澳门中文媒体的词汇演变——以《澳门时报》为例 *

澳门大学 王 珊 深圳大学/澳门大学 陈 钊
深圳大学/中国科学院 张昊迪

提要：词汇增长模型能够通过分析词种（Types）与词例（Tokens）之间的关系，揭示词汇演变。澳门作为一个多语言、多文化交汇的城市，其词汇使用能够反映社会热点。然而，目前关于澳门词汇演变的研究较少。本研究选择澳门中文媒体《澳门时报》构建新闻语料库，应用 3 种词汇增长模型拟合词汇变化，并根据效果最佳的 Heaps 模型，进一步探讨词汇变化与新闻内容之间的关联性。结果表明，澳门新闻词汇的演变与国内外热点事件、施政方针以及民生紧密相关。此外，通过分析打乱文本时序后的文本，验证了所采用方法的有效性。对《澳门时报》词汇演变的分析为深入理解澳门的语言生活提供了重要参考。

关键词：中文媒体、词汇、新闻、历时、语料库、澳门

1 引言

计量语言学领域对词种（Types）与词例（Tokens）之间的关系有广泛的探讨，词种词例比（Type-Token Ratio，缩写为词种词例比）常用于评估文本的词汇丰富度。然而，词种词例比只能反映词种与词例在某一时间点的关系，无法用于分析词汇的历时演变。为了解决这一问题，词汇增长模型应运而生，通过函数拟合词种与词例之间的增长关系，可以描述词种词例比的变化趋势。目前，主流的词汇增长模型包括 Guiraud 模型（Guiraud 1954）、Heaps 模型（Heaps 1978）和 Hubert 模型（Hubert & Labbé 1988）。这些模型能够根据词例的数量预测词种的增长情况，适用于分析有时间顺序的语料的词汇变化。

澳门作为中西文化的交汇地，推行“三文四语”政策，但目前关于澳门词汇演变的研究较少。媒体作为信息传播的主要渠道，是民众获取社会信息的重要来源，

* 本研究受教育部国家语委“十三五”科研规划项目“澳门旅游休闲汉语研究”（YB135-159）的阶段性成果。王珊和张昊迪为本文通信作者。
作者贡献：
王珊：选题构思、数据收集、研究方法、初稿撰写、修改润色、字数占比（40%）；
陈钊：数据收集、数据分析、初稿撰写、字数占比（30%）；
张昊迪：研究方法、修改润色、字数占比（30%）。

而新闻媒体能够在一定程度上客观地反映社会现象和发展趋势。本文通过构建澳门中文媒体《澳门时报》新闻语料库，运用词汇增长模型进行分析，以验证此类模型对研究澳门新闻词汇变化的有效性。

2 词汇增长和澳门词汇研究

2.1 词汇增长研究

词种词例比是一种用来衡量二者之间关系的指标，广泛应用于作者身份判定、语言掌握情况评估等领域。例如，Hoover（2003）分析了 12 位作者的作品，发现词种词例比可以有效地判断作者的身份；Yu（2010）认为词种词例比与写作和口语质量在统计学上存在显著的正相关性，语言能力较高的人其词种词例比值也相对较高，因此词种词例比被视为评估学习者语言掌握情况的重要指标；Mellor（2010）指出词种词例比经常用于衡量说话者和写作者的语言水平，语言水平越高使用的低频词汇就越多；Wang（2014）分析了英语二语学习者电邮的词种词例比与写作熟练度之间的关系。

尽管词种词例比在分析固定文本内部的词汇特征方面非常有效，但由于它只能表示特定时间点上的词种与词例的比值，这使得它在历时分析中存在局限。而词汇增长模型通过数学函数来学习词种与词例之间的数量增长关系，学习后预测的词种与词例比值与词种词例比值的原理相似，能够有效地反映词汇丰富度。例如，Savoy（2015）分析了 1790—2014 年美国国情咨文演讲中词汇量的增长情况，涵盖了 42 位美国总统的 225 篇演讲。结果显示，当白宫应对反复出现的问题时，总统更倾向于重复使用论点；而在面对新情况或提出新方案时，词汇增长的预测值与实际观测值的差异变大。此外，王珊、王会珍（2021）通过分析中国 1954—2018 年的政府工作报告，探讨了词汇增长与政策之间的关系，并验证了该方法在中文数据集中的适用性。这些研究表明，词汇增长模型在分析演讲、政府工作报告的词汇演变方面具有重要价值。

2.2 澳门华语词汇研究

大华语是以普通话为基础的全球华人共同语（李宇明 2017），其在语音、词汇、语法、语用上有一定弹性和宽容度（陆俭明 2017），存在多个变体（徐大明、王晓梅 2009；李宇明 2017；田静、苏新春 2018）。

澳门作为中西文化的交汇之地，已有不少关于其语言使用的研究。例如，邵朝阳（1999）通过实地语料收集，分析了澳门博彩业隐语的使用；黄翊（2005）探讨了澳门清代中文文档中具有地方特色的词语；黄翊（2007）进一步分析了澳门语言

生活的形成与发展；汤志祥（2008）认为香港与澳门的华语具有高度的共通性，但在词汇上仍存在一些差异；袁伟（2015）讨论了澳门中文平面媒体中的字母词规范问题；姚双云、黄翊（2014）对比了澳门与内地新闻词汇的差异；Wang & Luo（2019）分析了与澳门旅游相关的词汇使用情况；王珊、汤蕾（2022）从词汇来源、语音特点、构词方式和语义特征等方面分析了澳门华语的特色词汇，并考察了这些词在LIVAC 汉语共时语料库中的区域分布和词义变化。然而，对澳门词汇使用的研究主要集中在共时层面，缺乏基于语料库的历时词汇增长与社会热点关系的分析。对澳门这样一个多语言、多文化地区的词汇演变进行研究，不仅可以验证词汇增长模型在新闻领域的适用性，还能够填补澳门词汇演变研究的空缺，提供更深入的语言发展动态分析。

3 构建《澳门时报》语料库

3.1 语料选取

新闻具有客观性、公正性和真实性，能够一定程度上反映社会热点和重大事件。本研究对澳门的多家新闻媒体进行考察，包括《澳门日报》《市民日报》《澳门法治日报》以及《体育周报》等。然而，《澳门日报》和《市民日报》分别仅提供 2 年和 1 年的语料，数据量较小，难以支持历时分析。《澳门法治日报》和《体育周报》主要聚焦于法律和体育领域的报道，无法全面反映社会热点的动态变化。因此，本文选择了综合性报纸《澳门时报》作为研究对象。《澳门时报》原名《时事新闻报》，1972 年创刊，2015 年 9 月更名为《澳门时报》，2016 年 6 月 15 日由周刊改为日报。《澳门时报》以客观、专业的立场报道中国及全球的时事新闻和民生大事，紧密联系社情民意，关注社会民生。该报涵盖的语料内容广泛，包括法律、体育、民生以及国际热点等多个领域，视角全面而多元，适用于本研究的词汇分析。

3.2 语料预处理

由于中文文本中词与词之间没有间隔，进行词汇研究时需要首先对文本进行分词。本文选用了由北京大学语言计算与机器学习组开发的 pkuseg 分词工具包[1]（Luo *et al*. 2019）。pkuseg 是支持针对不同领域的个性化分词工具，提供了新闻、网络、医药、旅游及混合领域的分词模型。鉴于本研究所使用的《澳门时报》语料属于新闻领域，本文采用了 pkuseg 的新闻分词模型，以确保分词的准确性和可靠性。

由于 pkuseg 分词工具包仅支持简体中文分词，而《澳门时报》使用的是繁体中文，因此本文首先使用 OpenCC 工具将繁体中文转换为简体中文，然后利用

pkuseg 对简体文本进行分词。分词完成后，再将分词结果对应回《澳门时报》中繁体文本中的切分位置，以获得最终的分词结果。此方法可以避免繁简转换过程中出现的“一简对多繁”或“一繁对多简”等问题。本文还对分词结果进行了进一步的处理，去除包含阿拉伯数字的词语（如包含“百”“千”“万”“年”“月”和“日”）以及标点符号，例如删除“1 万”和“1996 年”这类的词语，而“千锤百炼”这类的词语则保留。在词语与短语的界定上，学术界有较多争议。董秀芳（2004）认为多音节结构应归为复合词，如“计算机病毒”“国家安全系统”“社会政治经济学”等。因此，本研究也将类似的结构视为词，如“半导体工厂”“新冠病毒核酸检测”等。

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（每年所有词的总数）、词种数（每年词例中不同词的数量）以及字数等数据。该语料库涵盖了从 2011 年 1 月 1 日—2021 年 6 月 7 日的内容，共收录了 50,411 篇文章，字数达 23,440,059 个，包含 10,303,095 个词例和 338,598 个词种[2]。其中，2011—2020 年的语料共计 47,856 篇文章，字数为 22,096,302 个，包含 9,711,067 个词例和 326,466 个词种[3]。

表 1　2011—2021 年《澳门时报》语料库概况

年份	文章数	词例数	词种数	字数
2011	1,205	360,657	30,972	793,163
2012	1,222	365,734	32,762	806,047
2013	1,482	444,130	36,152	982,346
2014	1,660	362,461	36,533	805,947
2015	2,148	437,260	45,121	982,622
2016	8,482	1,445,242	108,867	3,323,143
2017	11,801	2,043,629	135,444	4,683,304
2018	7,152	1,475,069	97,418	3,378,961
2019	6,066	1,306,562	80,320	3,016,205
2020	6,638	1,470,323	80,964	3,324,564
2021	2,555	592,028	47,228	1,343,757
总计	50,411	10,303,095	338,598	23,440,059

4 《澳门时报》的词汇增长

4.1 3 种词汇增长模型

词汇增长模型假定词例和词种之间存在一定的函数关系，通过拟合出来的模型预测在不同词例数量条件下的词种数。为了评估词汇丰富度的变化，通常会对比词种词例比与模型预测的词种值以及实际观测值之间的差异。现有的主要词汇增长模型包括 Guiraud 模型（Guiraud 1954）、Heaps 模型（Heaps 1978）和 Hubert 模型（Hubert & Labbé 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 的平方根的比值为常数。根据这个关系推导出预测词种数量V'与当前词例数量n之间的关系公式见(1)。

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

Heaps 是指数预测模型，公式见(2)和(3)。

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

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

其中，V'表示词种的预测值，a和C为模型的参数，n为词例数量。Heaps 模型能够很好地拟合词例与词种之间的关系。Tweedie & Baayen（1998）的研究进一步指出，a和C这两个参数与词例数量相关。Hubert & Labbé（1988）提出了更复杂的 Hubert 模型，认为词种可以分为通用词汇（general vocabulary）和专业词汇（specialized vocabulary）两类。专业词汇在文本中通常很少重复出现，因此其词种与词例之间的比值呈线性关系。基于此，Labbé *et al.*（2004）提出假设，专业词汇占文本词种数的比例为p，通用词汇占文本词种数的比例为$1-p$。该模型的公式见(4)。

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

V是文本的词种数，k是出现的最高词频，V_i是出现频次为i的词的数量，u表示预测语料占总语料的比例，即一篇文章中需要预测的词例数与文章总词例数的比值，而$(1-u)^i$表示频率为i的词种不出现在样本中的概率。此外，Hubert 模型假设词语在文本中的出现次数与文本的长度有关，并且词语的出现频率与其增长速率成正比关系，即出现频率越高的词语，其增长速率越快。

本文采用 Guiraud、Heaps 和 Hubert 三个模型对《澳门时报》语料库进行拟合分析，并使用 scipy[4] 框架，通过非线性最小二乘法对数据的残差平方和进行优化，见公式(5)。该函数用于评估模型预测值与实际观测值之间的误差，并对模型进行优化。其中，Guiraud 模型的参数为$\mathrm{r} = 98.028625$；Heaps 模型的参数为$\mathrm{a} = 2.785034$，

C = 0.727058；Hubert 模型的比例参数为p = 0.316723。本文将选择效果最优的模型进行误差分析。

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

4.2 《澳门时报》的词汇增长

本节对全部语料进行处理，以每 1,000 个词例为采样单位进行模型拟合。图 1 展示了《澳门时报》语料库中词种数量随词例数量增加的实际观测值，以及 3 个模型对这一变化的预测值。Tokens 的数量为x轴，Types 的数量为y轴的曲线。Guiraud 模型的预测曲线在整体上表现出明显的偏差。起初，词例数量较少时，预测值明显高于实际观测值，随着词例数量的增加，模型的预测值明显低于实际观测值。Heaps 模型和 Hubert 模型的预测值都能够较好地反映实际观测值的变化，但 Heaps 模型在整个过程中与实际观测值更为接近。

图 2 进一步展示了观测值与预测值之间的差异，其中数值越接近 0，预测的准确性越高。结果显示，Guiraud 模型的预测效果最差，且方差较大，与实际数据的平均差值为−7,414.67。Hubert 模型的预测效果相对较好，其与实际数据的平均差值为 1,729.56。Heaps 模型的预测效果最佳，与实际数据的平均差值仅为−526.22。从 3 条曲线中可以明显看出，Heaps 模型的预测效果优于 Hubert 模型，而 Guiraud 模型的效果最差。因此，本文在后续分析中将采用 Heaps 模型的预测结果进行分析。

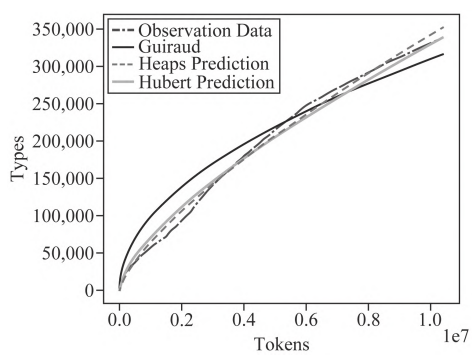

图 1　不同模型词种数量预测曲线

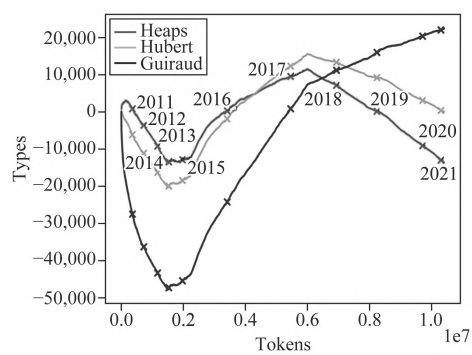

图 2　不同模型预测误差分布

5　澳门中文媒体词汇分析

表 2 统计了《澳门时报》2011—2020 年的数据，包括词例数（每年及其之前所有年份的词例总和）、观测词种数量（每年及其之前所有年份的词种数量总和，相同

词种在不同年份中只计算一次）、Heaps 预测值、观测值词种词例比、预测值词种词例比、预测误差（预测值–观测值）和新增词语数量（即每年新增的词种数，等于该年的词种观测值减去之前所有年份的词种观测值）。

表 2　每年观测值与 Heaps 预测值

年份	词例数	观测值（词种数）	Heaps 预测值	预测误差	观测值词种词例比	预测值词种词例比	新增词语数
2011	360,657	30,972	30,560	−412	0.086	0.085	—
2012	726,391	47,370	50,844	3,474	0.065	0.070	16,398
2013	1,170,521	62,510	71,927	9,417	0.053	0.061	15,140
2014	1,532,982	76,015	87,513	11,498	0.050	0.057	13,505
2015	1,970,242	93,893	105,029	11,136	0.048	0.053	17,878
2016	3,415,484	159,417	156,685	−2,732	0.047	0.046	65,524
2017	5,459,113	231,432	220,347	−11,085	0.042	0.040	72,015
2018	6,934,182	270,122	262,198	−7,924	0.039	0.038	38,690
2019	8,240,744	298,395	297,261	−1,134	0.036	0.036	28,273
2020	9,711,067	326,466	334,948	8,482	0.034	0.034	28,071

新增词语的数量范围大体在 1.3 万—7.2 万之间波动，新增词语数量较多主要是因为搜集到的《澳门时报》语料仅涵盖近十余年的新闻文本，该媒体没有提供在此之前的新闻，导致无法包括所分析的时间点之前的词汇。此外，每年新闻报道的主题不同，也会引发新增词语数量的显著变化。例如，当计算 2014 年的新增词语时，对比的是 2011—2013 年《澳门时报》的词语。尽管无法获得 2011 年之前的语料，但这并不妨碍本文对社会热点与近年来澳门词汇演变之间关系的分析。

接下来分析模型预测值与《澳门时报》实际观测值之间的关系。2011 年 Heaps 模型的预测值略低于实际观测值，观测值为 30,972，预测值为 30,560，前者多出 412。这表明 2011 年有较多的热点。例如："核电站"与日本福岛核电站的核泄漏事故有关，该事故在全球范围内引起了广泛关注，核能安全成为热门话题；"珠海北站"指的是珠海北站的正式投入使用；"本拉登"与他被击毙的事件相关，这一事件引发了全球范围内的反恐讨论和安全关注。2008 年底公布的《珠江三角洲地区改革发展规划纲要》，首次从国家发展战略层面明确澳门"世界旅游休闲中心"的发展定位[5]。2011 年，国家"十二五"规划提出"支持澳门建设世界旅游休闲中心""支持澳门推动经济适度多元化""加强内地和香港、澳门交流合作"[6]，进一步推动了澳门本地

旅游资源的开发和利用。在这一政策背景下，出现了如“控烟”“区域性”“国际休闲中心”“一程多站”等词语。“控烟”反映了澳门特区政府为提高公共健康水平，推行严格的控烟措施；“区域性”指的是澳门在区域合作中的角色和影响力不断增强；“国际休闲中心”体现了澳门作为全球旅游目的地的地位不断巩固；“一程多站”是指旅游线路规划中多个目的地的组合，增强了游客的体验。

2012 年预测值高于观测值，观测值为 47,370，预测值为 50,844，前者比后者少 3,474，表明这一阶段的词汇使用相对稳定。这一年《澳门时报》报道的全球和本地热点事件较少，报道主要集中在民生领域，例如“物业”“医疗”和“市场”等非新增词语的使用频率大幅增加。不过该阶段仍然出现了 16,398 个新增词语，反映了当时一些引发广泛关注的新闻事件。例如：澳门特区政府在这一年增加了对历史文化的重视，《文化遗产保护法》进入立法程序，催生了“文遗法”“物质文遗”和“文化遗产”等词语。此外，2012 年神舟九号载人飞船成功发射，这一重大航天事件带来了“神舟”“载人航天代表团”和“航天员”等与航天相关的词汇。澳门大西洋银行和中国银行发行了纪念钞“龙钞”，由此出现了“发钞”“新钞”和“龙钞”等与金融相关的词汇。同时，澳门大力推进 3G 信号的普及，这使得“2G”“3G”和“智能机”等与通讯技术相关的词汇频率大幅提高，反映了当时通讯领域的快速发展和技术升级。这些词汇的出现，体现了澳门在历史文化保护、航天事业、金融创新和通讯技术等方面的发展趋势。

2013 年预测值高于观测值，观测值为 62,510，而预测值为 71,927，前者比后者少 9,417。该年度新增词语 15,140 个。这一年是科技飞速发展的一年，许多与科技相关的词汇涌现出来。其间发生了几个重要事件，包括辽宁舰首次编队赴南海试验训练、神舟十号成功发射以及 2012 年底第二代北斗卫星开始为亚太地区用户提供区域定位服务。这些事件催生了一批相关词汇，如“北斗”“辽宁舰”“军舰”和“探月”等。2013 年也是澳门立法会选举年，与选举和立法相关的词汇如“选举法”和“选举日”等的使用频率显著上升。尽管这一年有许多新闻热点，但由于澳门政策的相对稳定以及热点新闻的减少，媒体报道更多集中在民生问题上。这导致 Heaps 模型预测的词汇增长速率高于实际观测值的增长速率，观测值与预测值之间的差距较上一阶段扩大了 5,943。这一现象表明，尽管科技和政治事件引发了部分新增词语的出现，整体词汇使用的变化趋于平稳，更多的词汇增长反映了澳门社会对民生问题的持续关注。

2014 年观测值比预测值小，观测值为 76,015，预测值为 87,513，前者比后者少 11,498，新增词语数量为 13,505 个。该年的新闻报道主要聚焦于民生问题，如治安、青少年事务和社会福利等话题。这一年也是澳门回归祖国 15 周年，催生了“爱国人士”等之前未出现的与爱国相关的词汇，反映了澳门社会对这一重要历史时刻的关注。此外，埃博拉病毒在全球范围内进一步蔓延，导致了“埃博拉”和“隔离带”

等与传染病相关的词语进入公众视野，反映了社会对公共卫生和传染病防控的高度关注。澳门 A 区的填海工程也成为媒体关注的焦点之一，“轻轨半岛”等词汇出现，反映了澳门城市基础设施建设的进展。此外，为了保障本地人才的培养，澳门特区政府在 2014 年提出了一系列长期发展计划，催生了“精英人才团队”和“人才发展委员会”等词汇。这些词汇反映了澳门在推动人才培养和提升城市竞争力方面的政策导向。同年，马航 MH370 航班失踪事件震惊全球，新闻报道中出现许多与航天和空难相关的词汇，如“空难”“尾翼”“雷达应答机”和“马航”等，显示出社会对这场空难事件的持续关注和讨论。总的来说，2014 年的词汇变化反映了澳门社会在关注民生的同时，对重大国际事件、公共卫生问题以及城市发展等领域的关注。

2015 年预测值大于观测值，观测值为 93,893，预测值为 105,029，前者比后者少 11,136，新增词语 17,878 个。这一年，澳门的博彩业收入创 5 年来新低，不过澳门特区政府一直倡导的经济多元化显示出成效，出现了许多其他行业的词汇。例如“纺织业”一词反映了澳门在制造业方面的复苏，“连锁书店”体现了零售和文化产业的扩展。值得一提的是，这一年电子商务领域在澳门获得了关注，支付宝于 2015 年进入澳门市场，成为首个在澳门提供扫码支付服务的第三方支付平台，“微商”代表了新兴经济模式在澳门的出现和发展；“网购”反映了跨境电商风潮下人们购物方式的转变；“亚马逊”和“淘宝”代表了全球和区域性的电子商务平台；“顺丰”作为物流快递的代表，反映了跨境电商对物流行业的需求增长。这些词不仅展示了澳门在经济多元化发展中的新趋势，还反映了科技进步和消费模式的转变对澳门社会的影响。此外，2015 年也是抗日战争胜利 70 周年，北京天安门广场举行了盛大的阅兵仪式以纪念这一历史事件，由此出现了一些之前阶段未出现的词语。例如：“慰安妇”引发了社会对历史正义的再度关注；“阅兵式”不仅描述了此次大型军事活动，还象征着国家力量；“南京军事法庭”与二战后的战争罪审判有关，引起人们对历史事件的反思和讨论。

2016 年预测值结果开始小于观测值，说明新增热点较多。预测值为 156,685，而观测值为 159,417，两者相差 2,732，新增词语 65,524 个。这一年，《澳门时报》对国外体育联盟的报道较多，大量体育相关的新增词语涌现。例如：“湖人”“西甲”“投篮”“拜仁”和“皇马”等词汇都与 NBA 和西甲联赛等体育赛事相关，反映了澳门社会对国际体育赛事的浓厚兴趣和关注。澳门在市场上 3 次发现禽流感病毒，引发了广泛的公共卫生关注。与此相关的词汇如“横琴检验检疫局”“传染病中心”和“病毒”成为新闻报道中的高频词，反映了澳门社会对传染病防控和公共健康的重视。此外，在第十一届中国国际航天博览会上，歼-20 战斗机首次公开亮相并进行飞行展示，这一重大军事事件引发了大量与之相关的词汇的出现，如“歼-20”“载弹量”和“超音速”等。这些词汇不仅反映了中国军事技术的进步，也引起了广泛的社会关注，成为当年的重要话题之一。总体而言，2016 年大量涌

现的词汇主要集中在体育和军事领域。《澳门时报》对 NBA 和西甲等国际体育赛事的广泛报道，以及对中国航天技术的关注，都是导致 Heaps 模型预测值低于实际观测值的主要原因。这一年新增词语的暴发，体现了澳门社会对全球事件的敏锐反应和高度关注。

2017 年预测值小于观测值，预测值为 220,347，观测值为 231,432，两者相差 11,085。这一差距可以归因于几个关键因素：首先，《澳门时报》从 2016 年 6 月中旬起，由周刊变为日刊，导致语料量大幅增加，直接影响了词汇的观测值；其次，澳门在 2017 年推出了一些新政策，特别是在智慧城市建设方面，采用了大数据和人工智能等技术来优化城市管理，并提供了多项相关服务。这些政策的实施催生了一些新增词语，如“智慧城市联盟工作组”“智慧城市委员会”和“澳门云计算中心”等，这些词语反映了澳门在智慧城市建设中的进展和创新。澳门的电话诈骗案件数量显著上升，媒体对此进行了大量报道，导致了“电骗”“电号”和“电人”等词汇的出现，反映了社会对电信诈骗问题的高度关注以及政府和公众对打击此类犯罪所做的努力。中国出现了“新四大发明”的概念，指代高铁、支付宝、共享单车和网购，这些新兴事物迅速融入社会生活，相关的词汇如“共享化”“摩拜”和“轻轨博物馆”等也随之进入了媒体报道的高频词汇中。这些词汇不仅反映了中国科技和创新的快速发展，还展示了这些新发明对社会的深远影响。总体而言，2017 年的词汇增长反映了澳门社会在多个领域的变化与发展，特别是在智慧城市建设和应对新兴社会问题方面。

2018 年预测值小于观测值，预测值为 262,198，观测值为 270,122，两者相差 7,924。该年度共出现了 38,690 个新增词语，大部分与美食、应急机制和人工智能相关，如“本澳防灾避险中心”“智慧警务所”“美食之都小组”“澳门美食网路”“美食旅游团”“阿里云计算有限公司”“港澳大湾区人工智能联盟”和“人工智能时代”等。这些词汇反映了当年澳门社会的几个重要热点事件。首先，2017 年澳门经历了一场严重的飓风灾害，造成了巨大的损失。这一背景下，澳门特区政府在 2018 年出台了“完善应急机制，强化公共安全”的政策措施，切实提高防灾减灾的能力和水平。新增词语如“本澳防灾避险中心”和“智慧警务所”反映了这些措施的实施，以及社会对公共安全和应急响应能力的关注。其次，澳门致力于打造“创意城市美食之都”，这一目标在 2018 年得到进一步推动。澳门于 1 月 17 日启动了“澳门美食年”项目，推广澳门特色美食，提升城市的国际知名度，从而推动旅游业的发展。新增词语如“美食之都小组”“澳门美食网路”和“美食旅游团”正是这一系列活动的产物，展示了澳门在美食文化推广方面的努力和成果。最后，智慧城市建设在 2018 年进一步推进，澳门开始更加重视人工智能、大数据和云计算技术的发展。这一趋势催生了诸如“阿里云计算有限公司”“港澳大湾区人工智能联盟”和“人工智能时代”等词语，反映了澳门在科技创新和城市智能化建设方面的积极探索。总体而言，

2018 年的词汇增长不仅反映了澳门在应对自然灾害、美食文化推广和智慧城市建设方面的努力，也展示了澳门在全球化和技术进步背景下的社会变革。

2019 年预测值小于观测值，观测值为 298,395，预测值为 297,261，两者相差 1,134。这表明模型预测值与观测值之间的误差逐渐缩小，反映出新闻报道内容开始趋于稳定。2019 年新增 28,273 个词语，其中许多与科技发展密切相关。2019 年被视为科技发展的重要一年，澳门在这一年迎来了多个重大科技事件。例如：澳门电讯成功拨通了澳门首个 5G 电话，这标志着澳门正式进入 5G 时代，推动了通信技术的快速发展；华为推出了自主研发的鸿蒙操作系统，作为应对美国制裁的一部分，这一事件引发广泛关注，并催生了与之相关的词语如“鸿蒙”和“美国科技战”等。澳门还举办了三大国际人工智能会议，这为澳门人工智能产业的发展注入了新的动力，带来许多与人工智能相关的词汇。例如：“人工智能科学”“人工智能控烟”和“人工智能门诊”等词汇展示了人工智能技术在医疗和公共健康领域的应用和扩展。此外，“微电子研发中心”体现了澳门在微电子领域的研发进展；“半导体工厂”反映了澳门在半导体制造方面的努力。新增词语不仅记录了澳门在科技创新领域的进展，也表明了澳门在面对全球科技变革时的应对策略。

2020 年预测值大于观测值，预测值为 334,948，观测值为 326,466，相差 8,482。这一年最大的事件无疑是新型冠状病毒的全球大流行。2020 年 1 月 22 日，澳门确诊了首例新冠肺炎病例，此后疫情迅速蔓延，全球经济受到严重冲击，许多国家和地区陷入衰退。面对这一前所未有的危机，中国展现了强大的动员能力，迅速采取措施应对疫情，并努力推动经济复苏。新增词语的增加主要集中在与新冠疫情相关的领域。例如：“新冠病毒核酸检测”反映了广泛应用于诊断新冠病毒感染的关键技术；“绿码”是健康码系统中的一个重要标志，用于显示个人的健康状态；“新冠病毒”成为了疫情报道中的核心词汇；“粤澳健康码”是粤港澳大湾区为应对疫情而推出的跨境健康码系统；“澳门应急医疗队”体现了澳门在应对公共卫生危机中的迅速反应；“紧急防疫指挥中心”代表了政府在疫情期间设立的危机管理机构。虽然 2020 年新增了 28,071 个词汇，但由于与新冠或医疗相关的词汇使用频率极高，这些词汇的增长速度超过了其他新增词汇的增加速度，导致整体词汇丰富度（词种词例比）略有下降。也就是说，尽管新增词汇数量显著增加，但这些高频词汇的使用导致了整体词汇分布的集中化，减少了词种词例比的变化幅度。这一现象反映了疫情对社会的深远影响，新冠相关词汇迅速占据了媒体报道的主导地位，成为这一年语言使用的显著特征。

6 澳门词汇增长的验证程序

上一节通过分析不同年份的新增词语和相关热点事件，探讨了预测值与观测值

之间的差异。然而，这些差异究竟是由新闻内容的变化引起的，还是由模型拟合误差导致的，还需要进一步验证。本节通过随机化方法对收集到的所有语料进行验证。如果误差的增大是由新闻内容引起的，那么随机化实验的结果中，误差应该有所减少（Savoy 2015）。验证方法为将整个语料库的所有词语顺序打乱，随机生成新的文本。这些新文本的词例总数和词种总数与原始语料库保持一致。如果打乱后的观测值与预测值的差异小于原语料库的差异，而唯一的变量是文本的原有时序语义信息，那么就可以证明，新闻内容的实时性是导致观测值与模型预测值之间误差的主要原因。

为了更好地观察预测值与观测值之间的差异，本文使用模型拟合后的结果计算预测值与观测值差异的标准分 Z-Score（Savoy 2015）作为衡量模型拟合效果的指标。这种方法能够更加准确地判断新闻内容的变化在多大程度上影响了模型的预测效果，并验证模型在处理不同语料时的可靠性。

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。训练时 Heaps 模型参数为$a = 35.708521$，$C = 0.566868$。可以看出，Heaps 模型拟合的曲线几乎与随机文本的增长曲线重合，其平均误差为-52.08，显著低于真实文本中的拟合误差。

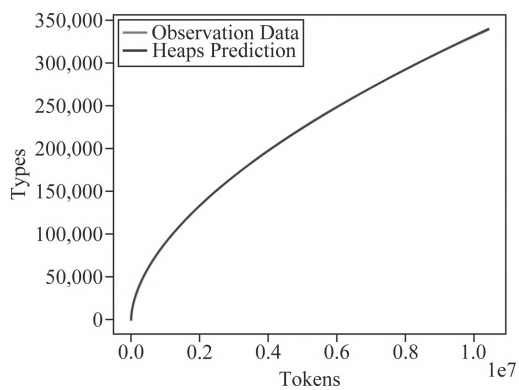

图 3　随机文本增长与预测曲线

图 4 和图 5 分别展示了真实文本与随机文本的预测误差 Z-Score 分布。在这两个实验中，大部分 Z-Score 数据都处于［-3σ，3σ］的范围内，这表明这两个实验的误差均在可接受的范围内。图 3 显示，随机文本的预测误差远小于真实文本的实验误差，且随机文本的误差表现出一定的随机性，从而进一步支持了这一结果：真实文本中 Heaps 模型的拟合误差主要来源于文本内容的变化，而不是模型的缺陷。

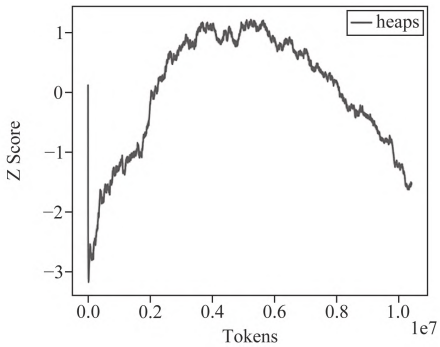

图 4　真实文本的预测误差 Z-Score 分布

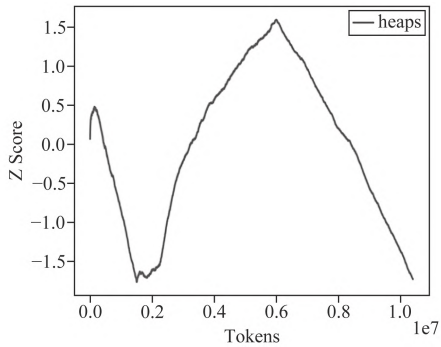

图 5　随机文本的预测误差 Z-Score 分布

7　结论

目前对澳门词汇的分析主要集中在共时层面，缺乏利用大规模语料库进行的词汇演变研究。首先，本文以《澳门时报》十余年的语料为基础，探索澳门词汇的变迁。其次，本文采用 3 种词汇增长模型即 Guiraud、Heaps 和 Hubert 模型，对词汇的历时演变进行了建模分析。结果显示，Heaps 模型的拟合效果最佳。通过对 Heaps 模型预测值与实际观测值之间的差异进行探究。本文发现，在政策稳定及世界热点新闻较为平稳的情况下，Heaps 模型的词种预测值往往高于观测值，反映出词语的重复使用率较高，新增词语较少，词种词例比值较低；相反，当出现重大热点事件或新政策出台时，观测值通常高于预测值，表明新增词语较多。再次，本文通过将语料库中的文本顺序随机打乱，使其不再具备原文的语义时序信息，并利用 Heaps 模型对乱序文本进行拟合分析。通过对比原始语料库与随机文本的标准分数，进一步证明了模型预测误差与文本内容之间的相关性，也验证了 Heaps 模型在新闻媒体语料历时分析中的有效性。本文是首项基于语料库开展的中国澳门词汇演变的研究，揭示了澳门的语言生活不仅关注本地民生，还密切关心国家大事和国际时事，展现了澳门作为中西交汇的社会所具备的家国情怀和国际视野。

注释

1　网址：https://github.com/lancopku/pkuseg-python。

2　将 2011 年 1 月 1 日—2021 年 6 月 7 日的词种数相加为 731,781 个，去除重复后为 338,598 个。

3　将 2011—2020 年每年全年词种数相加为 684,553 个，去除重复后为 326,466个。

4　网址：https://www.scipy.org/。

5　网址：https://www.gov.mo/zh-hans/news/48808/。

6 网址：https://www.gov.cn/2011lh/content_1825838_15.htm。

参考文献

GUIRAUD P. Les caractères statistiques du vocabulaire: essai de méthodologie [M]. Paris: Presses universitaires de France, 1954.

HEAPS H S. Information retrieval, computational and theoretical aspects [M]. New York: Academic Press, 1978.

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

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

LABBE C, LABBE D, HUBERT P. automatic segmentation of texts and corpora [J]. Journal of Quantitative Linguistics, 2004, 11(3): 193-213.

LUO R, XU J, ZHANG Y, et al. PKUSEG: a toolkit for multi-domain Chinese word segmentation [J]. ArXiv abs/1906. 11455, 2019.

MELLOR A. Automatic essay scoring for low level learners of English as a second language [D]. Sketty: Swansea University, 2010.

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

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: 323-352.

WANG S, LUO H. A corpus-based study of the vocabulary of Macao tourism Chinese [C]// TAO H, Chen H H-J. Chinese for specific and professional purposes. Singapore: Springer, 2019: 373-391.

WANG X. The relationship between lexical diversity and EFL writing proficiency[J]. University of Sydney Papers in TESOL, 2014, 9: 65-88.

YU G. Lexical diversity in writing and speaking task performances[J]. Applied Linguistics, 2010, 31(2): 236-259.

董秀芳. 汉语的词库与词法[M]. 北京：北京大学出版社，2004.

黄翊. 澳门语言研究[M]. 北京：商务印书馆，2007.

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

李宇明. 大华语：全球华人的共同语[J]. 语言文字应用, 2017,（1）：2-13.

陆俭明. “华语”的标准：弹性和宽容 [J]. 语言战略研究, 2017, 2（1）：1.

邵朝阳. 澳门博彩隐语研究[J]. 中国语文，1999，（4）：267-274.

汤志祥. 论“港澳词语”以及“澳门特有词语”[J]. 江苏大学学报（社会科学版），2008，（5）：24-29.

田静，苏新春. 文化互动视野下的“大华语”概念新探——兼谈华语社区词的文化间性[J]. 新疆社会科学，2018，（5）：142-148.

王珊，汤蕾. 澳门华语特色词汇研究[J]. 语言战略研究，2022，（2）：74-85.

王珊，王会珍. 中文词汇增长研究[J]. 中文信息学报，2021，（1）：17-24.

徐大明，王晓梅. 全球华语社区说略[J]. 吉林大学社会科学学报，2009，（2）：132-137.

姚双云，黄翊. 澳门与内地新闻语篇词汇差异的计量研究[J]. 语言文字应用，2014，（2）：27-37.

袁伟. 中国澳门特区中文平面媒体中字母词的规范研究[J]. 语言文字应用，2015，（3）：68-75.

通信地址：999078　澳门特别行政区　澳门大学人文学院（王珊、陈钊）

519000　广东省珠海市　珠海澳大科技研究院（王珊）

518052　广东省深圳市　深圳大学计算机与软件学院（陈钊、张昊迪）

200031　上海市　中国科学院上海脑科学与类脑研究中心（张昊迪）

includes quotes from both Chinese and foreign proverbs, laws, and regulations. Discourse strategies encompass nomination/referential strategies, predication strategies with adjectives and predicates, argumentation strategies involving influence and law, perspectivization strategies, and intensification/mitigation strategies. These contribute to constructing a positive image of China as a responsible country facing global health emergencies.

An accuracy analysis of the definite article in lexical bundles in Chinese EFL learners' essay writing

Most previous research has examined erroneous uses of the definite article within noun phrases, with little known about its use in lexical bundles. Using the Ten-thousand English Compositions of Chinese Learners (TECCL Corpus), this study investigates the types of definite article errors in four-word lexical bundles in Chinese EFL learners' writing. Our findings reveal that omission errors are as the most frequent error type, occurring mostly in noun-phrase bundles containing of-phrase fragments. These findings may provide useful insights into the teaching and research of definite articles.

The vocabulary growth in Macao's Chinese media: A case study of *Macao Times*

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 multilingual and multicultural integration, Macao's vocabulary use can reflect the focus of society, but there is no research on Macao's diachronic vocabulary growth. This paper constructed a diachronic corpus of Macao 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 Macao vocabulary is closely related to hot news, policy guidelines and people's livelihood. This research also used texts in random order after removing the timing information to verify the effectiveness of the method. This is the first study to investigate the evolution of Macao vocabulary based on a large-scale diachronic corpus, which is of great significance for the in-depth understanding of the development of Macao's language life.