国际视野

日本计量文体学发展历程: 从文章心理学走向数字人文

王子睿

摘　　要　研究通过文献分析法,从时间演变和内容分类两个角度,对日本计量文体学(stylometry)的发展历程进行系统梳理,并通过典型研究总结其特点。日本计量文体学的发展路径呈现出鲜明的本土化特征:20 世纪 30 年代,日本计量文体学在几乎未受西方理论影响的情况下,萌芽于心理学领域;60 年代前后,计量国语学会的成立推动了语言计量方法的普及;90 年代后,随着计算机和统计学的进步,该领域进入成果丰硕的繁荣期。在这一漫长积累过程中,日本计量文体学逐步形成了以计量为基础,语言学、心理学与文学等多个领域交叉,文理融合、问题导向的文学研究特色,积累了大量具有实用价值的研究成果,为国内文学研究领域提供了重要的参考。

关 键 词　数字人文;　计量文体学;　日本;　人文计算

分 类 号　H052;　I313.074;　TP391.1

作者简介　王子睿,北京语言大学博士研究生,Email: oushieiwzr@gmail.com。

0　引言

在 2013 年莫雷蒂(Franco Moretti)和乔克斯(Matthew L.Jockers)分别提出了远读法(distant reading)、常量分析法(Macroanalysis)后[1,2],计算方法的应用在世界文学研究领域引起瞩目,但实际上,使用计算进行文学探索的历史可以追溯到 19 世纪末通过统计莎士比亚作品的文体特征验证作者身份的研究。而据格日贝克(Peter Grzybek)的研究,1897 年卢托斯瓦夫斯基(Wincenty Lutosławski)最早在将这种计算和文学结合的研究方式明确定义为“计量文体学”(stylometry)[3]。与彼时不同,今日的计量文体学研究依赖于计算机技术,还纳

基金项目: 本文系 2024 年北京语言大学研究生创新基金(中央高校基本科研业务费专项资金)项目“文学与翻译计算批评方法的梳理及应用研究”(24YCX072)的阶段性研究成果;2024 年国家社科基金一般项目“人机互动日语口译教学模式建构与研究”(24BYY074)的阶段性研究成果。

入网络文学[4]等多种研究对象,被视作数字人文的重要分支。国内文学领域的文体计量研究起步稍晚,在"数字人文研究中国化"[5]的趋势和目标之下,别国研究的发展历程和学术特色值得总结、反思和借鉴。

起步于20世纪30年代的日本计量文体学,与欧美的发展轨迹有所不同,最早发源于心理学且被当时的研究者命名为"文章心理学"①,20世纪末伴随着信息技术的迅速发展和不断接受欧美现代统计学的理念,迎来飞速发展期。在日本计量文体学的发展中,涌现了一批以村上征胜、金明哲等出色研究者为核心的研究团队,诞生了以计量国语学会、行动计量学会、汉字信息处理学会、人文与计算机学会为代表的众多使用数字方法探索文学、语言学等的学术交流平台。日本计量文体学的大量学术累积为日本人文研究提供了一种客观数据与逻辑论证并重的实验式文学批评思路。但就目前国内的文献来看,对日本计量文体学的介绍、挖掘工作还不够深入,尤其是日本学者接受数字技术作为文学研究方法后解决了哪些传统文学研究的问题,是否提出了具有本土特色的研究方法等方面,需要进行梳理和总结,本文将在对日本计量文体学的发展历程进行回顾的基础上,通过对其重要研究成果进行分析案例可以促进对文学与文体研究产生新的思路。

1 计量文体学研究简介

当前的计量文体学(stylometry/computational stylistics),是基于现代统计科学知识、利用自然语言处理技术对各种文本②进行文体特征(style feature)测量和分析的研究领域,以解决关于作者、文体以及文本之间的诸多问题,被视作数字人文的一个重要分支③。计量文体学最早的和最常见的研究对象都是文学文本,研究方法上区别于基于经验式、主观内省式的传统文学研究,以定量研究为典型特征,通常定量、定性方法相结合。与计量文体学在研究方法上具有高度相似性的,还有语料库文体学、计算风格学、测量文体学以及语料库翻译学等领域。

研究者不满足于用感性解决文学问题,而使用计算方法来确定作者身份及其文体特征,带来了这一领域的许多经典案例。最早的手动计算探索可以上溯到Mendenhall(1887)的研究,Mendenhall将单词长度的分布作为识别一个作者文体的重要依据[6]。其他典型研究有如莎士比亚著作的文体识别问题,苏联著名文学作品《静静的顿河》的作者验证[7],对《红楼梦》前80回和后40回作者的判断等[8,9]。日本的计量文体学的开端同样可以追溯到1935年波多野完治对测量了谷崎润一郎和志贺直哉的文体特征的测量[10]。

以上经典案例也塑造了人们对于计量文体学的基本认知。不过,针对计量文体学本身的研究显示,这一领域内部存在着丰富的差异性,本文整理了斯塔马托斯(Stamatatos)、尼尔(Neal)等人、拉古蒂纳(Lagutina)等人、萨沃伊(Savoy)的研究成果后[11-14],提出了"二次分类法"(图3)。首先根据研究的目的不同,将计量文体学的研究分为"工学导向"和"人文导向":前者以解决构建识别模型、提高识别精度等工学问题为最终目的,后者以解决文学、语言学阐释等人文问题为最终目的,二者有相互交叉的情况。其次再从研究对象来分类:

① 一般认为1935年波多野完治《文章心理学》第一版的发行是日本计量文体学发展的起始点。
② 这里的"文本"包括文学类和非文学类,目前已经出现了判别AI写作和人类写作的相关研究。
③ 关于计量文体学的对应英文,笔者曾对Web of Science核心期刊集进行了计量分析,使用"stylometry"一词的论文数量较多,作者的学科背景也最为广泛。

“工学导向”中根据数据集对象以及分类要求的不同,可以分类为: (1) 闭集归属问题,(2) 开集归属问题,(3) 作者验证问题 “人文导向”可以分类为: (4) 个人文体问题,(5) 类型文体问题,(6) 文体变化问题①,但从在实践中往往会有两者交叉的现象,例如在作者验证问题中同时存在个人文体的探讨。

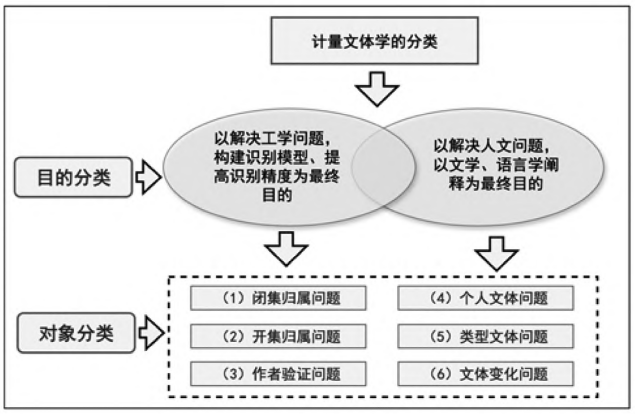

图 1　计量文体学的研究课题

2　国内日本计量文体学研究现状

国内学者已经对日本的计量文体学进行了一定介绍,魏育邻在著书《日语文体学》中以语言研究者的视角,将日语文体学分为“文学文体论”和“语言学的文体论”两类,认为前者关注文学批评和文学研究,而后者则奉行语言学“描写语言形态”的使命,但作者并没有对 1990 年后日本的计量文体学的最新研究成果进行介绍[15]。施建军在其著书《计量文体学导论》在介绍了日本部分研究成果的基础上,用中文文学作品的实际案例详细讲解了计量文体学的实验方法,并指出目前我国国内计量文体学的发展和世界先进水平还存在一定差距,指出国内相关研究在古典文献的作者识别方面未有突破、成果数量较少、尚未找到测量汉语的有效测量指标[16]。毛文伟将日本计量文体学归类为语料库个人文体研究和语料库类型文体研究,认为当前日本相关研究存在研究视野狭窄、理论研究不够深入、基础数据亟需充实的问题[17]。此外,一些学者也吸收了日本学者的计量方法对日本文学进行探索研究。例如李月平和毛文伟使用名词比、MVR 等 7 个指标测量了夏目漱石的 13 篇短篇小说[18]。李文平等基于依存距离和词汇丰富度等指标芥川龙之介的儿童文学与其他文学作品进行了系统比较,发现两者之间存在明显的文体差异[19]。

以上为数不多的研究,对于日本计量文体学的发展脉络及研究成果的介绍并不充分,对其研究的特点,尤其是可供中文文学研究参考的研究问题、测量指标方面,也尚需更深入的梳理与总结。因此,本文在吸收以上研究成果的基础上,将全面梳理至今为止的日本计量文体学发展脉络,对其主要研究成果进行整理、分析。

① (1) 闭集归属问题: 在一个固定的、已知的作者集合中,确定某篇作品的作者是谁。(2) 开集归属问题: 作品创作者未知。(3) 作者验证问题: 判断某篇作品是否由特定的作者撰写。(4) 个人文体问题: 研究某个特定作者的写作风格和特点。(5) 类型文体问题: 识别和分析不同类型或类别的文本风格,如小说、诗歌、新闻等。(6) 文体变化问题: 研究同一作者或者同一文体类型的作品在不同时期的变化。

3 日本计量文体学发展回顾

基于所搜集的日本文献,参考日本学者评述后,本文综合研究规模、统计方法的发展情况及计算机的使用程度等因素,将日本计量文体学的历史大致划分为三个时间段:1935—1960 年、1960—1990 年、1990 年至今。

3.1 开拓(1935—1960 年):文章心理学时代

在被作为一种独立的研究范式认识之前,计量文体学在日本经过了一段较长的发展过程,并且特别值得注意的是,其在日本发端于心理学研究,当时被称为“文章心理学”,标志是 1935 年波多野完治的《文章心理学》出版。波多野完治身为心理学研究者,积极将客观的研究方法应用在文章和文体分析上,不过其主要课题都是围绕心理学的解释和应用展开,安本美典评价其具有在心理学和文学的交界处开创新一门学科的能力[20]。

除研究者个人的志趣外,文章心理学的出现与 19 世纪末到 20 世纪初日本的“文言一致”运动存在一定关系。彼时日本文学家在文体观上百家争鸣,促使日语向现代白话文体转型,诸多不同的文体特征就此产生。因作者而异的多样文学风格的出现引发了研究者对个人文体的思考,1934 年出版的谷崎润一郎的《文章读本》说明当时日本文学界也在对文章和文体进行思考①。波多野完治注意到了社会、国家语言框架下被规制的文体与表达个体性格的个人文体之间的差别,便希望探究造成文体差异的个人心理原因,这也是他将心理学应用在文章研究中的一个重要原因。之所以以“文章”命名,而不是“文体”,波多野完治解释是因为“文章心理学”比“文体论心理学”更为宽泛②,“文体论”是研究风格的语言学领域,将“修辞学”或者“作文法”的心理学方面也结合起来,才能构成文章心理学[21]。这也从说明了波多野最初即是将文章和测量作为其心理学研究工具,最终目的仍然是解决心理学的问题。他在《文章心理学》中花费了大量的篇幅论证“文如其人”③这一观点的不足,认为虽然谁都可以撰写文章,但并非人人皆有文体④;并使用句长、顿号⑤、词性等特征对谷崎润一郎和志贺直哉两位作家的文章风格进行手工计算,得出了两者风格截然不同的结论。该书出版后在日本学界引起了巨大的反响,并很受欢迎,在很长的一段时间内不断再版。在不断完善理论基础和推出新见解的基础上,波多野完治于 1966 年推出了新著《现代文章心理学》,揭示了这种新的研究方法在其他文章体裁中的应用价值。杉浦清人指出,虽然波多野的研究并没有给在文学研究方面破解某部作品或者某个作家的心理带来

① 《计量国语事典》:“1934 年谷崎润一郎的《文章读本》出版之际,社会大众对文章和表达方式的关注度非常高。”[20] 253

② 《现代文章心理学》:“如上所述,我的‘文章心理学’比‘文体论的心理学’范围更广。‘文体论’是研究风格的语言学分支,但我希望将‘修辞学’或‘作文法’的心理学方面也包括进来,形成完整的‘文章心理学’。”[21] 10

③ 《文章心理学》:“文如其人”(文は人なり)这一说法源自法国生物学家布封(Georges-Louis Leclerc de Buffon),他提出这一观点是为了表达,尽管科学真理的发现是人类共有的财富,但每个人表达这些真理的方式都是不同的,后来意思不断通俗化,形容一个人写的文字是其个性的体现”,波多野完治使用“通俗化”来阐释文章和个性之间的关系。[10] 3-36

④ 《现代文章心理学》:“如‘文言一致’运动前的日语很难反映出文体个性,又如低年级班中一群学生的作文,文体几乎如出一辙,难以体现文体个性。”[21] 9

⑤ 现代日语中的顿号等于中文的逗号。

新的见解,但《文章心理学》出版以来,其本人不断尝试各种方法将文学和统计学接轨,将对象从文学作品拓展到新闻文本等文体上,其研究价值是值得肯定的。[23] 可以认为,在日本,不论是计量文体学还是计量语言学的发展都受到了波多野完治研究的影响。

将"文章心理学"继续往计量文体学方向推动的是同样身为心理学研究者的安本美典。安本美典在接受波多野完治的理论基础之上,吸收了统计学家耶莱(Yule)(1939)等欧美研究者将句子长度作为作者判别特征的方法,尝试通过手动计算来验证《源氏物语》中"宇治十帖"①的作者问题,这也是日本学者第一次将推断统计方法应用到作者识别任务上。1960 年,安本美典的《文章心理学的新领域》一书出版,她在书中详细地解释了各种统计学分布以及因子分析法与文体特征之间的推理关系和计算过程,同时指出,当时日本各种文理研究都积极接受计量作为客观验证研究方法,唯文学落后于其他学科②。安本美典继而将自己的研究定位为从统计理论的角度进一步推进波多野完治的方法③,为当时的研究者介绍了如何使用统计学研究方法进行文学作品的作者识别。可以说,安本美典的研究推动波多野完治建立起的文章心理学继续往更"科学"的道路上发展[23]。

文章心理学开启了日本人文研究的新思路,尽管"手算"的方式在现在看来非常朴素,但却为人文研究提供了一种谁都可以验证结论的科学方法④,无疑给当时日本人文研究者开拓了一种全新的视野。

3.2 确立(1960—1990 年):"文体计算"转向

尽管魏育邻也指出 20 世纪 60、70 年代通过安本美典的研究进一步地发展了文章心理学,但与其用"发展"来概括,不如将其形容为"转向的发展"更为合适。实际上,20 世纪 60 年代后,即安本美典的《文章心理学的新领域》出版以后,全文中出现"文章心理学"的相关论文并不多见,而使用"文体""计量"等关键词的论文日渐增多,同时魏育邻也指出了安本美典的主要关注点仍然也在文章类型上,在性格心理学和时代精神方面的成果并不多[15]220。可以说,计量文体学虽然诞生于心理学研究者的摇篮之中,但其研究者们组成背景逐渐地开始丰富,开始逐渐从以心理学解释为目的的文章心理学脱出,慢慢走向语言学和文学研究导向的文体计量研究。

继 1948 年国立国语研究所成立,1956 年末计量国语学会创建⑤,这标志着计量研究方法在日语语言学领域得到了极大的重视。在这一阶段,最值得关注的是大野晋、桦岛忠夫、寿岳章子等人对文体特征的探索。不同文体的文章中的词性占比作为文体特征指标最早被关注。如大野晋研究了《万叶集》《枕草子》《源氏物语》

---

① 《源氏物语》第四十五帖至第五十四帖的主要舞台为日本宇治市,因而被称为"宇治十帖"。安本美典最早提出了"宇治十帖"可能不是出自紫式部之手的观点。

② 《文章心理学的新领域》:"或许由于用数字研究文学的难度较大,在统计学的应用方面比其他科学稍微落后了一步"。[23]9

③ 《文章心理学的新领域》:"我的研究主要是在统计理论的层面上,进一步推进了波多野完治先生的文章心理学方法"。[23]6

④ 《文章心理学的新领域》:"所谓科学,就是即使是天才所提出的想法,只要按照其逻辑进行推理,都能够理解"。[23]1

⑤ 据《计量国语学事典》荻野纲男、伊藤雅光撰序,该学会 1957 年创刊《计量国语学》,在之后近 70 年时间内坚持每年出版四期。在欧美,类似的学会直到 1960 年代才开始出现,而目前唯一仍然活跃于国际舞台的是 1994 年成立的国际计量语言学会(International Quantitative Linguistics Association,IQLA),其成立时间比日本计量国语学会晚了近 40 年。欧美的计量语言学研究虽然占据主流,但日本计量国语学会在这一领域的开创性工作和持续的研究活动,使其在国际计量语言学的发展历程中拥有不可忽视的地位。

等古典文学作品,发现动词和名词占比呈相反关系,这一语言现象被称为“大野法则”[24],并由水谷静夫进行了公式化。

1965 年出版的桦岛忠夫和寿岳章子的著作《文体的科学》[25],是这一阶段的代表作。桦岛忠夫认为“文体”分为两种,一种是“文言一致”运动前的汉文体、汉文训读体、和文体、和汉混淆文体、候文体;第二种则是个性的文体表达,具体而言是“文言一致”运动后语体统一后,人们可以随心所欲地用口语表达观点、表现自己的个性,这种文体也被称为个人“写作癖好”,受到作者的性格、经历、语言背景、思维方式甚至主观意图等多种因素影响,而为了研究这种独特的文体,就必须将文章的内容与文体进行区别①。这已经清晰展示出桦岛忠夫和寿岳章子对文体的思考在一些方面已经超出了文章心理学的范畴,朝向语言学发展。基于这样的文体思想,桦岛忠夫和寿岳章子根据文体的不同将文章分为“要约型”和“描写型”两种不同倾向,而后者又可以分为“动态描写”和“样态描写”两类。② 在这种分类的基础上,两人提出了一种日语文体特征计算指标——名词率与 MVR③,而且至今仍被使用。当然,随着研究的发展,MVR 值具体在多大程度上能说明读者对文章的真实感受也在不断被检验,例如龙太井关等人的研究[26]。

1960 年后,计算机开始进入各个研究领域。1965 年左右,国立国语研究所引入了计算机,当时的计算机体积巨大,国语研究者还是主要靠卡片来储存数据[27],但在日语的计量分析上却没有什么进展,相关文献数量也较少。村上征胜指出,这主要因为当时的计算机无法识别日语单词,这一问题一直持续到 1990 年后 Mecab、Cabocha 等分词工具和对应的词库出现后才得到了解决。[28] 在这一阶段,日本学者在研究中除了使用词频、词性等欧美计量文体学研究中的普遍指标外,还提出了众多具有日语语言特色的特征值。

3.3 突破(1990 年至今):数字技术助力时代

20 世纪 90 年代以后,计算机本身的进步,各种文学数据库的建立和日语自然语言处理技术的发展,尤其是光学符号识别、分词工具、分词词典以及句法分析④工具等的诞生,为日本的计量文体学的发展提供了强有力的支持,研究者从理论构建投入到了使用数据库和机器学习技术进行文体特征挖掘等工作中。

在这个阶段,首先是各种文学数据库和资源库建立起来,为相关研究者提供了直接动力和研究基础。上阪彩香指出,1995 年日本文部省科学研究费资助及川昭文主持的“人文科学和计算机”项目是日本古典文献数据库建设的重要标志点。[29] 近藤みゆき统计了 1990 年代建立的众多数据库,比如 1990 年长濑真理制作的日英对译《源氏物语》数据库[26],1999 年日本国文学研究资料馆公开的《日本古典文学全文数据库》《角川古典大观源氏物语》等一系列 CD-ROM 作品,角川书店 1996 年出版的《新编国歌大观》CD-ROM。[31] 除了以上这些由大型出版公司或语言资源机构建设的各种古典文献数据库,还有不少研究者出于个人研究目制作的文学

① 《文体的科学》:“为了分析文体,就必须将文体从内容中剥离出来”。[25]9

② 《文体的科学》一书中,桦岛忠夫举了三个文章段落的例子来说明不同类型的文章:大冈升平著《武藏野夫人》着重描写了事件的大体经过,而省略了事件的内部细节,此类文章定义为要约型文章;佐藤春夫的《田园的忧伤》详细描写了虫子的颜色、触角的长短、虫子爬行的模样等细节,所以属于描写型文章;久米正雄《破船》中一段文字着重于描写室内的昏暗、灯光的朦胧的样子,属于样态描写,另一段描写了临终前老师的呼吸、喉咙的动态,属于动态描写。[25]16-25

③ 公式为:M/V* 100,体现形容词、形容动词、副词、连体词等(统称为 M)与动词(V)之间的关系。[25]30

④ 日文和中文相同,没有空格作为词与词的分割标识,必须经过分词才能够让计算机识别为某个词组。

作品语料库,如村上征胜 1994 年在数理研究所承担的基盘项目《源氏物语的计量分析——基于统计分析的关于作者及创作过程的新研究》①,为《源氏物语》的作者身份鉴定问题提供了更确切的结论。另外,1997 年,日本最大的基于 Web 的文学文献全文数据库"青空文库"建立②,数字化并公开了大量日本古代以及近代以来的文学作品。

日语自然语言处理技术的发展,促使日本计量文体学走向数字时代。黑桥祯夫、长尾真开发的 JUMAN③,松本裕治研究室的 ChaSen④ 和 MeCab⑤ 以及 GINZA⑥ 等分词工具先后诞生;小木曾智信团队开发了专精于某一时代的词典。

在这一转变的过程中,统计学领域的研究者率先开始探索如何将机器学习方法应用在文学研究中,比如村上征胜。安本美典指出村上征胜虽然是统计学者,但在文学方面也十分有造诣。[32] 这种优势使其在 1994 年承担了前文所述的《源氏物语》项目后,1996 年又承担了"人文科学和计算机计量研究"项目,不仅针对《源氏物语》《紫式部日记》以及井原西鹤作品集、梵语大乘佛典等进行了数据库建设,同时展开了非文本对象的数字人文研究,例如根据浮世绘美人画中人物的面部数据进行创作者分析,根据寺院外形数据进行建造年代分析。而单纯统计学和理工科背景的研究者更偏向于方法探索,比如统计方法讨论,不同文体特征对文体风格识别的效度验证等等。这方面日本信息处理学会做出了许多具有代表性的研究。研究者普遍使用有监督学习方法,统计方法不断演化、不断科学化[33]。村上征胜在人文领域内使用计量方法的成果和探索在一定意义上直接推动了同志社大学将文化情报学作为独立学部成立,可以说是了解日本数字人文发展和计量文体学发展过程中不可忽略的重要研究者。

最近 20 年,各种便捷的文本分析工具相继诞生。例如开源的非结构化文本处理工具 KH-Coder,是由樋口耕一于 2001 年开发并持续完善的一款文本挖掘软件,支持日语、汉语、英语、韩语等多种语言的分词,使用多种统计方法并对结果进行可视化,呈现非结构化的文本数据中的各种信息[34],可支持多领域的人文社科研究。到 2024 年 10 月为止,通过 KH-Coder 来完成的研究超过 7000 项。⑦ MTMineR⑧ 是金明哲使用 JAVA、R 等语言制作的交互应用[35],可以支持多语种的分词、词性赋码,抽取 N-gram、句法树,支持向量机模型、决策树模型、LDA 线性分析等多种机器学习方法,还可以进行无监督分析、半监督分析。这些文本分析工具使得文科背景的研究者能较容易地使用计算方法,拉近了愈来愈复杂的数字方法和传统人文研究之间的距离,推动了日

① 具体可以参见"本科研经费数据库"中该课题的结项信息,KAKIN: https://kaken.nii.ac.jp/ja/report/KAKENHI-PROJECT-06451163/064511631996kenkyu_seika_hokoku_gaiyo/。

② 青空文库是富田伦生、野口英司等人倡导建立的大规模免费在线文学作品库,目前也面临着一些挑战,如校正和维护人员减少、20 余年前创立的运作模式难以继续、版权保护时间延长 20 年导致很多作品无法免费公开,等等。

③ 具体可参见京都大学语言媒体研究室官网: https://nlp.ist.i.kyoto-u.ac.jp/edit.php?JUMAN。

④ 具体可参见奈良先端科学技术大学自然语言处理研究室官网: https://cl.naist.jp/?%BC%F5%BE%DE%B0%EC%CD%F7。

⑤ MeCab 是由京都大学信息学研究科与日本电信电话株式会社通信科学基础研究所共同研究开发的,具体可参见: https://taku910.github.io/mecab/。

⑥ GiNZA 是由 Megagon Labs 与国立国语研究所合作开发的,具体可参见: https://megagon.ai/jp/projects/ginza-install-a-japanese-nlp-library-in-one-step/。

⑦ 到 2024 年 10 月 14 日为止,有 7193 项研究是基于 KH-coder 完成,参考: https://khcoder.net/bib.html?year=recent&auth=all&key=。

⑧ MTminer 详情请参见: https://a3hsn.org/mt.html。

本数字人文的发展。

3.4 重要发展节点和成果

为更好地展示日本计量文体学的发展脉络,本文将以上回顾中的重要研究或事件进行了梳理,如表1所示。

表1 日本计量文体学的重要发展节点

年	研究者或机构	研究成果或事件	意义
1935	波多野完治	《文章心理学》出版	日本计量文体学的开端
1956	大野晋	大野晋提出的词汇法则	最早对词性的探索
1956	计量国语学会	计量国语学会成立,同年创刊《计量国语学》	重要学术平台
1957	安本美典	《源氏物语》"宇治十帖"作者判别	日本学者第一次用计算方法进行作者身份识别
1965	桦岛忠夫、寿岳章子	《文体的科学》中提出MVR	重要文体指标
1965	水谷静夫	将大野晋词汇法则进行了科学的公式化	重要文体指标
1970	宫岛达夫	宫岛达夫提出了"词汇相似度"计算公式	为文体识别提供了重要计算公式
1991	村上征胜、伊藤瑞叡	日莲著作作者身份判别	使用现代统计学方法进行计量文体学
1993	金明哲	应用标点符号进行文体判别	重要文体指标的提出
1998	汉字文献信息处理研究会	汉字文献信息处理研究会成立,创刊《电脑中国学》	为汉学研究交流计量方法提供了重要平台
2000	近藤みゆき	《古今和歌集》中的男女用词差异	重要的日本古典文学计量研究案例
2002	近藤泰弘、近藤みゆき	NGSM古典文献分析法	为汉学和日本古籍研究提出了重要文体指标
2002	师茂树	使用N元词进行佛教典籍《般若心经》的计量分析	在宗教典籍中的应用
2004	山田崇仁	中国战国、先秦时期的汉语词汇测量分析	汉籍的测量研究
2004	樋口耕一	非结构化文本分析工具KH Coder公开	出现了更多的基于软件的计量文体学案例
2005	同志社大学	文化情报学作为独立学科成立,村上征胜任当时的第一任学部长,学科定位为文理融合	日本数字人文教育发展的重要节点
2009	计量国语学会	《计量国语学会事典》出现了"计量文体学""计量文献学"条目	被承认为一个独立的研究领域
2016	金明哲	计量文体分析工具MTminer公开	重要的计量文体学研究工具

4 日本计量文体学代表性研究

从上一节的发展回顾和表 1 可以看出,《源氏物语》、和歌、汉籍、佛教典籍和近代作家是日本计量文体学的重要研究对象,产生了许多有代表性的研究成果。这一节将从领域视角出发对这些成果进行分析,以进一步总结日本计量文体研究的特点。

4.1 《源氏物语》系列研究

在日本计量文体学的发展中,对被誉为日本古典文学最高峰的《源氏物语》展开的研究形成了一条独立的支线,贯穿着整个过程。其研究中主要集中在闭集归属研究和文体变化研究方面。今西祐一郎和室伏信助将《源氏物语》的文体问题总结为以下四点: (1) 《源氏物语》初卷的成立顺序; (2) 初卷的排列顺序; (3) 故事的发展线是如何成立的; (4) 作者的身份问题。[36] 从 20 世纪 50 年代开始,日本学者就着手探索这些问题。安本美典(1958) 抽取直喻、拟声拟态词、色彩词、心理描写等文体特征,基于卡方检验、因子分析得到了《源氏物语》前四十四帖和“宇治十帖”不是同一作者所作的结论。[37] 新井皓士追溯了日本平安时代文字的发展,并基于五十音图的首辅音行和母音行的频率数据测量了《源氏物语》的部分章节,认为“宇治十帖”不存在作者著作权的问题。[38] 因人力和自然语言处理技术的有限,在 1999 年之前,研究者还没有对《源氏物语》进行过全文分词处理。村上征胜和今西祐一郎制作了 37.6 万词的《源氏物语》语料库,并首次使用了全文分词、词性标注的方法,测量了助动词使用频率,认为第一部“玉发十六帖”很可能是在第二部完成后才写的,并提出仅仅通过测量助动词数量无法得出“宇治十贴”是紫式部之外的人所作这一结论。[39] 土山玄对《源氏物语》和《宇津保物语》等古典文学作品进行了计量分析,最后也否定了“宇治十贴”有另外作者的说法。[40] 小野洋平(2015) 从统计学角度重新验证了村上征胜、今西祐一郎的研究,指出后者在创作时间上的问题。[41] 日本研究者对《源氏物语》的一系列计量文体研究中开拓了许多文体特征,并且对这些文体特征的效度、相关性进行了讨论,安本美典的观点不断被检验,目前得到的基本观点是《源氏物语》是同一人的作品,但是章节顺序和创作时间仍存在争议。可以预见,《源氏物语》仍将是日本计量文体学研究者未来的重要课题。

4.2 和歌、汉籍、佛教典籍三领域

和歌、汉籍、佛教典籍作为古代典籍,也是日本计量文体学的重要研究对象。研究者主要是传统人文学者,研究问题主要集中于个人文体和类型文体,注重对文体特征的描述以至借助数字方法来解决文学问题。

和歌研究受益于前文所述 1990 年后日本出版商积极推动的古典文献数字化。基于古典文学数据库,大量使用计量方法的和歌研究出现,其中最具代表性的是近藤みゆき承担的“关于平安时代和歌资料中特殊词汇提取的计量研究及其工具的公开”的项目。近藤みゆき和近藤泰弘使用 N-gram 分析日本和歌男女作者的用词文体差别是其中一项的研究。该研究发现,男作者和女作者的用词有较明显的差异: 男性和歌用词更倾向于向目标主动表达爱慕之意; 而女性和歌中多出现泣鹿和鸟啼等意象,用来映射受爱情困扰的命运。[42] 这一研究方法被汉字信息处理学会的研究者们注意到,他们又基于 N-gram 推出了新的指标,并应用到了其他方面

的探索中。不同年代的词性构成、不同和歌作者的文体特征判别也是该领域的重要研究问题,例如山元启史和ボルホドシチェク选取了六位和歌作家的作品,使用了 Jaccard 指数和 Dice 系数计算了作家之间的相似度。[43]

日本的中国学擅长以定性为主的考据方法,但仍有部分研究者引入了计量方法,这部分研究说明汉学也是日本数字人文的一个应用领域。汉字文献信息处理学会是汉籍计量研究的重要场所,学会的《汉字文献信息处理研究》《电脑中国学》等杂志发表了一些现代和古代汉语的处理方法研究,石井公成、近藤泰弘、师茂树和山田崇仁教授在这方面付出了诸多努力。山田崇仁使用 N-gram 对《孟子》《中庸》《周礼》以及战国时期的汉字词汇进行了计量分析,成果丰富;[44] 此外还对《孙子兵法》进行测量,证明了其成书时期可能是公元前 3 世纪①。下西纪子通过 N-gram 模型对《山海经》的文本数据进行量化处理,然后运用层次聚类分析和判别分析等统计技术来识别不同篇章的执笔者,最终得出了《山海经》的编著者不止一人的结论,如《五藏山经》和其余部分的编著者不同,《五藏山经》的五篇是同一编著者,其他海外经和海内经的编著者也各有不同。[45]

佛教典籍研究领域,研究者一方面积极推动建立数据库,一方面使用计量方法开展本土宗教典籍和外来典籍译者研究。古濑顺一对日莲和亲鸾文章的文体特征进行了计量分析,发现日莲的文章更注重理论性;另外,相较于给男性写的信,日莲给女性写的信中使用了更柔和、更平易近人的语言。[46] 村上征胜等对日莲的 23 篇文章进行了计量分析,发现其文体特征发生变化的时间与传统人文研究者认为的流放时期不吻合。[47] 石井公成使用汉字文献信息处理学会研究者们共同提议的 NGSM② 比较了《大乘起信论》与其他文献在用词和语法上的一致性和差异。[48]

4.3 近代作家相关研究

近代作家的文体判别也是最近 20 年日本计量文体学的主要研究主题。这方面,金明哲团队成果丰硕,其研究主要关注文体变化和闭集归属问题,在算法和特征值的提出方面也有不少成果。尾城奈绪子和金明哲针对“青空文库”中的 225 篇太宰治的文章进行分析,得出太宰治吗啡戒断反应期间部分文体特征发生了明显变化的结论。[49] 刘雪琴观察到宇野浩二在患上精神疾病后文体发生了显著的变化,于是从文字、符号、词汇以及句法方面进行了计量分析,证明了此观察:宇野浩二病后作品中逗号和名词的使用率提高,文体从口语语体变化为一种近书面语的语体。[50] 孙昊对川端康成的诸多有代笔疑问的作品进行了计量分析,提出《古都》很可能是川端康成和学生北条城、泽野久雄合作完成的观点,并认为《花日记》很有可能是中里恒子和川端康成共同执笔的小说,以及三岛由纪夫不是《山之音》作者。[51] 柳烨佳和金明哲注意到《受难华》可能不是菊池宽的作品、而是由横光利一代笔的争议,抽取了《受难华》的标点、词性、多元词等特征进行了作者身份计量分析,得到了作者是菊池宽本人的结论。[52] 目前,日本近代作家的文体问题还有许多研究空间,比如结合语言学和社会心

① 详情可以参考山田崇仁的个人博客“睡美人亭”:https://www.shuiren.org/profile.htm。

② NGSM(N-gram Based System for Multiple Document Comparison and Analysis)是以近藤泰弘为首的汉字文献信息处理研究会的研究者们开发的文献比较分析方法。

理学等进行综合分析。

4.4 日本计量文体学常用的文体特征指标

以上介绍的研究中使用了大量的文体测量指标,日本研究者在文章心理学时代波多野完治、安本美典所使用的各种文体特征——例如表示心理描写的 'おもす、おぼす、覚ゆ、思し煩ふ"("我想" "我觉得" "感到"一类词) 的基础上,提出了很多其他的文体指标,例如在 N-gram 的基础上衍生出了基于词性的 N-gram,基于助词和标点符号的 2-gram。限于篇幅,本文无法对特征值的应用进行一一总结,现参考浅石卓真的整理[53],将日本计量文体学研究中常用的文体指标划分为长度、频率、特定形式三种进行了汇总。

表 2 日本计量文体学研究常用文体指标

类型	文体特征名称	中文解释
长度	単語の長さ	单词长度
	文の長さ	句长度
	段落长	段落长度
	会話文と地の文	引语和其他非引语部分
词频	品詞の使用率	不同词性的词占比
	色彩語と比喩語	色彩词语和比喻词,如"红" "黑"
	指示詞率	指示代词占总词量的比例
	名詞比率	名词占总词量的比例
	漢字含有率	汉字的含有量
	主辞の現れ方	主语的使用频率与使用方法
	直喩の使用度、声喩の使用度	のような、のごとし等比喻提示词,なよなよ、はらはら、ほろほろ等拟声拟态词
	心理描写(おもす、おぼす、覚ゆ、思し煩ふ等等)	表示心理描写的"我想" "想" "思考"等词或字
	4 字以上の漢字列の割合	4 个字以上的汉字词比例
特定形式	MVR	动词/形容词* 形容动词* 副词* 连体词
	N-gram	N 元词组
	NGSM	滑动窗口 N 元词组
	文頭文末のパターン	句首与句末出现的词语
	TF-IDF	词频与逆词频
	語彙の豊富さ	词汇丰富度
	句型の豊富さ	句型丰富度
	係り受け距離	依存距离
	音素	母音和子音

5 结语

本文在对计量文体学的内涵进行分析和对国内学者的相关研究进行整理的基础上,对日本计量文体学的发展历史、主要研究成果进行了梳理。计量文体学在日本发展历史可以大致分为“文章心理学”开拓时代、“文体计算”转向阶段和“数字技术”助力的时代,典型研究案例集中在《源氏物语》、和歌、佛学典籍、汉籍和近代作者文体的计量分析上。

在整理、回顾的基础上,可以发现日本计量文体学具有起步时间早、成果也较为丰富的特点。这与日本的学术评价体系、教育体系下形成的“强目的导向”“弱学科导向”不无关系,这一点有待继续探讨。

在当下以大语言模型为代表的人工智能技术飞速发展的数字时代,传统人文研究式微,突显人文学科的独特性成为重要话题。正如日比嘉高所言,人文学科若想抹去学科的夕阳感,就必须走数字人文的道路。[54] 计量文体学不仅仅是工科领域的作者识别技术等的代名词,它也为人文学者提供了从人文问题出发、通过数字方法阐释文学的新途径。日本计量文体学的发展经验和案例为我国的文学研究提供了一种新的视角。计量文体学在我国的文学领域未耕耘之地广阔,未来值得期待。

参考文献

[1] MORETTI F.Distant reading [M] .London: Verso Books,2013.

[2] JOCKERS M L.Macroanalysis: digital methods and literary history [M] .Champaign: University of Illinois Press,2013.

[3] LUTOSLAWSKI W.Principes de stylométrie appliqués a la chronologie des œuvres de Platon [J] .Revue des études grecques,1898,11(41):61-81.

[4] 刘洋,余思琪.文学计算的理论、方法及发展挑战——刘洋博士访谈 [J] .数字人文研究,2023,3(4):37-48.

[5] 张旭,洪逸暄,尤剑.我国数字人文研究热点与趋势 [J] .图书馆工作与研究,2023(10):69-76.

[6] MENDENHALL T C.The characteristic curves of composition [J] .Science,1887,9(214):237-246.

[7] 李之基.《静静的顿河》著作权问题新说 [J] .文史哲,1993(5):95-98.

[8] 施建军.基于支持向量机技术的《红楼梦》作者研究 [J] .红楼梦学刊,2011(5):35-52.

[9] 陈大康.从数理语言学看后四十回的作者——与陈炳藻先生商榷 [J] .红楼梦学刊,1987(1):293-318.

[10] 波多野完治.文章心理學入門 [M] .东京: 新潮文庫,1953.

[11] STAMATATOS E.A survey of modern authorship attribution methods [J] .Journal of the American Society for Information Science and Technology,2009,60(3):538-556.

[12] TEMPESTT N,KALAIVANI S,ANEEZ F,et al.Surveying stylometry techniques and applications [J] .ACM Computing Surveys,2018,50(6):86.

[13] LAGUTINA K,LAGUTINA N,BOYCHUK E,et al.A survey on stylometric text features [C] //2019 25th Conference of Open Innovations Association (FRUCT),Helsinki,2019:184-195.

[14] SAVOY J.Machine learning methods for stylometry: authorship attribution and author profiling [M] .Berlin: Springer,2020.

[15] 魏育邻.日语文体学 [M] .吉林: 吉林教育出版社,2002.

[16] 施建军.计量文体学导论: 卷 1 [M] .北京: 北京大学出版社,2016.

[17] 毛文伟.日本的语料库文体学研究: 进展、问题及展望 [J] .外国语(上海外国语大学学报) ,2021,44(3) : 82-90.

[18] 李月平,毛文伟.小说文体的量化研究——以夏目漱石的短篇小说为例 [J] .外语电化教学,2011(1) : 51-55.

[19] 李文平,劉海濤,呉長紅.芥川龍之介の児童文学の文体に関する計量的分析 [J] .計量国語学,2022,33(7) : 526-540.

[20] 計量国語学会.計量国語学事典 [M] .东京: 朝倉書店,2009.

[21] 波多野完治.現代文章心理学 [M] .东京: 大日本図書株式会社,1966.

[22] 杉浦清人.文学研究におけるデジタル・ヒューマニティーズの可能性 I: 文章心理学・計量文献学・マクロ分析 [J] .れにくさI: 現代文芸論研究室論集,2017,7: 80-96.

[23] 安本美典.文章心理学の新領域 [M] .东京: 東京創元社,1960.

[24] 大野晋.基本語彙に関する二三の研究——日本の古典文学作品に於ける [J] .国語学,1956(24) : 34-46.

[25] 樺島忠夫,寿岳章子.文体の科学 [M] .京都: 綜芸舎,1965.

[26] 龍太井関,理紗菊池,正哉望月,等.品詞構成に基づく文体指標は読者の印象とどのように関わるか [J] .計量国語学,2022,33(7) : 493-509.

[27] 丸山直子.日本語学とコンピュータ [C] //ソフトウェアエンジニアリングシンポジウム2011 論文集,2011: 1-4.

[28] 村上征勝.文章の計量分析 [J] .計測と制御,2000,39(3) : 216-222.

[29] 上阪彩香.古典文を対象とした計量的研究の現状 [R] //情報処理学会研究報告.人文科学とコンピュータ研究会報告,2015,(7) : 1-2.

[30] 長瀬真理.「源氏物語」ハイパー・テキストの開発の試み [R] //情報処理学会研究報告.IM, [情報メディア] ,1995,95(1) : 11-18.

[31] 近藤みゆき.Nグラム統計処理を用いた文字列分析による日本古典文学の研究——『古今和歌集』の「ことば」の型と性差 [J] .千葉大学人文研究,2000(29) : 187-238.

[32] 安本美典.文化を計る: 文化計量学序説 [J] .行動計量学,2003,30(1) : 166-166.

[33] 杉浦清人.文学研究におけるデジタル・ヒューマニティーズの可能性 I: 文章心理学・計量文献学・マクロ分析 [R] //れにくさI: 現代文芸論研究室論集,2017,7: 80-96.

[34] 樋口耕一.言語研究の分野におけるKH Coder 活用の可能性 [R] //計量国語学,2017,31(1) : 36-45.

[35] 金明哲.教育と研究のためのテキストマイニングツールMTMineR(5. 4) [J] .日本計算機統計学会大会論文集,2016,30: 113-116.

[36] 小野洋平.源氏物語成立論の統計科学的再考察 I: 村上・今西(1999) を中心に [J] .計量国語学,2015,29(8) : 296-312.

[37] 安本美典.文体統計による筆者推定 [J] .心理学評論,1958,2(1) : 147-156.

[38] 新井皓士.源氏物語・宇治十帖の作者問題 I: 一つの計量言語学的アプローチ [J] .一橋論叢,1997,117(3) : 397-413.

[39] 村上征勝,今西祐一郎.源氏物語の助動詞の計量分析 [C] //情報処理学会論文誌,1999,40(3) : 774-782.

[40] 土山玄.計量文献学による『源氏物語』の成立に関する研究 [D] .京都: 同志社大学,2015.

[41] 小野洋平.源氏物語成立論の統計科学的再考察 I: 村上・今西(1999) を中心に [J] .計量国語学,2015,29(8) : 296-312.

[42] 近藤みゆき.Nグラム統計処理を用いた文字列分析による日本古典文学の研究——『古今和歌集』の「ことば」の型と性差 [J] .千葉大学人文研究,2000,(29) : 187-238.

[43] 山元啓史,ボルホドシチェク.八代集「桜の花」歌における作者の分類 [C] //じんもんこん2018 論文集,2018: 175-180.

[44] 山田崇仁.N-gram 方式を利用した漢字文献の分析 [J].立命館白川靜記念東洋文字文化研究所紀要,2007(1):1-23.

[45] 下西紀子.統計的手法による『山海経』編著者の識別 [J].Core ethics:コア・エシックス,2015,11:107-121.

[46] 古瀬順一.日蓮遺文の文体:計量分析を通して [J].統計数理,1988,36(1):89-97.

[47] 村上征勝,岸野洋久,伊藤瑞叡.日蓮遺文の計量分析 II:思想の変化と文体の変化(統計数理研究所研究活動(研究会報告 文献情報のデータベースとその利用に関する研究会)) [J].統計数理,1990,38(2):311.

[48] 石井公成.『大乘起信論』の用語と語法の傾向 [J].印度學佛教學研究,2003,52(1):293-287.

[49] 尾城奈緒子,金明哲.太宰治の文体の計量的分析——助詞と使役,受身,授受構文の関連を中心として [J].計量国語学,2016,30(7):457-458.

[50] 劉雪琴.宇野浩二の文体的特徴に関する計量的研究 II:文体変化を中心に [D].京都:同志社大学,2019.

[51] 孫昊.川端康成の代筆問題及び文体問題に関する計量的研究 [M].东京:同志社,2017.

[52] 柳燁佳,金明哲.菊池寛「受難華」の代筆問題の研究 [J].データ分析の理論と応用,2020,9(1):1-11.

[53] 浅石卓真.テキストの特徴を計量する指標の概観 [J].日本図書館情報学会誌,2017,63(3):159-169.

[54] 日比嘉高.日本近现代文学研究者用计算机想做什么?不想做什么? [J].数字人文研究,2021,1(3):89-92.

The Development of Stylometry in Japan: From Article Psychology to Digital Humanities

Wang Zirui

Abstract This study aims to systematically trace the history and development of stylometry in Japan from the perspectives of temporal evolution and content classification through a literature review. It also analyzes representative cases and characteristics of stylometry in Japanese literary research. The evolution of Japanese stylometry has exhibited distinct localized features: in the 1930s, it emerged in the field of psychology with minimal influence from Western theories; around the 1960s, the establishment of the Mathematical Linguistic Society of Japan facilitated the popularization of quantitative linguistic methods; and after the 1990s, advancements in computer technology and statistics ushered in a period of prolific research output. Throughout this extensive accumulation process, Japanese stylometry has gradually developed interdisciplinary characteristics that intertwine linguistics, psychology, and literature, establishing a research paradigm that merges the sciences and humanities and emphasizes problem-oriented inquiry. This has resulted in a wealth of practical research findings that serve as important references for domestic literary studies.

Key words digital humanities; stylometry; Japan; quantitative research