每 590 个点选取一个作为采样点，一共选取 1000 个点，用于 Heaps 模型与 Hubert 模型的拟合。对于拟合得到的曲线，以及这 1000 个采样点，绘制得到图 1 与图 2。

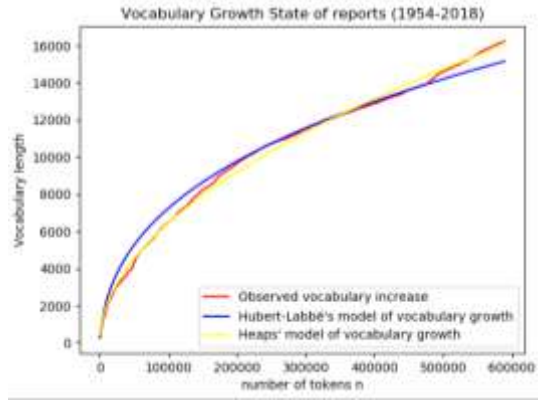

图 1 两种模型预测曲线

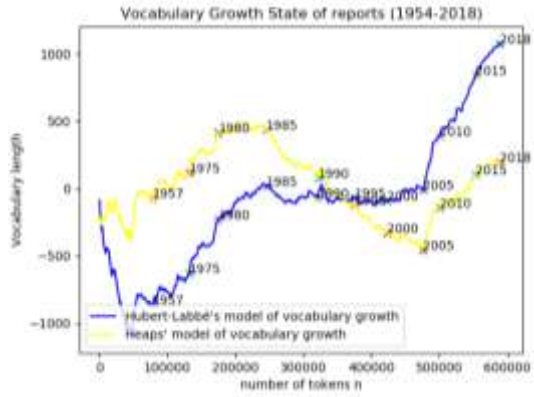

图 2 两种模型预测值与现实观测值之差

图 1 显示了两种预测曲线与现实曲线，表现了现实中与两种模型种词种类与词数量的增长关系。在词语数量较小时，词语类别数量随其迅速增长，而当词语数量较多时，其增长的速度会降低。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词语分布的特点决定的。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。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，对于观测值与 Heaps 模型的预测值进行分段分析，结果如表 1。

表 1 每阶段现实观测值与 Heaps 预测值

时间段	词总（tokens，含重复）	词种（types，不重复）	Heaps 预测值	观测值-预测值	观测值 TTR	预测值 TTR	新词数量	新词类占总此类比例	该阶段政府工作报告数量
1953-1957	78987	5710	5641	69	0.07140	0.07220	\	35.59%	4
1958-1964	129096	7349	7435	-86	0.05760	0.05690	（在上一个阶段没出现的词）1639	10.22%	4
1966-1975	132162	7439	7554	-115	0.05710	0.05620	（在前两个阶段没出现的词）100	0.62%	1
1976-1980	174941	8591	9002	-411	0.04910	0.05140	1152	7.18%	5
1981-1985	245450	10224	10662	-438	0.04334	0.04165	（在前 3 个阶段没出现的词）1633	10.18%	5
1986-1990	321936	11752	11829	-77	0.03674	0.03650	1528	9.52%	5
1991-2000	424164	13541	13207	334	0.03113	0.03192	1789	11.15%	10
2001-2005	476151	14369	13915	454	0.02922	0.03017	828	5.16%	5
2006-2010	500679	14745	14604	141	0.02916	0.02944	376	2.34%	5
2011-2015	553933	15633	15531	-102	0.02822	0.02803	888	5.53%	5
2016-2018	589990	16042	16236	-194	0.02751	0.02719	668	4.10%	3

表 1 对中国政府工作报告，在不同阶段的词数量、词种类、Heaps 模型预测值，阶段新增词数等信息进行了详尽的展示。在中国建国初期，受三年饥荒、文化大革命等事件影响，政府报告在 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代表这一阶段的特征。如 1975 年政府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 5710，模型预测值为 5641，前者较后者多 69 个词类，说明此时有更多的新词汇，这与当时中国所处的历史背景是有关的。1953 年到 1957 年，中国进行了第一次工业化建设，同时，中国对于农业、手工业、工商业进行了社会主义改造，初步建立了社会主义制度。这些巨大的社会变化需要更多的词汇去描述，这些新词汇主要包括如下几类：生产计划提及的具体内容，社会主义改造的形式，如“油菜籽”、“烧碱”“公私合营”、“合作小组”、“改造”等。此外，该时期的政府工作报告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，截至到 1964 年，现实观测值为 7349 词，模型预测值为 7435 词，前者较后者少 86 词，说明该阶段中国的政策相对稳定。历史上，该时期中国进行了人民公社化，大跃进等运动，国家以快速的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，没有提及其他方面如教育、医疗、交通、服务业等的发展在 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 1639 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），中国进入了持续十年的文化大革命，政府工作的中心从经济建设转移到了阶级斗争。该时期仅在 1975 年有政府工作报告，主要对十年期间的文化大革命以及世界局势做出总结，这篇报告的预测结果要大于现实观测值，截止到 1975 年，现实观测值为 7439 词，模型预测值为 7554 词，前者较后者少 115 词。该报告的主要内容为对反革命集团的批评，对国际形势的分析等，而未对于国家建设进行具体地规划，导致了其词汇上的单调性，因而产生了较少的 TTR 值。随后的中国第五个五年计划（1976-1980）中，中国处于从混乱中恢复的时期，此时对于过去十年的建设进行了反思与总结，将工作重心回归到

了经济发展上来。此时政府工作相对稳定。该阶段出现的新词共 1152 种，如“四人帮”、“篡党夺权”、“党羽”等，反应了时代的特点。

在中国第六个五年计划时期（1981-1985），政府工作报告中的词类别数量呈现稳步的趋势，现实观测值与模型预测值差值的扩大趋势明显减弱，5 年中该差值 421（预测值 9495，观测值 9074）变化到 438（预测值 10662，观测值 10224），可以认为是差值的变化为正常浮动。这是因为这段时期中国的制度改革与对外开放逐步加深，市场经济得以承认，一些经济特区也先后开放，人口、教育、外交、能源、交通方面的陈述内容也有增加，出现的新词共 1633 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，导致了该阶段需要更多的词汇去描述新的发展方向。这一阶段词类数量，与模型预测值相比，表现出了迅猛的上升趋势，现实预测值的增长量较预测值多 354 词。此阶段出现新词共 1528 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），是改革开放推进最快的时期，现实观测值保持上升趋势，预测值由 12013 增长到 13207，观测值由 12027 增长到 13541，观测值始终高于预测值，二者差距迅速从 14 词扩大至 334 词。该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。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，就业方式进一步发展，对此，江泽民书记提出了三个代表等思想。此时期的词汇类别数量继续保持稳步上升趋势，共增加新词 1789 种。出现的“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，预测值小于现实观测值，且后者保持稳定的增长趋势预测值由 13318 词增长到 13915 词，观测值由 13689 词增长到 14369 词，二者且差距从 371 词扩大至 454 词。中国加入了世界贸易组织，同时，中共中央总书记胡锦涛提出了科学发展观的战略思想，对于城乡发展，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“信息化”、“数字化”、“青藏铁路”、“西气东输”、“反垄断法”等新词共 828 种，极大的扩展了词汇类别的数量。此阶段词汇类型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，对于新政策的阐述，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，次贷危机引起了国际金融市场的震荡，导致了全球经济发展的停滞与后退。此阶段的预测值保持着下降的趋势，预测值由 13981 词增长到 14313 词，观测值由 14405 词增长到 14559 词，二者差距从 424 词缩小至 246 词，新增词汇仅 376 种，与此阶段中国 GDP 的增长率下降保持一致性。随后的第十二个五年计划与十三个五年计划期间（2011-2018），习近平主席提出了“全面深化改革”，在这段期间中国 GDP 增长率也趋于平缓。此时预测值与观测值较为接近，二者的差值较小，保持在 200 以内。新增词汇 1556 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增测所产生的影响。

从整体上看，中国政府工作报告展示出，自 1954 年至今可以分为四个阶段。第一个阶段为 1954-1965 年，在这段时间内中国进行了社会主义改造并进行了初步的发展建设，该阶段的现实观测值以预测值为中心，随着时代特征而上下波动，二者差值始终在 200 词以内。第二个阶段为 1966 年至 1984 年，该阶段为文化大革命及其后中国经济的恢复期。此阶段中国政府工作报告中对于经济建设具体化、数字化的指标占比较少，因而新词出现较少。与 Heaps 模型的预测值相比，本阶段的现实观测值持续降低，差距最大时较预测值少 465 词。第三个阶段是自 1985 年至 2007 年的增长期，随着改革开放程度的加深，中国在经济、教育、交通、医疗等方面进行了取得了重大的成就，不少新政策提出，如“科学发展观”等，导致了新词的大量递增，使得现实观测值迅速增长，远远超过了模型预测值，最大时观测值较预测值多 454 词。最后一个阶段为 2008 年至 2018 年，现实观测值保持下降趋势，速率相对平缓。受次贷危机影响，中国自 2008 年后 GDP 增长率有所放缓，总体趋于稳定。在此阶段，政府政策相对稳定，与模型预测的趋势相一致，二者差值保持在 200 词以内。此外，不同领导人或不同写作团队对于工作报告的用词可能存在影响，这可能也是导致新词出现的一个原因。

4 验证程序

上文以五年计划作为分析的基本时间单位，分析了词汇类别 Heaps 模型预测值与现实观测值的差距。二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：对于经过分词并人工审核修改后的文本，以词语为单位，随机化地打乱顺序，生成新的文本。新生成的随机文本其词语总量，词语种类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。

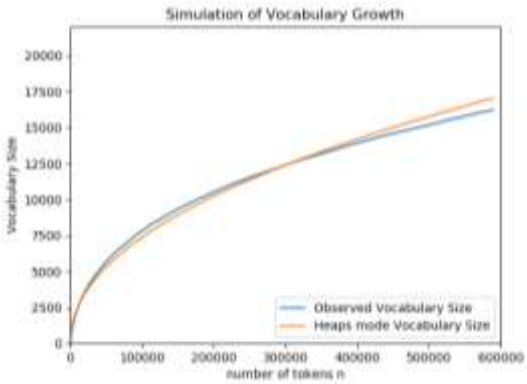

图 3 随机文本预测结果曲线

考虑到每次生成的随机文本可能有偶然性，我们随机生成 1000 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 1000 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 曲线模型如图 3 所示。

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 $a=e^{3.48}$，$C=e^{0.4714}$。通过图 3 可知，此时 Heaps 模型已经较好的拟合了观测值绘制的曲线。为了更好的观测二者差值，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''（见式 7），作为衡量拟合程度的指标。

$$V''(u)=\frac{V(u)-V'(u)}{\sigma V'(u)} \qquad (7)$$

图 4 随机文本预测变化

图 4 反映了乱序后的文本，其标准分数（Z-Score）随词数量增长的关系。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（百分之 99 以上的数据均在-3*σ 到 3*σ 的范围内）。

图 5 表现了 1954-2008 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 到 2）远小于乱序前政府工作报告中的差值（-4 到 5）（图 5）。故而可以说明，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。

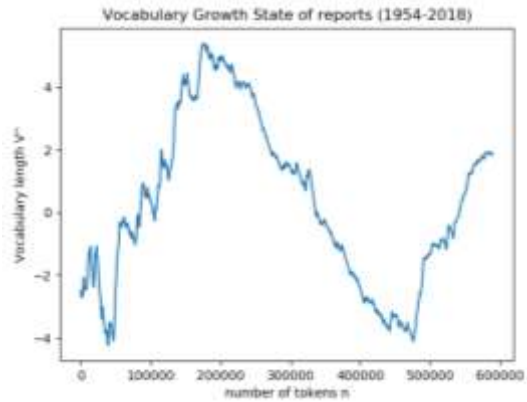

图 5 现实文本预测变化

以上对于中国政府工作报告（1954-2018）的随机化模拟实验证明了词汇增长模型的变化受到了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间差距的变化，是由国家发展重心、所处时代特点等决定的。

5 结论

本文选取从 1954 年至 2018 年中国政府工作报告为语料，根据 Heaps 模型与现实观测值之间的差距，分析了政府工作与此差值的关系，发现了

二者之间的联系：在中国建设全面展开，制度改革频繁，新政策频出的时期，需要更多的词汇去描述工作的具体目标，以及改革的方法制度，此时现实值要高于 Heaps 模型的预测值，而在政策相对稳定的阶段，新增的内容较少，重复原有方针较多，对于旧词的复用频率更高，此时现实观测值要低于 Heaps 的预测结果。

这一结论，也说明了使用 TTR 预测曲线与观测值差值，分析文本词汇丰富度变化的可行性。在第五章中，我们将文本中词的顺序随机打乱，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。

TTR 领域的之前的研究，大多采用英语、法语等使用拉丁字母的印欧语系语言。该类语言本身具有词划分的特性。而在汉藏语系等语言中，文字始终连续无划分，其 TTR 的统计需要进一步的分词处理。本文采用了中文语料作为研究对象，补充了 TTR 研究在这一方向的缺失，证明了中文语言中 TTR 分析的可行性。

本文仅使用 TTR 作为衡量词汇丰富度的指标，只考虑单个词出现的规律。其它词汇丰富度衡量方式，如语言熵，MATTR，Trigram 等，均可用于实验作为后续的研究方向。对于其他需要分词的语言，如日语，韩语等，其语料也可以用于此类研究，以丰富本文研究在不同语言中的结果。

致谢：感谢张林峰同学为本文的数据处理所做的贡献。

参考文献

[1] ARNOLD, E. and LABBE, D., 2015. Vote for me. Don't vote for the other one. *Journal of World Languages 2*, 1, 32-49.

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

[3] HEAPS, H.S., 1978. *Information retrieval, computational and theoretical aspects*. Academic Press, Cambridge.

[4] HOOVER, D.L., 2003. Another perspective on vocabulary richness. *Computers and the Humanities 37*, 2, 151-178.

[5] HUBERT, P. and LABBE, D., 1988. A model of vocabulary partition. *Literary and Linguistic Computing 3*, 4, 223-225.

[6] JUOLA, P., 2008. Assessing linguistic complexity. *Language complexity: Typology, contact, change*, 89-108.

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

[8] LABB¢, C., LABB¢, D., and HUBERT, P., 2004. Automatic segmentation of texts and corpora. *Journal of Quantitative Linguistics 11*, 3, 193-213.

[9] MCKEE, G., 2000. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing 15*, 3, 323-338. DOI= http://dx.doi.org/10.1093/llc/15.3.323.

[10] SAVOY, J., 2010. Lexical analysis of US political speeches. *Journal of Quantitative Linguistics 17*, 2, 123-141.

[11] SAVOY, J., 2015. Vocabulary growth study: an example with the State of the Union addresses. *Journal of Quantitative Linguistics 22*, 4, 289-310.

[12] SAVOY, J., 2017. Trump's and Clinton's Style and Rhetoric during the 2016 Presidential Election. *Journal of Quantitative Linguistics 25*, 2, 168-189.

[13] TWEEDIE, F.J. and BAAYEN, R.H., 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32*, 5, 323-352.

[14] ZHANG, Y., 2014. A corpus based analysis of lexical richness of Beijing Mandarin speakers: variable identification and model construction. *Language Sciences 44*, 60-69.