近十年来澳门的词汇增长[1]

王珊 [1,2]　　陈钊 [3,1]　　张昊迪 [3,4]

[1] 澳门大学人文学院，澳门大学，澳门，999078

[2] 珠海澳大科技研究院，珠海, 519000

[3] 深圳大学计算机与软件学院，深圳，518052

[4] 上海脑科学与类脑研究中心，上海，200031

shanwang@um.edu.mo, hdzhang@szu.edu.cn

摘要

词汇增长模型可以通过拟合词种（types）与词例（tokens）之间的数量关系，反映某一领域词汇的历时演化。澳门作为多语言多文化融合之地，词汇的使用情况能够反映社会的关注焦点，但目前尚无对澳门历时词汇演变的研究。本文首次构建澳门汉语历时语料库，利用三大词汇增长模型拟合语料库的词汇变化，并选取效果最好的 Heaps 模型进一步分析词汇演变与报刊内容的关系，结果反映出澳门词汇的变化趋势与热点新闻、澳门施政方针和民生密切相关。本研究还采用去除文本时序信息后的乱序文本，验证了方法的有效性。本文是首项基于大规模历时语料库考察澳门词汇演变的研究，对深入了解澳门语言生活的发展具有重要意义。

关键词：澳门；词汇增长；历时变化；语料库

Macau's Vocabulary Growth in the Recent Ten Year

Shan Wang[1,2]　　Zhao Chen[3,1]　　Haodi Zhang[3,4]

[1]Faculty of Arts and Humanities, University of Macau, Macau

[2]Zhuhai UM Science & Technology Research Institute, Zhuhai, 519000

[3]School of Computer and Software, Shenzhen University, Shenzhen, 518052

[4]Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, 200031

shanwang@um.edu.mo, hdzhang@szu.edu.cn

Abstract

[1] 基金项目：本研究受教育部国家语委“十三五”科研规划项目资助，项目编号：GP-023-2020。

第二十届中国计算语言学大会论文集，第350页-第360页，呼和浩特，中国，2021年8月13日至15日。

Vocabulary growth models can reflect the diachronic change of vocabulary in a certain field by fitting the quantitative relationship between word types and tokens. As a place of multi-language and multi-cultural integration, Macau's vocabulary use can reflect the focus of society, but there is no research on Macau's diachronic vocabulary growth. This paper constructed a diachronic corpus of Macau Chinese for the first time, used three vocabulary growth models to fit the vocabulary changes in the corpus, and selected the Heaps model with the best effect to further analyze the relationship between vocabulary change and newspaper content. The results reflect that the changing trend of Macau vocabulary is closely related to hot news, Macau's policy guidelines and people's livelihood. This research also uses the out-of-order text after removing the text timing information to verify the effectiveness of the method. This is the first study to investigate the evolution of Macau vocabulary based on a large-scale diachronic corpus, which is of great significance for the in-depth understanding of the development of Macau's language life.

Keywords: Macau, vocabulary growth, diachronic change, corpus

1.引言

计量语言学有许多关于词种（Types）与词例（Tokens）之间关系的研究，二者之间的比例被称为 Type-Token-Ratio（TTR），通常用于衡量文章中词汇的丰富度，如对于小说而言，不同作者往往有不同的语言风格。TTR 值，可以用作判断文本作者的指标（Labbé et al., 2004）。但是 TTR 只能表示某种状态下词种与词例的关系，不能进行历时发展的分析，而词汇增长模型能够通过函数拟合词种和词例之间的增长关系，从而拟合到 TTR 的变化情况。目前主流词汇增长模型有三种：Guiraud 模型（Guiraud, 1954），Heaps 模型（Heaps, 1978），Hubert 模型（Hubert & Labbe, 1988），它们能够通过词例的数量预测词种的数量，可用于研究以时间为顺序的语料。

澳门是中西文化交汇之地，实施"三文四语"的政策，但是目前尚无对澳门历时词汇演变的研究。新闻一定程度上可以客观公正地反映社会现象及社会发展趋势，对历时新闻语料进行词汇增长模型构建，能够通过模型的预测反映出社会的发展，对以澳门为代表的多语言多文化融汇之地进行历时词汇演变研究具有重要的意义。本文的研究问题包括：（1）如何构建具有代表性且体现历时特点的澳门汉语语料库？（2）如何对语料库进行词汇增长模型建模并分析其反映的词汇变化？（3）如何验证词汇增长模型对澳门词汇历时发展的有效性？

本文的主要结构如下：第二部分是与词汇增长和澳门词汇的研究，第三部分是历时澳门汉语语料库的构建，第四部分词汇增长实验，第五部分是对澳门汉语的词汇增长的分析，第六部分采用验证程序检验词汇增长模型的有效性，第七部分的本文的结论。

2．相关工作

2.1 词汇增长研究

TTR 能够反映词种（Types）与词例（Tokens）之间关系，能够用于判定作者身份，了解语言掌握情况等，如 Hoover（2003）对十二位作者的作品词种数量进行统计，发现 TTR 可以用于判断作者的身份。Yu（2010)指出 TTR 与写作和口语的质量在统计学上具有显著的正相关性，证明语言能力较高的人其 TTR 值也会相对较高，TTR 是考察学习者对一门语言掌握情况的指标。Mellor（2010）发现 TTR 也经常用于衡量说话者与写作者的语言程度，语言程度越高则使用的低频词较多。X. Wang（2014）分析了英语二语学习者的电子邮件文章的 TTR 值与写作熟练度之间的关系。

TTR 值只对于某一固定文本内部的词汇特征分析很有效，但由于它只能表示某一个时间点词例与词种之间的比值关系且词种数量的增加一定会使得 TTR 值下降，这种特性 使得 TTR 值无法用于历时分析。而词汇增长模型能够利用数学函数表示词例与词种数量之间的关系，且本质上词汇增长模型是学习词种与词例之间的数量增长关系，其学习后的预测的词种与词例的比值与 TTR 值原理一致，能够反映词汇丰富度，如 Savoy（2015）分析了 1790 年至 2014 年间美国国情咨文演讲词汇量增长情况，语料涵盖了 42 位美国总统的 225 篇演讲。结果显示，在白宫面临反复出现的问题的

情况下，美国总统更倾向使用重复的论点或解释；而在面对出乎意料的情况或提出新的解决方案的时候，词汇增长的预测值与观测值的差异变大。王珊、王会珍（2021）通过分析中国 1954-2018 年的政府工作报告，分析词汇增长与政策之间的关系，验证了在中文数据集上运用该方法的可行性。可见，词汇增长模型对于演讲、政府工作报告乃至英语语言中的词汇历时变化研究非常有效。不过，目前尚无对澳门报刊的词汇增长的研究。

2.2 澳门词汇研究

华语是以普通话为基础的全世界华人的共同语言（李宇明, 2017）。华语有多个变体，包括大陆普通话、台湾国语、港澳华语、新马印尼文莱华语等,另外北美华语正在形成,欧洲华语略有雏形。不同地区的华语有不同的特点，其中面向港澳与内地的华语研究较多。澳门是中西文化交汇之地，实施三文四语的政策，目前有不少关于澳门语言使用的研究。邵朝阳（1999）通过实地收集语料并研究澳门赌博隐语的使用情况；黄翊（2005）的研究分析澳门清代中文文档中具有澳门地方色彩的词语；汤志祥（2008）认为香港与澳门两地的华语具有高度的共通性，但也存在一些词汇差异；袁伟（2015）讨论澳门中文平面媒体的字母词规范；姚双云、黄翊（2014）对比澳门与内地新闻词汇的差异；S. Wang and Luo（2019）利用自建语料库分析与澳门旅游相关的词汇。

然而，总体上目前对澳门社会词汇使用的研究多为例举式，尚无历时词汇增长与社会热点关系的分析。近年来澳门发展十分迅速，对以澳门为代表的多语言多文化融汇之地进行不同时期的词汇增长研究，不但能够验证词汇增长模型在新闻领域的有效性，还有助于弥补澳门词汇演变研究的空缺。

3 构建历时澳门汉语语料库

3.1 语料选取

新闻在一定程度上有客观性、公正性与真实性，可以反映社会热点与重大事件。Franklin（Franklin, 2008）指出报纸内容不仅会关注具有新闻价值的人物、公司、事件和社会现象，还会对其进行深度分析。如（Johnson & Bevitori, 2017）使用词汇增长对美国和英国的特定领域的若干报纸进行分析，探讨气候变化对人类迁移的影响是如何在报道中体现的。

本研究考察了澳门的各大报纸，包括《澳门日报》、《市民日报》、《澳门法治日报》、《体育周报》等,其中《澳门日报》、《市民日报》分别提供 2 年和 1 年内的数据，因数据量太小无法做历时分析，而《澳门法治日报》和《体育周报》偏向法律与体育方面的专门领域报道，不能全方位展示社会热点的变化，因此本文选择了综合性报纸《澳门时报》。《澳门时报》原名为《时事新闻报》，创刊于 1972 年，在 2015 年 9 月改名为《澳门时报》，并且在 2016 年 6 月 15 号从周刊改为日刊。《澳门时报》以客观、负责及专业的态度报道中国及世界时政要闻与民生大事，贴近社情民意、关注民生。该报的网站提供了 10 余年的语料，本文获取了 2011 年 1 月 1 日到 2020 年 12 月 31 日的全部电子报刊语料，内容覆盖法律、体育、民生与国际热点等领域。

3.2 语料预处理

由于中文文本的词与词之间不存在分隔符，开展词汇增长研究需要对文本进行分词。本文选取 pkuseg 分词工具包[2]，它由北京大学语言计算机与机器学习组研制（Luo et al., 2019），其性能与 jieba、 THULAC 等分词工具包相比，在细分领域和跨领域测试的表现上是效果最好的[2]。pkuseg 支持不同领域的个性化分词工具。根据分词的领域不同，可以选择不同的分词模型，目前 pkuseg 支持新闻领域、网络领域、医药领域、旅游领域以及混合领域的分词模型。因本研究的语料选自《澳门时报》，属于新闻领域，故本文采用该工具的新闻领域分词模型进行分词。

pkuseg 分词工具包仅支持简体中文分词而《澳门时报》的语料库是繁体中文字体，本文使用 OpenCC 工具包[3]先将原文转化为简体中文。本文利用 pkusge 进行分词，然后再根据切分的位置，对应到《澳门时报》的繁体版，获得对原文的分词结果。

[2] https://github.com/lancopku/pkuseg-python
[3] https://github.com/BYVoid/OpenCC

此外，我们还去除了带阿拉伯数字的词语（如带“百”、“千”、“萬”、“年”、“月”和“日”以及标点符号，例如本文仅删除“1 萬”“1996 年”这类词语，而不删除“千錘百鍊”这样的词语。在词与短语的界限上，汉语学界存在较大的争议，董秀芳（2004）把多音节的结构归为复合词，如“计算机病毒”、“国家安全系统”、“社会政治经济学”等，因此本研究也将这样的结构，如“半導體工廠”、“新冠病毒核酸檢測”等视为词。

3.3 《澳门时报》语料库的信息

表.1 展示了《澳门时报》语料库的统计信息，包括年份、文章数、词例数、词种数和字数等统计信息。该语料库共包含 2011-2020 年 10 年的语料，共收集 47856 篇文章，高达 22096302 字，9711067 词例（tokens）和 684553 词种（types）。

年份	文章数	词例数	词种数	字数
2011	1205	360657	30972	793163
2012	1222	365734	32762	806047
2013	1482	444130	36152	982346
2014	1660	362461	36533	805947
2015	2148	437260	45121	982622
2016	8482	1445242	108867	3323143
2017	11801	2043629	135444	4683304
2018	7152	1475069	97418	3378961
2019	6066	1306562	80320	3016205
2020	6638	1470323	80964	3324564
总计	47856	9711067	684553	22096302

表.1 《澳门时报》语料库(2010-2020)概况

4. 词汇增长模型实验

4.1 三种词汇增长模型

词汇增长模型假设词例与词种之间存在着某种函数关系，通过拟合出来的模型，预测在不同词例下的词种的数量，通过分别分析 TTR 与词种的预测值与观测值之间的差值，反映词汇丰富度的变化，分析其特点。目前主流的词汇增长模型主要有三个：Guiraud 模型（Guiraud, 1954），Heaps 模型（Heaps, 1978），Hubert 模型（Hubert & Labbe, 1988）。

Guiraud 模型对词种与词例数量的比值关系进行建模，认为词种数量 V 与词例数量 N 的平方根比值为常数。根据这个关系得出预测词种数量V'与当前词例数量n之间的关系表达式为：

$$V'(n) = c \cdot \sqrt{n} \tag{1}$$

Heaps 是指数预测模型，其表达式如下：

$$\boldsymbol{V' = an^C} \tag{2}$$

$$ln(V') = ln(a) + Cln(n) \quad 0 < C < 1 \tag{3}$$

其中V'代表词种的预测值，而 a,C 为模型的参数，n 为词例的数量。Heaps 可以很好的拟合词例与词种之间的关系，并且 Tweedie and Baayen（1998）的研究进一步表明，其 a,C 这两个参数与词例的多少相关。

Hubert 提出了更复杂的模型，Hubert 认为词种应该有两种类型——常用词与非常用词。非常用词在文本中往往很少出现重复，使得它的词种与词例之间的比值为线型关系，而常用词则代表哪些词例数量远远大于词种数量的词。由此，Labbé et al.（2004）假设当一个词出现时，它是非常用词的概率为 p，常用词的概率为 1-p。其模型公式如下：

$$V'(u) = puV + (1-p)\left[V - \sum_{i=1}^{k}[V_i(1-u)^i]\right] \tag{4}$$

其中i是词语出现的次数，V_i是出现频次为 i 的词的数量。p 是指语料中只指词例为 1 的词语所占的比例。而 u 是预测预料占总语料的比例，即一篇文章中需要预测的词例数除以文章总词例数。$(1-u)^i$表示预测文本相对词数量和词在文本中出现次数与减少的词种数量的关系。其中 u 为预测文本词例数量与语料库文本词例数量之比，i 代表着到达该词例数量的预料库中出现 i 次的词种数量。Hubert 模型假设词语在文本中出现的次数与文本的长度相关，并且其增长速率与词语在文本中出现的次数具有正相关关系，即出现次数越多的词语，其增长的速率越快。

对比三个模型，其中 Guiraud 模型太过于简单，表现效果也是最差的模型，由于根号函数的性质，其增长率会快速下降。Hubert 模型（公式 4）由两部分组成，第一部分为非常用词的预测，第二部分为常用词预测。其中非常用词预测采用的是线性增长函数(puV),常用词预测是指数增长函数。在 u 较小的时候，相比指数部分，线性部分的增长速率较慢；在 u 较大的时候，相比指数部分，线性部分的增长速率较快。所以相对于 heaps 函数，在 p 较大的情况下，可能会出现前期预测值较小，后期预测值较大的情况。据 Petersen et al.（2012）研究发现，当词种的数量越大的时候，其增长速率越小，而 Hubert 中非常用词部分的增长部分是恒定的，与此结论相反且 p 值为恒定值，使得训练出来的 Hubert 模型可能存在预测值与观测值在 h 在前期增长速率过慢，而后期增长速率过快。Guiraud 模型是根据词种数量与词例数量的平方根之间的比值为某个常数 r 的假设建模的。Heaps 模型则是假设在双对数空间中与词种与词例的数量的比值存在某个常数 a。

本文使用 Guiraud、Heaps 与 Hubert 三个模型对《澳门时报》语料库进行拟合，使用 scipy[4]框架，并用非线性最小二乘法的方法，对数据所得出的残差平方和进行优化（见公式 5）,该函数用于衡量模型预测值与观测值的误差，从而优化模型。其中，Guiraud 模型的参数 r=98.028625；Heaps 模型的参数 a=2.785034，C= 0.727058；Hubert 模型中的比例参数为 p=0.316723。本文将会在下面的实验效果中选择效果最好的算法以作误差分析。

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

4.2 《澳门时报》的词汇增长

2011 年 1 月 1 日-2020 年 12 月 31 日的《澳门时报》电子报刊语料共含总词例数 9711067。本次实验将每 1000 个词例采样一次，共得到 9711 个样本用于模型的拟合。图.*1* 展示了三个模型的预测值与语料库的 Token-Type 曲线（这里不是 TTR，而是以 Token 为 x 轴，type 为 y 轴的曲线）。可见三个模型的预测值在前期都比观测值高，而到中期比观测值低，到后期又反超观测值。其中可以看出对于《澳门时报》来说其词种与词例的数量关系接近指数函数，到后期虽然有变缓的趋势，但总体变化不大。

图.*2* 展示的是观测值与预测值的差，其值越接近 0，预测越准确。Guiraud 模型预测整体效果最差，且方差大，其平均值达到-7414.67。从 Hubert 的曲线可以看出 2019-2020 数据 Hubert 预测较

[4] https://www.scipy.org/

准确，但总体效果表现不好，其差值平均值为 1729.56。Heaps 模型预测值与观测值相差较小，p 平均值为-526.22。从三条曲线中可以看出 Heap 模型的预测效果比 Hubert 模型好，而 Guiraud 模型的预测效果最差，故下文采用 Heaps 模型的预测结果以作分析。

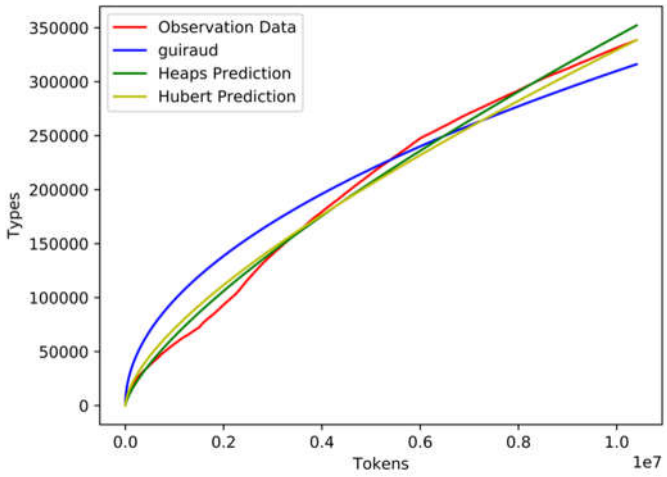

图.1 不同模型词种数量预测曲线

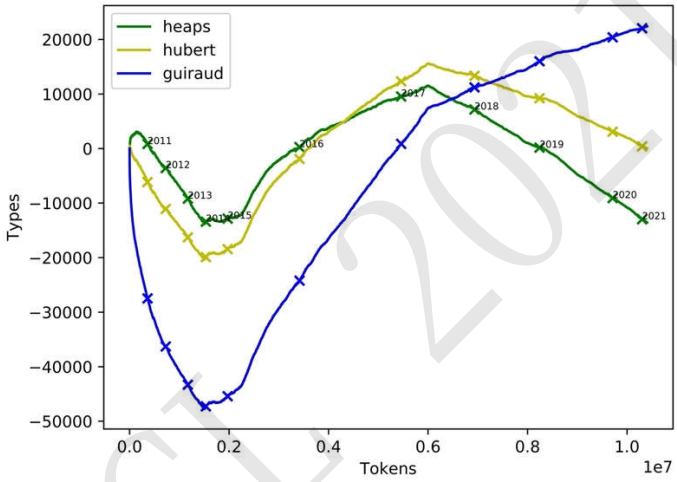

图.2 不同模型预测误差分布

5. 澳门汉语的词汇增长分析

表.2 展示了《澳门时报》不同年份的数据情况，包括词例、观测值（语料库词种数）、Heaps 预测值、观测值 TTR、预测值 TTR 和预测值的误差、新词语数量、文章数等统计信息。

年份	词例	观测值（词种数）	Heaps 预测值	预测值-观测值	观测值 TTR	预测值 TTR	新词语数量	文章数
2011	360657	30972	30560	-412	0.086	0.085	-	1205
2012	726391	47370	50844	3474	0.065	0.070	16398	1222

2013	1170521	62510	71927	9417	0.053	0.061	15140	1482
2014	1532982	76015	87513	11498	0.050	0.057	13505	1660
2015	1970242	93893	105029	11136	0.048	0.053	17878	2148
2016	3415484	159417	156685	-2732	0.047	0.046	65524	8482
2017	5459113	231432	220347	-11085	0.042	0.040	72015	11801
2018	6934182	270122	262198	-7924	0.039	0.038	38690	7152
2019	8240744	298395	297261	-1134	0.036	0.036	28273	6066
2020	9711067	326466	334948	8482	0.034	0.034	28071	6638

表.2 每年观测值与 Heaps 预测值

本节分析模型预测值与《澳门时报》的实际观测值的关系。2011 年 Heaps 模型的预测值略小于观测值，观测值为 30972, 模型预测值为 30560，前者较后者多 412，说明该阶段出现了较多的新词。这是由 2011 的许多热点事件导致的，如“核电站”、“珠海北站”与“本拉登”等词语反映了日本福岛核电站泄漏、珠海北站投入使用、本拉登之死等事件。此外，澳门政府提出要推动区域合作发展，实现经济多元化的政策，将澳门定位为旅游城市，大力发展澳门本地旅游资源。此外，澳门不断拓展自身发展空间，提高区域合作的层次。在这一背景下，出现了许多有代表性的词，如 “控煙”、“區域性”、“國際休閒中心”、“一程多站”和“粵澳合作園區”等词汇。

2012 年预测值大于观测值，观测值为 47370，模型预测值为 50844，前者比后者少 3474，反映出该阶段的词汇使用相对稳定。这一年《澳门时报》报道的世界与本地热点的变化性相对较少，主要集中在民生方面，例如“物業”、“醫療”、“市场”等非新词的词频大幅提升。不过该阶段出现了 16398 个新词，反映了当地关注的新闻热点，如澳门政府增加了对历史文化的重视程度，其中《文化遺產保護法》進入立法程序，诞生了“文遗法”、“物质文遗”与“文化遗产”等新词。而 2012 年，神州九号载人飞船上天，出现了“神舟”、“载人航天代表团”与“航天員”等与航天相关的新词。澳门大西洋银行及中国银行发行龙钞，出现大量“發鈔”、“新鈔”与“龍鈔”等新词。澳门推进 3g 信号的普及率，使得 “2g”、“3g”与“3g 智能機”这类与通讯相关的名词频率大大提高。

2013 年预测值大于观测值，观测值为 62510，预测值为 71927，前者比后者少了 9417，新增词种 15140 个。该年度是科技大发展的一年，该阶段出现了许多与科技相关的新词。这期间发生了几件大事，包括辽宁舰第一次远赴南海执行任务、神州十号发射成功、第二代北斗卫星开始为亚太地区用户提供区域定位服务等，由此产生了一批与之相关的新词，如“北斗”、“遼寧艦”、“軍艦”、“探月”等。而 2013 年的立法会选举，使“選舉法”与“選舉日”等与选举和立法有关的词的使用频数上升。澳门开始推行 4g 信号，使得“4g”、“信號”、等词的词频提升。虽然这期间发生了许多新闻热点，但由于澳门政策的稳定，且热点新闻减少，更多的是报道关于民生的问题，使得 heaps 观测值的增长速率比预测的增长速率低，且观测值与预测值的差距相较于上一年提升了 5943。

2014 年观测值比预测值小，观测值为 76015，预测值为 87513，新增词种 13505 个，其主要报道依然关注于民生问题，例如治安、青少年及民生福利等相关问题。该年度是澳门回归祖国十五周年，诞生了“爱国人士”等之前的阶段没出现的与爱国相关的新词。而这段时间也发生了许多其他事件，如埃博拉病毒的进一步失控，出现了“埃博拉”、“隔离带”等与传染病相关的词语。澳门 A 区填海工程的相关报道中出现了“轻轨半岛”等新词。这一年澳门政府为保障本地人培养长效机制，提出

了一系列计划，出现了“精英人才團隊”、“人才發展委員會”等新词；而 2014 年的马航事件也使得《澳门时报》上与航天和空难相关的词汇出现，例如：“空難”、“尾翼”、“雷達應答機”、“马航”等。

2015 年词种预测值比观测值大，观测值为 93893，预测值为 105029，新增词种 17878 个。这一年澳门博彩业收入降至四年最低，而澳门此前推崇的经济多元化优势至此体现出来，出现了很多其他行业的词语，如“鉅記手信店”、“紡織業”、“連鎖書店”、“洗衣店”、“優步”、“微商”等。2015 年是抗日战争胜利 70 周年，北京天安门广场举行大型阅兵仪式以纪念二战的胜利，由此出现了一些之前的阶段未出现的新词，如“慰安妇”、“阅兵式”与“南京军事法庭”等。支付宝在这一年进入澳门市场，澳门出现了许多网购平台，新增了许多与此相关的新词，如“网购”、“亚马逊”、“淘宝”与“顺丰”等。

2016 年测值结果开始小于观测值，预测值为 156685，而观测值为 159417，两者相差 2732,新增词种 65524 个。2016 年该报对国外一些体育联盟的报道较多，出现了较多体育相关的新词，如“湖人”、“西甲”、“投篮”、“拜仁”与“皇马”等。2016 年澳门三次在市场发现禽流感病毒，该报也做了相关报道，出现“横琴檢驗檢疫局”、“傳染病中心”与“病毒”等相关的词语。第十一届中国国际航天博览会上歼-20 首次进行飞行展示，“殲 20”、“载弹量”与“超音速”等词进入人们的眼帘。总体上，这一年新词大爆发的点主要集中在体育方面，《澳门时报》对于 NBA、西甲这类体育赛事相关的词汇增长是预测值大于观测值的一个主要原因。

2017 年预测值小于观测值，预测值为 220347，观测值为 231432，两者相差 11085。导致两者之间出现这么大差距的原因可能是因为澳门在该年度出现了更多新政策，其中最主要的是智慧城市的建设，使用大数据与人工智能等技术有效地改善城市管理，并且提供一系列与之相关的服务，一些新词反映了这些现象，如：“智慧城市联盟工作组”、“智慧城市委员会”、“澳门云计算中心”等与智慧城市相关的词语。此外，2017 年澳门电话诈骗数量增长显著，出现了“电骗”、“电号”与“电人”等相关的新词。而 2017 年开始，中国出现了新的四大发明的概念化，出现了许多相关的词汇，如：“共享化”、“摩拜”、“輕軌博物館”等与新四大发明相关的词汇。大量与之相关的新词出现，导致 TTR 上升的速率加快。

2018 年预测值小于观测值，预测值为 262198，观测值为 270122，两者相差 7924。这一年出现的新词一共 38690 个，多数与美食、应急机制和人工智能相关，包括“本澳防災避險中心”、 “智慧警務所”、“美食之都小組”、“澳門美食網路”、“美食旅遊團”、“阿里雲計算有限公司”、“港澳大灣區人工智能聯盟”、“人工智能時代”等。它们反映出如下热点事件：第一，由于受 2017 年飓风灾害的影响，澳门遭受巨大损失，澳门政策新增“完善应急机制，强化公共安全”一项，切实提高防灾减灾的能力和水平。第二，澳门全力建设“创意城市美食之都”，并于 2018 年 1 月 17 日开启“澳门美食年”项目，推广澳门特色美食，从而推动旅游业的发展。第三，智慧城市进一步发展，使澳门注重发展人工智能、大数据与云计算技术。

2019 年预测值小于观测值，观测值为 298395，预测值为 297261，两者相差 1134。其预测值开始与观测值之间的误差缩小，反映出新闻报道开始趋于稳定的状态。这段时间新增词语 28273 个。2019 年是科技发展大年，澳门电讯拨通澳门首个 5G 通话、华为鸿蒙操作系统、美国制裁华为等事件使得与此相关的新增词语出现。2019 年 4 月，三大人工智能国际会议在澳门举行，使得澳门人工智能产业的发展更进一步，出现了很多与人工智能相关的应用，从出现的新词中能够体现这一趋势，例如“鸿蒙” 、“美國科技戰” 、“微電子研發中心” 、“半導體工廠”、“人工智能科學”、“人工智能控煙”、“人工智能門診”等，这段时间虽然出现了很多与科技相关的新名词，但是其澳门政策开始趋于稳定，与前两年的政策一致，所以使得观测值与预测值的误差开始变小。

2020 年预测值大于观测值，预测值为 334948，观测值为 326466，相差 28273。这一年爆发的新冠病毒肺炎席卷全球，世界各地的经济都跌落至低谷。为了应对这种情况，中国发挥了互联网的优势及强大的动员能力使得中国能够快速地度过危机，恢复经济。这一年的新词主要集中在新冠方面，例如“新冠病毒核酸檢測”、“綠碼”、“新冠病毒”、“粵澳健康碼”、“澳門應急醫療隊”、“緊急防疫指揮中心”等与新冠相关的词。但是由于新增的词种都主要集中于新冠病毒相关的词，虽然词种的数量增加 28071 个词，与此相关的一些与医疗相关但并非新增词的增长速度比新词的速度更快，导致 TTR 存在一定的波动。

6. 词汇增长的验证程序

本文第 5 部分以每一年的语料为基本分析单位，通过定性分析不同年份的新增词语与相关的热点来分析预测值与观测值产生差异的原因。但是二者的差距是由新闻的内容造成的，还是因为模型拟合的误差所造成的，需要进一步验证。本部分通过随机化的方法进行定量验证，若是由新闻内容造成误差的增大，则随机化实验的实验结果误差应当有所减少（Savoy, 2015）。验证方法如下：随机打乱整个语料库内所有词语的顺序，生成新的文本。新生成文本的词例总数、词种总数与原语料库相同，仅使其语序打乱，使得打乱后的语料不再具备原文的语义信息及 TTR 值。若证明随机打乱后的观测值与预测值的差值比原语料库小，而仅有的变量为文本原有的时序语义信息，即可证明新闻内容的实时性是造成观测值与模型预测值误差的主要原因。为了更好地观察预测值与观测值的差值，我们使用模型拟合后的结果计算预测值与观测值差的标准分数（Z-Score, 见式 6）作为衡量拟合效果的指标。

$$diff(n) = V(n) - V'(n) \tag{6}$$

$$Z_score = \frac{diff(n) - \mu_{diff}}{\sigma_{diff}} \tag{7}$$

图.3 展示了随机文本的增长曲线与模型拟合后的预测曲线，其训练的 Heaps 模型参数为 a=35.708521, C=0.566868。从图.3 中可以看出 Heaps 拟合的曲线几乎与随机文本的增长曲线重合，其平均值为-52.08，远远小于现实文本的拟合误差。

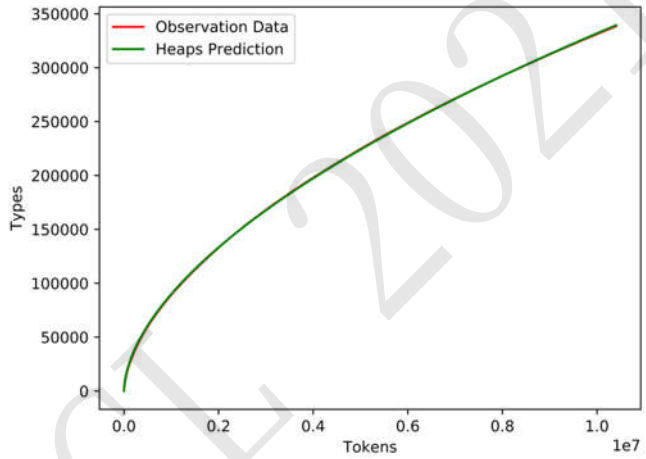

图.3 随机文本增长与预测曲线

图.4 与图.5 分别展示了现实文本数据与随机文本数据的预测误差 Z-Score 分布，这两个实验的 Z-Score 中的大部分数据都处于$(-3\sigma, 3\sigma)$的范围内，即可以证明这两个实验的误差是可接受的。图.3 展现出来的随机本文的预测误差远小于现实文本的实验误差，随机文本的实验误差显示了一定的随机性，且现实文本的误差符合定性分析实验的分析结果，因此可以证明现实文本中 Heaps 拟合效果的误差主要来源于文本的内容。

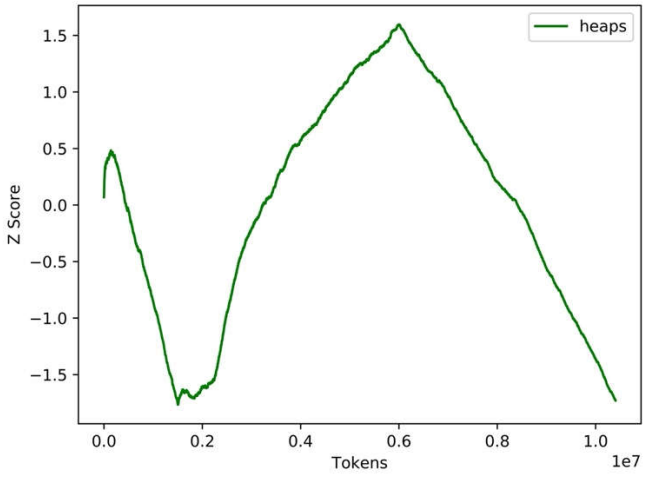

图.4 现实文本的预测误差 Z-Score 分布

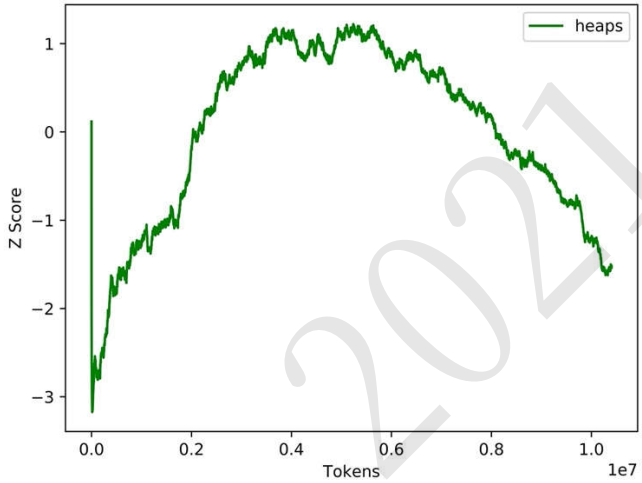

图.5 随机文本的预测误差 Z-Score 分布

7. 结论

澳门作为多语言多文化的社会，具有中西交融的特色，澳门的华语词汇使用颇受关注，但是目前对澳门词汇的研究集中在例举式共时分析，缺少利用大规模语料库进行的历时词汇增长考察。词汇增长模型能够通过词例的数量预测词种数量在不同时期的变化，通过预测值与实际观测值之间的差异，展示词汇在不同时期的演变。首先，本文以《澳门时报》10 余年的报刊语料首次构建澳门汉语历时语料库，语料库的规模高达 22096302 字，9711067 词例和 684553 词种，用探索近 10 年来澳门的词汇演变。其次，本文采用三种词汇增长模型 Guiraud、Heaps 和 Hubert 建模 ,发现 Heaps 模型的拟合效果最好。根据 Heaps 模型的词种预测值与语料库词种观测值之间的差值，对语料进行深入分析后发现：在政策稳定及世界热点新闻的领域相对稳定的情况下，词种的 Heaps 模型预测值高于观测值，反映出词语重复使用率较高，新增词语较少，TTR 值较低；当热点频出或新政推出时，观测值高于预测值，新增词汇较多。再次，本文将语料库中文本的顺序随机打乱，使语料不再具备原文的语义信息，使用 Heaps 模型对乱序文本进行拟合，通过分析原始语料库的标准分数与随机文本的标准分数的对比，进一步说明了模型预测的误差与文本内容之间具有相关性，同时也首次证明了 Heaps 模型用于新闻语料中进行历时分析的可行性。

本研究是首项对澳门词汇的历时演变研究，反映出澳门语言生活不但关注本地民生，也重视国家大事以及国际时事。澳门作为中西交汇的多语社会, 具有丰富的特色词汇以及国际视野，对其进行历时研究能更深入地了解澳门的热点变化与社会发展。在未来研究中，我们将进一步扩大领域，

如对澳门施政报告、司法领域、旅游业等的词汇增长的考察，澳门与香港、内地的词汇变化的关联性等，更充分地反映澳门语言使用现状、演变趋势、社会的变迁和与其他华语区的互动关系。

参考文献

董秀芳. (2004). *汉语的词库与词法*: 汉语的词库与词法. 北京：北京大学出版社.

黄翊. (2005). 清代中文档案中的澳门汉语词汇. *华东师范大学学报(哲学社会科学版), 37*(3), 56-56.

李宇明. (2017). 大华语:全球华人的共同语. *语言文字应用*(01), 2-13.

邵朝阳. (1999). 澳门博彩隐语研究. *中国语文, 4*.

汤志祥. (2008). 论“港澳词语”以及“澳门特有词语”. *江苏大学学报(社会科学版)*(05), 24-29.

王珊、王会珍. (2021). 中文词汇增长研究. *中文信息学报, 35*(1), 17-24.

姚双云、黄翊. (2014). 澳门与内地新闻语篇词汇差异的计量研究. *语言文字应用*(02), 27-37.

袁伟. (2015). 中国澳门特区中文平面媒体中字母词的规范研究. *语言文字应用, No.95*(03), 68-75.

Franklin, B. (2008). The Future of Newspapers. *Journalism Practice, 2*(3), 306-317.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*: Presses universitaires de France.

Heaps, H. S. (1978). Information retrieval, computational and theoretical aspects. New York: Academic Press.

Hubert, P., & Labbe, D. (1988). A Model of Vocabulary Partition. *Literary and Linguistic Computing, 3*(4), 223-225.

Johnson, J., & Bevitori, C. (2017). Human Mobility and Climate Change at the Crossroad. A Diachronic Corpus-Assisted Discourse Analysis of the Nexus in UK and US Newspaper Discourse. In K. E. Russo & R. Wodak (Eds.), *The Representation of "Exceptional Migrants" in Media Discourse. The Case of Climate-induced Migration* (Vol. 21, pp. 7-22).

Labbé, C., Labbé, D., & Hubert, P. (2004). Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics, 11*(3), 193-213.

Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. (2019). PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation.

Mellor, A. (2010). *Automatic essay scoring for low level learners of English as a second language.*(Ph.D.). Swansea University, Sketty.

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports, 2*(1), 943.

Savoy, J. (2015). Vocabulary Growth Study: An Example with the State of the Union Addresses. *Journal of Quantitative Linguistics, 22*(4), 289-310.

Tweedie, F. J., & Baayen, R. H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities, 32*(5), 323-352.

Wang, S., & Luo, H. (2019). A Corpus-based Study of the Vocabulary of Macao Tourism Chinese. In I. H. T. H.-J. H. C.(Eds.)(Ed.), *Chinese for Specific and Professional Purposes*(pp. 373-391). Singapore: Springer.

Wang, X. (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL, 9*.