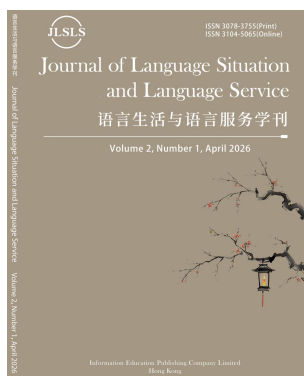




JLSLS



Journal of Language Situation and Language Service

JLSLS, Vol. 1, No.1, 2026, pp.57-70

Print ISSN: 3078-3755; Online ISSN: 3104-5065

Journal homepage: <https://www.lslsjournal.com>

DOI: <https://doi.org/10.64058/jls260105qsdew>

广州市政府工作报告词语发展变化的计量研究

侯仁魁(Hou Renkui), 张锦程(Zhang Jincheng)

摘要: 本文收集 1993-2024 年广州市政府工作报告建立语料库,选择表示主题的实词,包括名词和动词,基于向量空间模型建立文本向量,以文本聚类 and 对应分析为手段,考察报告中三十年来表示社会主题的名词和动词的发展变化。文本聚类结果显示,政府工作报告大致分为五个阶段,分别是 1993-2000, 2001-2008, 2009-2012, 2013-2019 和 2020-2024。从对应分析结果图中能够看出有些词语在某个阶段的报告中出现频率较高。接着以前一阶段报告为参照语料库(Reference Corpus),计算当前阶段关键词,关键词或者是在当前阶段报告中第一次出现,或者出现频率大幅提高,更能明确地观察到词语的变化,从而了解报告专注点的变化。最后,采用计量语言学方法抽取每个阶段报告文本的主题词(thematic word),计算它们的交集,从而发现所有阶段恒定不变的关注点,如“发展”、“经济”、“教育”、“改革”和“文化”等。

关键词: 广州市政府工作报告, 文本挖掘, 主题, 历时演变

作者简介: 侯仁魁, 通讯作者,副教授, 广州大学人文学院、国家语言服务与粤港澳大湾区语言研究中心, 应用语言学, 电邮:hourk0917@163.com; 张锦程: 研究生, 广州大学人文学院, 语言学及应用语言学, 电邮: 2112412080@e.gzhu.edu.cn。

Title: Quantitative Study on Diachronic Changes of Words in Work Reports of Guangzhou Municipal Government

Abstract: This paper collected 32 Work Reports of Guangzhou Municipal Government from 1993 to 2024, to establish the corpus. The themes of the report were represented by the nouns

and verbs which represent the concrete and abstract objects and taken actions. Text clustering and correspondence analysis were used to explore the diachronic changes of themes represented by nouns and verbs. The results showed that Work Reports were divided into five phases which are 1993-2000, 2001-2008, 2009-2012, 2013-2019 and 2020-2024 respectively. The result of correspondence analysis demonstrated that some content words occur more frequently in some reports than in others. The keywords of Work Reports at each phase showed the changes of its focus. Lastly, we extract the thematic words of Work Report at each phase from the perspective of quantitative linguistics. The intersection of thematic words at each phases reveals the consistent and stable focuses of report over years, such as “发展 (development and develop)”, “经济(economy)”, “教育(education)”, “改革(reform)” and “文化(culture)”. This paper explored the development and stability of focused themes of Work Reports of Guangzhou Municipal Government over 32 years.

Keywords: Work Report of Guangzhou Municipal Government, Text Mining, Thematic words, Diachronic change

Author Biographies: Renkui Hou, corresponding author, Associate Professor, School of Humanities, Guangzhou University; National Research Center for Language Service and Languages of the GuangDong-Hong Kong-Macao Greater Bay Area, Guangzhou, China; Applied Linguistics, E-mail: hourk0917@163.com. Jincheng Zhang, Master, School of Humanities, Guangzhou University, Guangzhou, China, Linguistics and applied linguistics, E-mail: 2112412080@e.gzhu.edu.cn.

一、引言

语言随着社会的发展、文化的变迁而发展变化。语言系统的各个子系统与社会发展联系有很大不同，词汇与社会的联系最为紧密，一般认为其发展变化速度最快（叶蜚声、徐通锵，2010）。随着社会新事物的产生、旧事物的消失，新词语出现频率不断增加，旧词语出现频率不断降低，直至消失。

作为一种典型的、每年更新的政治话语，各级《政府工作报告》中词语和词语属性的历时演变是一种什么状况呢？这值得我们从社会发展、语言演变的角度去进行分析。

（一）文献综述

多个世纪以来，语言的历时演变一直被世界各地的语言学家所关注，同时也是传统语言学研究的一个重点。早些年，人们更多的关注语言中某个特定语言单位，如某个语音或者某个词语，以及某个特定语言结构的历时演变。Lieberman 等（2007）考察了过去 1200 年来的英语动词的规则化过程，以及词语使用频率对规则化过程的影响。Baker (2011)在研究 20 世纪英国英语变化时，考察了词语使用频率和意义的历时变化。Campos 等（2020）建立了一个基于语料库的方法来自动计算语言的历时距离，分别计算了英语、西班牙语和葡萄牙语在不同时期的距离。

人们对政治话语中的语言运用特点研究也由来已久 (Van Dijk 1997)。如 Savoy (2018) 考察了 2016 年美国总统初选候选人的言语风格和修辞特点, Yu (2013) 考察了政治话语中语言使用和性别之间的相互关系, 结果发现, 相对于男性政治人物, 女性政治人物较多使用情感词语和人称代词。

同时, 也有一些研究使用计量方法来考察政治话语中语言的历时演变。Jiang 等 (2020) 以词汇丰富度和词汇复杂度为计量指标, 研究结果发现, 英国女王的话语在过去的 200 年发生了很大的变化, 慢慢的变得越来越平民化。Savoy (2015) 分析了从 1790 年到 2014 年来的词汇增长特点, 据此对社会的发展变化进行了分析。Savoy (2017) 考察了政治话语的语言风格和修辞的历时演变, 结果发现随着时间的发展句子变短, 词语的复杂度降低等, 都是为了便于听众的理解。

近年来, 对我国政治话语中语言运用特点的研究也引起学者的兴趣。胡开宝和田绪军 (2018a) 采用语料库方法, 从高频词、关键词的应用等角度, 分析了中国外交话语英译本所塑造的中国外交形象。胡开宝和田绪军 (2018b) 分析了 2000-2016 年《政府工作报告》英译本的语言特征及其文本效果。澎湃新闻借助国产大模型分析能力, 使用文本向量化的数据分析方法, 抽取了 2026 年国务院《政府工作报告》的关键词——“发展”、“建设”和“推进”, 分析了 2026 年国家关注的重点。同时, 进一步回顾了 1978-2026 年《政府工作报告》中的关键词, “发展”、“建设”和“经济”等词语一直是历年报告中的高频词, 提出了“常青词”的概念, 分析了历年报告中稳定不变的关注点¹。

刘熠和王晓红 (2025) 分析了国内政治话语的研究现状, 指出政治话语研究 学科交融多元化、宏观与微观并行的特点。宏观研究主要集中于政治学、新闻传播学等领域, 微观研究主要集中于语言学领域, 从认知语言学、批评话语分析及语用学等视角入手, 挖掘语言背后的政治、权力因素。

Hou et al. (2020) 基于分句和词语之间的关系考察了政府工作报告中语言的历时演变, 结果发现拟合两者之间关系的函数参数可以区分不同时期的政治话语。Hou et al. (2022) 基于 Menzerath-Altmann 定律不仅考察了汉语政治话语的历时演变, 同时也对同一时期政治话语在不同地域的变体进行了对比, 同时考察了政治话语的历时演变与共时差异。

国外对政治话语的共时研究主要分析了不同政治人物话语风格的不同, 而历时研究主要集中于语言运用特点的差异, 如词汇丰富度、词汇复杂度等, 并未考察历年来政治话语中展示社会关注点的具体词语的变化。而国内关于政治话语研究或者较少从语言学角度进行分析, 或者从语言学角度进行的分析多为共时研究, 分析政治话语中的高频词、关键词等, 较少见的历时研究考察了历年国务院《政府工作报告》中的“常青词”, 没有关注具体词语的发展变化, 或者是基于计量语言学定律考察了政治话语中拟合参数的发展变化。另外较多 分析了《政府工作报告》外语译本语言特点的分析, 并为关注汉语《政府工作报告》本身的语言分析。

与上述研究不同, 本文以 30 年的《广州市政府工作报告》为研究对象, 聚焦于反映政府和社会关注点的词语的演变, 从而观察社会的发展变化。

(二) 研究问题和研究方法

¹ https://www.thepaper.cn/newsDetail_forward_32652126, 2026 年 6 月 10 日访问。

基于已有研究较少从历时角度考察地方政治话语具体词语出现频率的变化, 本文采用计量方法, 基于语料库和文本挖掘技术, 考察 30 多年来的《广州市政府工作报告》反映社会发展变化的词语的历时演变。

从广州市人民政府网站收集 1993-2024 年政府工作报告建立语料库, 以名词和动词表示文本, 通过文本聚类 and 对应分析观察历年来词语是否具有变化, 根据变化的幅度, 是否可以将报告分为不同的阶段。

将前一个阶段的报告为参照语料, 提取当前阶段报告的关键词, 从而考察和验证词语的变化、政府和社会关注点的变化。另一方面, 基于计量语言学方法, 抽取每个阶段报告的主题词, 通过对比, 提取各阶段共同的主题词, 观察变化中的“不变”。

以日本同志社大学开发文本挖掘软件 MTMineR 为工具建立文本向量, 文本聚类、对应分析及主题词的提取都在 R 语言中实现, 关键词使用日本早稻田大学开发的索引和文本分析工具 Antconc (A freeware corpus analysis toolkit for concordancing and text analysis) 进行抽取。

二、语料库建立

广州市政府工作报告由市长代表广州市人民政府向广州市人民代表大会 (及其常务委员会) 汇报, 并同步提交给政协委员和列席人员审议, 完成宪法规定的“政府向人大述职、接受监督”的法定环节。我们可以通过语言分析考察报告关注的重点, 从中观察社会的发展变化。

从广州市人民政府网站人工收集 1993 年至 2024 年共 32 年的广州市政府工作报告¹, 删除其中空白行、注释和名词解释等信息, 仅保留报告文本。每一个报告对应一个文本文件, 建立语料库。

使用日本早稻田大学开发的多语言词法分析系统 TagAnt (Multi-language segmenter and Part-of-Speech tagger, Anthony 2024) 对报告文本进行分词和词类标注。该系统基于 SpaCy 自然语言处理框架², 具有工业级的可靠性, 对中文文本进行分词和词性标注时, 加载预训练中文模型 zh_core_web_sm-3.4.0。Spacy 的中文模型采用了先进的神经网络算法, 在标准中文语料库上进行了大规模训练, 具有较高的分词和词性标注准确率。

对语料规模进行统计, 最长为 1998 年报告, 文本长度为 12739 个词语, 最短为 2015 年报告文本, 文本长度为 7925 个词语, 报告文本平均长度为 10515 个词语, 所有报告文本总长度为 336486 个词语。

三、报告词语历时变化

政府工作报告谈论的是政府、社会和人们关注的事物, 具有明确的主题, 集中在信息的核心。一般情况下, 文本的主题使用语言中的名词、动词和形容词来表示, 名词代表报告文本中具体和抽象的事物, 形容词代表事物的属性, 动词代表对名词表示的事物施加的动作 (Popescu et al.

¹ <https://www.gz.gov.cn/zwgk/zjgb/zfgzbg/index.html>, 最后访问时间为 2025 年 2 月 23 日。

² <https://www.laurenceanthony.net/software/tagant/releases/TagAnt211/help.pdf>, 最后访问时间为 2026 年 6 月 11 日。

2009)。我们选择在所有报告中出现的名词和动词表示文本¹，计算每个词语的相对出现频率来建立文本向量。

（一）文本聚类

聚类分析是研究物以类聚的一种有效方法，可以用来初步探测文本之间的关系，是一种探索性数据分析技术。事先不知道类别的个数与结构，仅根据在数据中发现的对象本身的信息和对象之间关系的信息，将对象进行分类，使得同一类簇（cluster）中数据点之间的相似性高于不同类簇之间数据点的相似性。一般来说，一个理想的聚类结果可以这样定义，同一类簇中的数据具有最大的相似性，不同类簇之间的数据具有最大的差异性（Manning & Schutze 1999）。

我们事先并不清楚不同年份的政府工作报告所用的表示主题的名词和动词之间的相同和不同之处，因此，为了探索历年报告之间的关系，聚类分析符合我们的需要。本文选择凝聚层次聚类（Agglomerative Hierarchical Clustering）来对报告进行分析，该方法首先将单个文本作为个体簇，然后每一步合并两个最接近的类簇，直到只剩下一个类簇。该聚类方法的关键是计算文本之间的距离，以及类簇之间的距离。我们选择相对熵（Kullback-Leibler Divergence）计算文本之间的距离，选择离差平方和计算类簇之间的距离。已有实验结果表明，采用这两种方法计算两类距离，可以取得较好的聚类结果（Dhillon et al. 2003）。

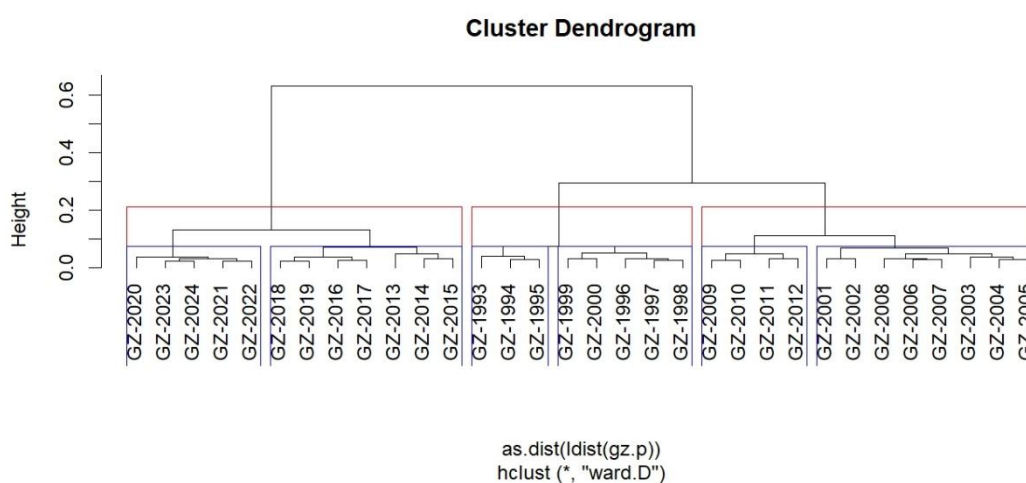


图1 以名词和动词表示广州市政府工作报告的聚类结果

(Figure 1: Clustering Results of the Guangzhou Municipal Government Work Report Represented by Nouns and Verbs)

以名词和动词表示文本，聚类结果以树状图（Cluster Dendrogram）表示，如图1。树状图为一个倒立的树，树中叶子节点表示文本，内部节点表示文本形成的类簇，两个文本或类簇的祖先节点的高度表示两个文本或者类簇之间的距离，高度越高表示两个类簇或者两个文本之间的距离越大，相似度越小。

¹ 经过了 TagAnt 词法分析之后，词类标记为 NN（一般名词），NR（人名、地名、机构名），VV（动词）的词语。

从图 1 可以看出, 32 年的广州市政府工作报告按照时间顺序大致可以分为三个大的阶段, 第一个阶段为 1993 年至 2000 年, 共 8 年, 其可以分为两个小的阶段, 分别为 1993-1995 以及 1996-2000, 这两个类簇之间距离较小; 第二个阶段为 2001 年至 2012 年, 同样可以分为两个小的阶段, 分别为 2001-2008 和 2009-2012; 第三个阶段为 2013 年至 2024 年, 其同样分为 2013-2019 和 2020-2024 两个小的阶段。

每个阶段报告的关注点应该具有系统的相似性, 而不同阶段的报告之间具有系统的差异性。每个大的阶段下的两个小阶段报告之间的差异要小一些, 相似度更大些。第一阶段和第二阶段报告之间的距离远小于两者与第三阶段报告之间的距离, 这说明, 第一阶段与第二阶段之间关注的主题虽有变化, 但没有那么大, 而这两个阶段到第三个阶段则有了非常大的变化, 这说明从 2013 年开始广州市政府工作报告关注的主题和社会热点等发生了较大的变化, 这与社会的发展具有一致性。

(二) 对应分析

图 1 所示的聚类结果只能说明报告之间存在差异, 却无法显示差异体现在哪里。我们使用对应分析来初步观察报告和表示主题的名词与动词之间的相互关系, 即某些名词或动词在某些报告中出现频率相对较高, 而在另外一些报告中出现频率则相对较低。

对应分析提供了一种使用低维图形探索变量和文本之间关系的直观方法 (Baayen 2008)。两个报告所关注主题的差别越大, 则两个文本之间的距离越大, 两个文本在图形中的距离就越远。同样, 差别较大的变量, 在图形上的距离较远。对应分析能够确保低维图形中文本和词语之间的位置关系尽可能地反映文本和词语之间的相互关系, 如果某个词语在某个阶段的报告文本中相对较多出现, 而在其它阶段的文本中相对较少出现, 则它和前者在图形中距离就很近, 甚至重合, 与后者之间的距离则较远。

我们可以通过对应分析的结果观察不同阶段报告文本之间的距离、不同实词之间的距离以及报告文本和词语之间的相互关系。

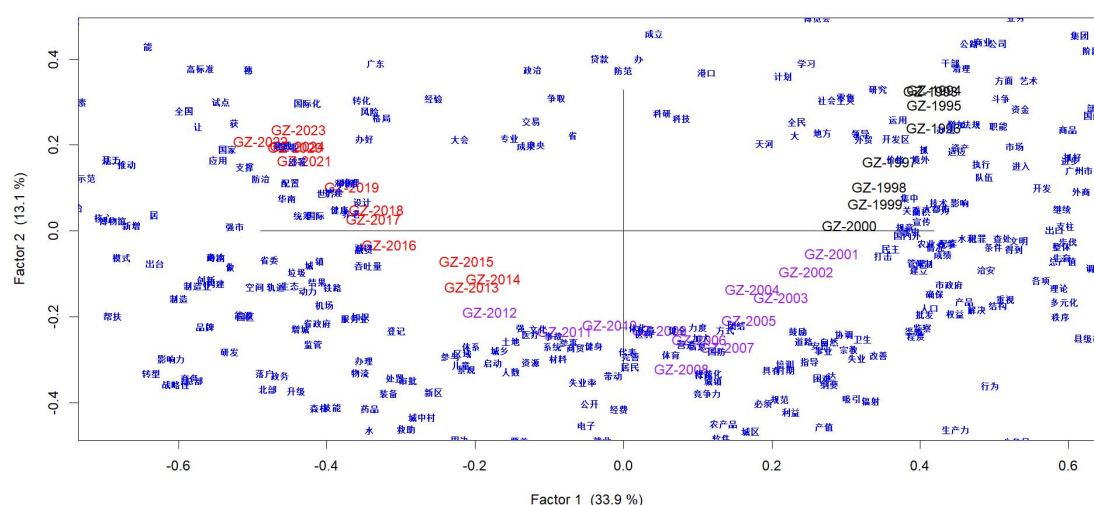


图 2 以名词和动词表示广州市政府工作报告的对应分析结果
(Figure 2: Correspondence Analysis Results of the Guangzhou Municipal Government Work Report Based on Nouns and Verbs)

我们使用 R 语言中 *LanguageR* 包中的函数 *corres.fnc* 对以名词和动词表示的报告文本进行对应分析，对应分析结果如图 2 所示¹。从图 2 中可以看出，报告文本之间的距离与图 1 基本一致，都是可以分为三个大的阶段。对应分析和聚类分析所得结果一致，两者相互验证，展示了所得结果的可靠性。这说明以动词和名词来表示文本，观察报告关注的社会热点在 32 年的变化，是可行的。

更重要的是，图 2 也展示了报告文本和实词（名词与动词）之间的相互关系，如“外贸、市场、商品、进步、适应”等词语在第一阶段报告文本中出现频率相对较高，“电子、竞争力、资源、完善、商贸、鼓励”等词语在第二阶段报告文本中出现频率相对较高，而“推动、试点、国际、高标准、健康、养老、生态、创新”等词语在第三个阶段报告文本中出现频率相对较高。

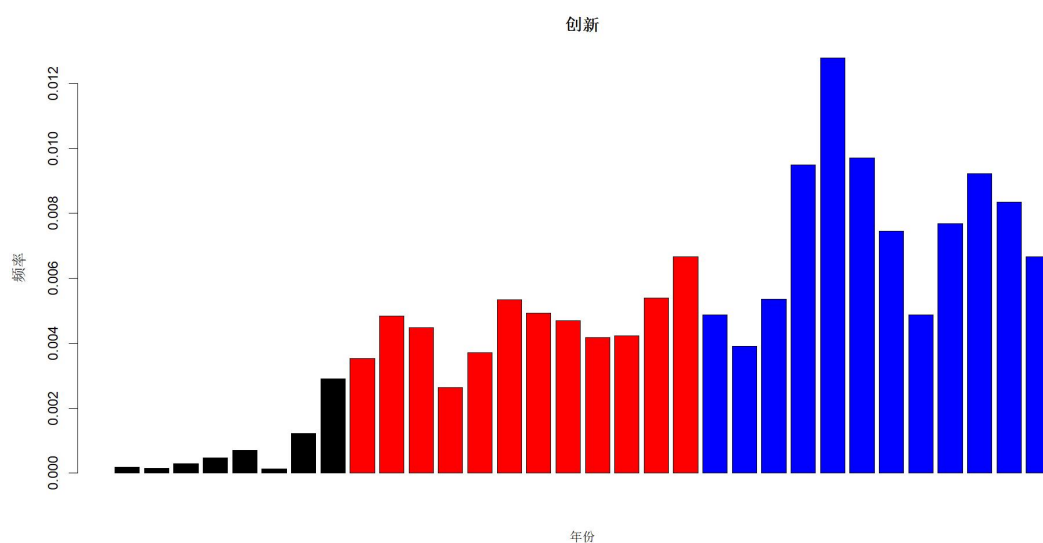


图 3 “创新”在历年报告中出现频率的分布图

(Figure 3 Distribution of the Frequency of the Word "Innovation" in Annual Work Reports)

我们以“创新”为例，使用柱状图展示其在报告中的出现频率分布，如图 3 所示。从中我们可以看到“创新”在第一阶段（1993-2000）政府工作报告中的出现频率要远远低于其在第二和第三阶段报告中的出现频率，而整体上看在第二阶段的出现频率也小于在第三阶段的出现频率。这说明随着时间的发展，政府工作报告更加注重包括“科技、技术”等各领域的“创新”。

（三）关键词分析

为了考察政府工作报告中表示关注点的名词和动词的变化，我们使用日本早稻田大学开发的软件 AntConc 来计算每一阶段相对前面一个阶段的关键词（Anthony 2024），从中更直观地观察这种变化。关键词是指在当前语料库中出现频率较高，而在参照（Reference）语料库中出现频率较低的词语，因此它可以反映不同阶段报告中关注点的差异。而主题词是在当前语料库中出现频率较高的名词和动词，只能反映词语在当前语料库中的出现情况，不与其它语料库对比。

使用 AntConc 计算关键词时，Log-Likelihood Measure 设置为“Log-Likelihood （4-term）”，

¹ 为了方便观察报告和主题词之间的关系，图 2 中仅显示小部分的名词和动词。

阈值设置为“ $p < 0.05$ (3.84 with Bonferroni)”, Effect Size Measure 设置为“Dice”。

我们首先将第一阶段, 即 1993-2000 年广州市政府工作报告, 作为参照语料库 (Reference Corpus) 来计算第二阶段, 即 2001-2008, 的关键词, 依据 *Keyness (Likelihood)* 值对关键词进行排序。从中可以看到, 相对于第一个阶段, 第二个阶段更加注重“创新、体系、社区、公共、提升、推进、中心镇、就业、信息化、农村、完善、会展、物流、研发、建设”等¹, 其中, “中心镇、会展、研发”以及“品牌、应急、城中村、亚运会、平台”等在第一个阶段报告中都没有出现。从图 3 中也可以看到“创新”这个词语的出现频率在当前阶段有了大幅提升。

然后以第二阶段的报告为参照语料, 计算第三个阶段, 即 2009 到 2012 年广州市政府工作报告的关键词, 同样依据 *Keyness (Likelihood)* 来进行排序。我们可以看到, 相对于第二阶段报告, 第三阶段的 4 年广州市政府工作报告更加注重“低碳、幸福、民生、提升、预计、保障性、服务、转型、危机、世界、城乡”等, 其中“低碳、幸福、保障性、危机”等在第二阶段 8 年的报告中都没有出现, 仅在当前阶段提出。

接着将第三阶段 (2009-2012) 的广州市政府工作报告作为参照语料库, 计算第四阶段, 即 2013-2019 年报告的关键词, 同样按照 *Keyness (Likelihood)* 值对关键词进行排序。相对于第三阶段, 第四阶段广州市政府工作报告更加注重“枢纽, 推进, 治理, 稳, 创新, 美丽, 推动, 互联网, 试点, 自贸, 清单, 试验区, 大学, 科技”等, 其中“稳、自贸, 清单”在第三阶段报告中都没有出现, 而“枢纽、治理、创新、推动、美丽”出现频率和分布年份都是大大增加, “推动”交通和“创新”“枢纽”建设, 关注“科技”“创新”, “创新”人才的培养, 各领域“试验区、试点”建设的“推进”, “美丽”乡村、“美丽”广州建设等。

最后, 以第四阶段七年报告 (2013-2019) 为参照语料库, 计算最后一个阶段 (2020-2024) 的关键词。关键词包括“大湾区, 高, 疫情, 全球, 推动, 大学, 智能, 数字, 枢纽, 打造, 政治, 防控, 试点, 融合, 国家, 实验室”等, 其中“大湾区, 实验室, 提质”等都是在这一阶段的报告中首次出现。从中可以看出, 全球疫情时期, 关注“疫情”防控, 并同时从“国家”视角关注“大湾区”建设, 推动“人工”“智能”和“数字”经济的“高”质量发展, 老城区和新领域的“出新、出彩”, 各领域的广深“双城”“联动”, “打造”“活力”“湾区”等。

从各阶段的关键词, 我们能够更加清晰的看到各个阶段报告关注点和所采取的举措的发展变化, 从中可以观察到社会的发展变化。

四、主题词抽取与分析

文本的主题使用名词、动词和形容词来表示, 但并不是每一个名词、动词和形容词都可以代表文本的主题, 只有一些特定的名词、动词、形容词能够体现文本的主题。Popescu 等 (2009) 认为一个文本的词频表中 *h* 点之前的动词、名词、形容词可以表示文本的主题, 它们能够反应文本的主题内容。

h 点是文献计量学中的一个概念, 用来表示一个学者的学术影响力, 后来被 Popescu (2006) 引入语言学, 并获得进一步发展。将一个文本中所有词语按照出现频次 (*frequency*) 来进行排序

¹ 我们仅列出其中的少数关键词, 来说明报告关注点和主题的变化。

(rank), 出现频次最高的排在第一位(rank=1), 其出现频率为 $f(1)$, 出现频次第二高的排在第二位(rank=2), 出现次数为 $f(2)$ 。依此类推, 随着 r 的增大, $f(r)$ 不断减小, 就会出现一个点 $r=f(r)$, 这个点就被称为 h 点。 h 点之前的动词、名词和形容词被称为主题词(thematic words)。考虑到政府工作报告的特点, 我们选择 h 点之前名词和动词作为主题词。

$f(r)$ 随着 r 的增加而降低, 是一个离散函数, 可能在这些点上不存在一个点使得 $r=f(r)$ 。在这种情况下, 使用下面的公式(1)计算 h 的值。

$$h = \frac{f(r_1) \times r_2 - f(r_2) \times r_1}{r_2 - r_1 + f(r_1) - f(r_2)}$$

公式(1)

其中 r_1 和 r_2 是两个相邻的词序(rank), $r_2 - r_1 = 1$, 而 $r_1 < f(r_1)$, $r_2 > f(r_2)$, $f(r_2) < f(r_1)$ 。

接下来我们以 1993 年广州市政府工作报告为例, 来展示如何计算文本的 h 点, 并抽取其主题词。首先统计每个词语出现频次, 建立 1993 年广州市政府工作报告中的词频表, 按照词语的出现次数进行排序, 如表 1 所示。

表 1 1993 年政府工作报告词频表¹

(Table 1 Word Frequency List of the 1993 Government Work Report)

Rank	词语	Frequency	Rank	词语	Frequency
1	的/DEG	215	16	在/P	45
2	和/CC	209	17	教育/NN	44
3	的/DEC	124	18	社会/NN	43
4	经济/NN	97	19	市/NN	43
5	了/AS	92	20	有/VE	41
6	建设/NN	77	21	城市/NN	39
7	要/VV	63	22	等/ETC	36
8	企业/NN	62	23	管理/NN	36
9	发展/NN	60	24	广州/NR	35
10	增长/VV	59	25	对/P	34
11	市场/NN	56	26	科技/NN	34
12	改革/NN	54	27	年/M	32
13	发展/VV	48	28	加快/VV	31
14	加强/VV	47	29	提高/VV	29
15	工作/NN	46	30	技术/NN	28

从表 1 中, 我们可以看出 h 值为 29, 因为当 r 为 29 的时候, $r=f(r)$ 。 h 点前共有 14 个名词(词类标记为 NN 和 NR), 6 个动词(词类标记为 VV), 它们分别是:

¹ 仅显示出现频次最高的前 40 个词语, 词语后面的英文字母表示该词语的词类

经济, 建设, 企业, 发展 (名词), 市场, 改革, 工作, 教育, 社会, 市, 城市, 管理, 广州, 科技

要, 增长, 发展 (动词), 加强, 加快, 提高

从中可以看到该年度工作报告关注的重点, 推动“经济”的“增长”, “加强”“加快”“改革”的步伐, “市场”的“建设”, “科技”和“企业”的“发展”等。

这个抽取主题词的计量语言学方法 (h 点方法), 计算复杂度较低, 具有较高的可解释性, 比较适合语言的历时研究。可以通过各阶段主题词的对比, 来确定各阶段共同的主题词, 以观察报告所关注主题的稳定性的。

从第 3 部分文本聚类、对应分析和关键词, 我们观察到了广州市政府工作报告关注点的历年变化, 那么 32 年来的有没有哪些关注点是不变的呢? 接下来采用以上方法来分别抽取这 5 个阶段的主题词, 观察恒定不变的关注点。

每个阶段主题词见附录中表 1, 5 个阶段的主题词共有 108 个。有些词语在某个阶段的报告中是主题词, 而在其它某些阶段的报告中可能不是主题词。那么, 这些词语就不是每个阶段都被关注的。如果某词语在每个阶段都是主题词, 那么就可以认为它是被 5 个阶段共同关注, 可以认为是稳定的。我们通过计算 5 个阶段主题词集合的交集来确定这些词语, 结果显示共有 32 个词语在 5 个阶段中都是主题词 (如下), 因此可以认为它们是 5 个阶段报告中不变的关注点。

“工作/NN”, “建设/NN”, “企业/NN”, “经济/NN”, “管理/NN”, “加强/VV”, “发展/NN”, “市/NN”, “社会/NN”, “城市/NN”, “改革/NN”, “发展/VV”, “广州/NR”, “教育/NN”, “加快/VV”, “增长/VV”, “推进/VV”, “工程/NN”, “产业/NN”, “开展/VV”, “水平/NN”, “项目/NN”, “环境/NN”, “政府/NN”, “完善/VV”, “服务/NN”, “促进/VV”, “文化/NN”, “中心/NN”, “机制/NN”, “国际/NN”, “实施/VV”

我们选择“发展”、“改革”和“文化”作为例子, 使用柱状图, 如图 4, 展示它们在历年广州市政府工作报告中的出现频率, 直观地观察它们的出现频率有没有随着时间的推移发生较大的变化, 是否具有稳定性。

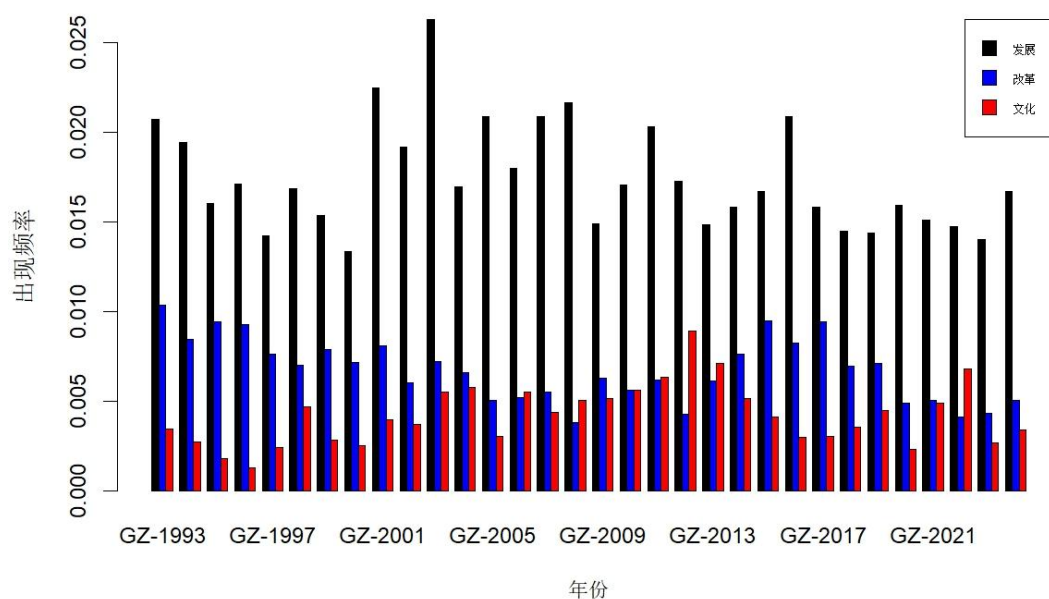


图 4 5 个阶段稳定不变的主题词出现频率柱状图（在词性标注时，“发展”为兼类词，这里没有考虑词类的差异）

(Figure 4 Bar Chart of the Frequency of Stable Subject Terms across Five Stages (Note: In part-of-speech tagging, "development" is a multi-category word; differences in word categories are not considered here))

从图 4 中我们可以看到，虽然三个主题词在某年报告中出现频率存在差别，可是某个特定主题词，如“发展”，在历年报告中的出现频率是基本稳定的，或者是关注某些事物的发展，如“民营经济发展”、“开放型经济发展”、“经济贸易发展”、“农村经济发展”“边远地区发展”等，或者是发展某事物，如“发展经济”、“大力发展新一代信息技术”、“大力发展社区卫生服务”、“大力发展跨境电子商务”、“大力发展海洋经济”、“大力发展应用型专科教育”等。

再如“改革”，历年关注多个领域的改革，如最早 1993 年报告的“科技体制改革进一步深化”，到 2024 年的“国家出台支持南沙放宽市场准入与加强监管体制改革、开展土地管理综合改革试点意见”，以及历年来的“继续深化城市管理体制改革”、“深化行政管理体制改革”、“深化卫生管理体制”、“积极推进农村经济体制改革”、“深化投资体制改革”等。总之，“改革”作为历年政府工作报告的关注点一直稳定不变。

再看“文化”，虽然其在每年报告中的出现频率要低于“发展”和“改革”，可是在每年报告中也都得到了体现，如最早 1993 年的“改造和新建了一批文化娱乐设施和场所，加强了对社会文化市场的管理。对外文化活动蓬勃发展”、“要继续深化文化艺术体制改革”，到 1994 年的“推进了文化体制改革”，到 2024 年的“推进文化强市建设，增强城市文化综合实力”，“坚定文化自信自强”，“提升历史文化名城影响力”，再如其它年份的“推进农村和社区文化建设，丰富群众文化生活”，“推动文化繁荣发展”，“推动文化创新，培育文化产业”等等，无不显示出报告对与

“文化”相关领域事业的重视。

上面抽取的是 5 个阶段报告的共同主题词进行的分析，接下来，我们抽取每年报告的主题词，发现其交集只有四个词，即这四个词语在所有的年份中都是被关注的重点。

“建设/NN”“发展/NN”“广州/NR”“加快/VV”

同样使用柱状图展示这些词语在每年政府工作报告中的出现频率，如图 5 所示。从中能够观察到它们在历年报告中出现频率的稳定性，可以看出它们每年都是报告的关注点。每年的政府工作报告都关注“广州”各方面，如环境、廉政、法治的“建设”和“发展”等，同时也在“加快”推进这种建设和发展。

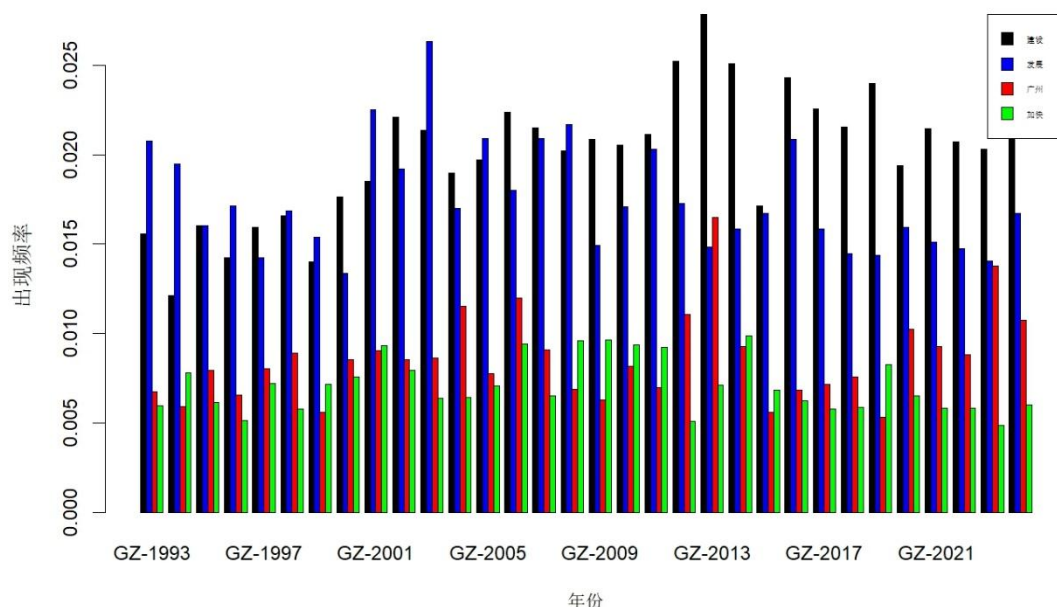


图 5 每年恒定不变的主题词的出现频率

(Figure 5 Occurrence Frequency of Invariable Subject Terms in Each Year)

五、结论

本文以文本挖掘为手段考察了广州市政府工作报告中词语的发展。首先以名词和动词来表示报告文本，对其进行聚类分析，发现 32 年的广州市政府工作报告大致可以分为三个大的阶段和五个小的阶段，第一个阶段为 1993 年到 2000 年的 8 年，第二个阶段为 2001-2008 年，第三个阶段为 2009 年到 2012 年的 4 年，第四个阶段为 2013 年到 2019 年的 7 年，最后一个阶段为 2020 年到 2024 年。不同阶段出现的表示报告主题的名词和动词的出现存在着差别。

对以名词和动词表示的报告文本进行对应分析，对应分析结果显示出报告文本和主题词之间的相互关系。从该相互关系中，能够看出每个阶段报告所关注的主题词是不同的。如第一阶段关注“市场”、“商品”、“外商”等，而第五阶段则关注“健康”、“生态”等。

关键词的分析，更能直观反映出特定阶段报告，相对于前一阶段报告，所关注的社会事物、以

及政府采取的措施等所发生的变化。这里建立的以聚类分析、对应分析、关键词计算观察报告主题变化的方法具有普适性。

最后采用计量语言学方法，计算了每个阶段报告的 h 点，抽取了每个阶段报告的主题词，词频表中位于 h 点之间名词和动词，对每个阶段的主题词集合进行了对比，找出了它们的交集。我们认为，交集中的主题词是每个阶段报告共同关注的主题，如“发展、建设、改革、教育、文化”等，是稳定不变的。时间的变化，社会的发展，报告中表示主题的词语在发生变化，但也有一些表示主题的词语是稳定的，即每个阶段中主题词的交际。

基金项目：本文系广东省社科项目“粤港两地政治话语历时演变与共时差异的计量研究”（项目编号：GD24CZY03），“粤港澳大湾区语言资源库建设研究”（项目编号：GD22XZY03）成果。

Conflicts of Interest: The authors declare no conflict of interest.

ORCID

侯仁魁 ^{ID} <https://orcid.org/0000-0003-2510-6277>

张锦程 ^{ID} <https://orcid.org/0009-0001-9048-0583>

References

- 胡开宝，田绪军.(2018a).中国外交话语英译中的中国外交形象研究——一项基于语料库的研究.中国外语.15(6): 79-88.
- [Hu Kaibao, Tian Xujun. (2018a) "A Corpus-based Study of the Image of Chinese Diplomacy in the English Translations of Chinese Discourse on Diplomacy", *Foreign Language in China*, 15(6): 79-88. DOI: <http://doi.org/10.13564/j.cnki.issn.1672-9382.2018.06.012>.]
- 胡开宝，田绪军. (2018b) 《政府工作报告》英译文本的语言特征与文本效果研究——一项基于语料库的研究. 外国语文. 34(5): 1-11.
- [Hu Kaibao, Tian Xujun. (2018b) "A Corpus-based Study of the Linguistic Features and Textual Effects of the English Translation of the Government Work Report". *Foreign Languages and Literature*. 34(5): 1-11.]
- 刘熠，王晓红. (2025). 基于文献可视化分析的国内政治话语研究现状与启示. 辽宁大学学报（哲学社会科学版）. 53（3）: 119-126.
- [Liu Yi, Wang Xiaohong (2025) "A Visual Analysis of Bibliometrics of the Research on Political Discourse in China and its Implications", *Journal of Liaoning University (Philosophy and Social Science)*. 53（3）: 119-126. DOI : <http://doi.org/10.16197/j.cnki.lnupse.2025.03.006>.]
- 叶蜚声，徐通锵（著），王洪君，李娟（修订）. (2010) 语言学纲要（修订版）. 北京：北京大学出版社.
- [Ye Feisheng, Xu Tongqiang (2010) "Outline of Linguistics (Revision Version, by Wang Hongjun and Lijuan)" Beijing University Press, Beijing, China]
- Anthony, L. (2024). TagAnt (Version 2.1.1) [Computer Software]. Tokyo, Japan: Waseda University.

<https://www.laurenceanthony.net/software/TagAnt>

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University.

<https://www.laurenceanthony.net/software/AntConc>

Baker, P. (2011). Times may change, but we will always have money: diachronic variation in recent British English. *Journal of English Linguistics*, 39(1), 65–88. DOI: <https://doi.org/10.1177/0075424210368368>.

Baayen R, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Cacoullos, R. T. (1999). Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. *Language variation and change*, 11(2), 143-170. DOI: <https://doi.org/10.1017/S095439459911202X>.

Campos, J., Otero, P., & Loinaz, I. (2020). Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, 26(4), 433-454. DOI: <https://doi.org/10.1017/S1351324919000378>.

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). Enhanced word clustering for hierarchical text classification. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 191-200. DOI: <https://doi.org/10.1145/775047.775076>.

Hou, R., C.-R. Huang & K. Ahrens. (2020). Language change in Report on the Work of the Government by Premiers of the People's Republic of China. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. Pp.100-111. DOI: <https://aclanthology.org/2020.paclic-1.12.pdf>.

Hou, R., C.-R. Huang & K. Ahrens. (2022). Regional varieties and diachronic changes in Chinese political discourse. *Humanity & Social Science Communication*. 9: 466. DOI: <https://doi.org/10.1057/s41599-022-01488-8>.

Jiang X, Jiang Y & Hoi C K W (2020) Is Queen's English Drifting Towards Common People's English. –Quantifying Diachronic changes of Queens's Christmas Messages (1952-2018) with Reference to BNC. *Journal of Quantitative Linguistics*. DOI: <https://doi.org/10.1080/09296174.2020.1737483>.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Savoy, J. (2015). Vocabulary Growth Study: An Example with the State of the Union Addresses. *Journal of Quantitative Linguistics*. 22(4): 289-310. DOI: <https://doi.org/10.1080/09296174.2015.1106270>.

Savoy, J. (2017). Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries. *Glottometrics*. 38:55-76.

Savoy, J. (2018). Analysis of the style and the rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities*. 33(1): 143-159. DOI: <https://doi.org/10.1093/llc/fqx007>.

Yu, B. (2013). Language and gender in congressional speech. *Literary and Linguistic Computing*. 29(1) :118-132. DOI: <https://doi.org/10.1093/llc/fqs073>.

Van Dijk TA. (1997). What is political discourse analysis? *Belgian Journal of Linguistics* 11: 11–52.