外国文学的计量研究：研究背景、发展现状及研究路径

Quantitative Studies of Foreign Literature: Background, Development and Approach

王　永（**Wang Yong**）

内容摘要：计算机技术的发展不仅使社会生活发生了重大变革，也为学术研究带来了很大的便利。借助数据库，研究者可以节约大量耗费在文献检索方面的时间，并且可以通过数据分析发现传统研究无法发现的特征。然而，外国文学界对此关注较少，产出的相关成果不多。本文通过对文学计量研究成果的综合分析，阐明在外国文学研究中运用计量方法的必要性与可行性，同时，结合相关研究详细介绍文学计量研究的步骤和方法。本文不仅有助于外国文学研究者了解数据、统计方法及文学研究的关系，还可以为其提供具体的研究路径，推动外国文学计量研究成果的产出。

关键词：外国文学；计量方法；数据库；统计分析

作者简介：王永，文学博士，浙江大学外国语言文化与国际交流学院教授、博士生导师，主要从事俄罗斯诗歌的计量研究及俄罗斯艺术研究。本论文为国家社科基金重大项目“中国外国文学研究索引（CFLSI）的研制与运用”【项目编号：18ZDA284】以及浙江文化研究工程（第二期）重大项目“浙江文学翻译家年谱”【20WH60068ZD】的阶段性成果。

Title: Quantitative Studies of Foreign Literature: Research Background, Development and Approach

Abstract: The development of computer technology has not only made significant changes in social life, but also brought great convenience to academic research. With the help of databases, researchers can save a lot of time spent on literature retrieval, and can discover features that cannot be discovered by traditional research through data analysis. However, in the foreign literary circles, the usage of databases hasn’t been drawing much attention, thus having produced fewer related achievements. Through a comprehensive analysis of the literary measurement research results, this article clarifies the necessity and feasibility of using measurement methods in foreign literature research. At the same time, it introduces the steps and methods of literary measurement research in detail in conjunction with related research. This

article not only helps foreign literary researchers to understand the relationship between data, statistical methods and literary research, but also provides them with specific research paths to promote the output of quantitative research results in foreign literature.

Keywords: foreign literature; quantitative method; corpus; statistic analysis

Author: Wang Yong, Ph. D., is Professor at School of International Studies, Zhejiang University (Hangzhou 310058, China). Her major fields of inquiry include Russian literature and quantitative linguistics (Email: wangyongzju@163.com).

大数据、云计算、人工智能、数字人文，是新世纪尤其是近十年来的学术热点话题之一。面对大数据时代，有的高校和学术机构积极相应。2015 年 3 月，复旦大学中文系启动“语言 • 大脑 • 计算”交叉学科平台。2020 年 1 月，清华大学联合中华书局主办的《数字人文》创刊号发行。各种与大数据、数字人文相关的译著、编著先后出版。

毋庸置疑，计算机技术已渗透到社会的各个领域，且不论“全球化本身是由数字技术的崛起所推动的”（奥恩 19），即使从眼前发生的事来看，大数据在新冠疫情期间所发挥的重要作用有目共睹。可以说，“计算机已经达到改变世界的‘全力’发展阶段”“大数据正在彻底改变从社会科学到商业的各个领域”（奥恩 IV-V）。

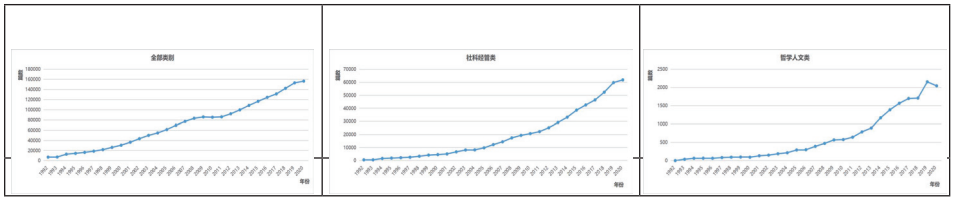
然而，迄今为止，数据这块石头尚未在外国文学界激起千层浪，不少学者持保留乃至质疑的态度。这一方面或许由于部分研究者过分夸大了定量研究的价值和有效性；另一方面，基于数据的研究往往停留在图表呈现及数据的陈述上，而进一步结合文学内容展开深度阐释的成果较少，难以充分显示出新方法对文学研究的价值。但最重要的原因，恐怕是由于对文学的计量研究了解不多之故。那么，面对大数据时代日新月异的计算机技术及其为各学科领域带来的丰硕成果，外国文学界该如何看待？文学的计量研究已取得哪些成果？如何开展外国文学的计量研究？本文将从外国文学开展计量研究的大背景、计量研究发展现状及具体的研究路径等方面展开阐述。

一、外国文学计量研究的背景

有学者断言，“‘大数据’时代的很多学科都将发生巨大甚至是本质性的变革和发展，进而影响人类的价值体系和知识体系，当然也影响到我们的学术研究”（郑永晓 143）。事实确实如此，计算机技术的发展正在影响着一个又一个学科。近二十年来，国内出版的相关著作及发表的论文增长迅速。自然科学自不必说，社科领域的应用也是如火如荼。在人文领域，历史学、传播学、语言学的定量研究已产出较为可观的成果，与此相比，文学领域的差距明显。

从出版的著作看，通过当当网搜索[1]关键词“大数据”，有 292068 个结果，排名靠前的有《大数据时代》《数据化决策》《大数据架构详解：从数据获取到深度学习》等；搜索“云计算”，有 108268 个结果，如《云计算：概念、技术与构架》《云计算通俗讲义》《图解云计算架构》等；搜索“人工智能”及“AI”，共计 138469 个结果，如《人工智能 从小白到大神》《人工智能：国家人工智能战略行动抓手》《AI· 未来》《AI 经济》等；搜索“数字人文”，种类显著减少，仅有 1552 个结果，且很多条目与数字人文无关，并有不少条目重复，主要文献有以下几种：《数字人文：改变知识创新与分享的游戏规则》《数字人文：数字时代的知识与批判》《赛博文化与数字人文》《< 献帝春秋 > 钩沉——从数字人文角度刊古籍辑佚》《面向知识挖掘句法语料库构建研究：数字人文视角下的史部典籍信息组织》，以及清华大学和中华书局合办的辑刊《数字人文》，复旦大学国家文化创新研究中心推出的文集《数字人文研究》。

从产出的论文看，通过知网搜索关键词“数据”（自动关联“大数据”等相关关键词），可以检索到论文总数 2156099 篇。其中社科经管类 590527 篇，占总数 27.39%；哲学人文类 19366 篇，占总数 0.90%。[2] 年度论文发表趋势图如下表：



数据显示，计算机技术在整体学术领域的应用上升趋势明显。1992 年的论文仅 7046 篇，2000 年增至 30515 篇，2010 年为 85260 篇，2020 年达 156523 篇，30 年内增加到 22 倍。2000 年前的 9 年，年均增加 2607.67 篇，年均涨幅 37.01%；此后的两个十年，年均增加的论文数分别为 5474.5 及 7126.3 篇，年均涨幅分别为 17.94% 及 8.36%。除了个别年度，每年发表的论文数绝对值不断增加。社科领域的发文数虽然增加的绝对值不如整体发文数，但增幅远大于前者。1992 年的论文仅 513 篇，2000 年增至 4562 篇，2010 年为 20693 篇，2020 年达 61857 篇，30 年内增至 120 倍。2000 年前，年均增加的论文数是 449.9 篇，年均涨幅 87.70%；此后两个十年，年均增加的论文数分别为 1613.1 及 4116.4 篇，年均涨幅分别为 35.36% 和 19.89%。哲学人文领域的发文数虽明显逊色，且前期进展速度较慢，但近些年有较大幅度的增长。1990 年的论文 34 篇，2000 年仍是二位数的 93 篇，2010 年为 577 篇，2020

1 截止 2021 年 8 月 14 日 13 时 10 分。

2 截止 2021 年 8 月 11 日 16 时 48 分。

年达 2044 篇，30 年内增至 60 倍。2000 年前增长缓慢，年均增加的论文数是 5.9 篇，年均涨幅 17.35%；此后两个十年，年均增加的论文数分别为 48.4 及 146.7 篇，年均涨幅分别为 52.04% 和 25.42%。

从上述检索结果可以看出，计算机技术引导的成果非常显著。尤其是新世纪以来，随着大数据、人工智能、云计算等技术越来越多地在人们的生活中得到运用，各学科领域越来越认识到计算机技术在学术研究中的重要性，开始大规模运用数据开展研究，相关成果显著增加，至今保持持续增长。即使是落后于其他学科的哲学人文领域，近十余年来的成果已有较大幅度的增长。

与此相比，外国文学研究领域对计算机技术的接受明显滞后。尽管各种会议、各种论坛上“大数据”“数字人文”“跨学科”的字眼并不鲜见，但在实际研讨过程中，更多的跨学科关注点集中于文学与历史学、政治学、经济学、法学的跨界研究，而对大数据、数字人文、数据库关注度较小，对数据能如何用于外国文学研究更是了解不多。即使与文学学科内的中国文学相比，也落在后面。知网上通过关键词“数据”可以搜索到的文学类（世界文学及中国文学）中文论文只有区区 480 篇，而世界文学更是仅占其中的 18.75%，90 篇。

然而，我们身处一个数字化、互联网的时代，人工智能发展迅速，Alpha-Go 的围棋水平令人心服口服；人机协同创作的诗歌、AI 书法、AI 绘画、AI 主播，让人难辨真伪。面对强大的计算机技术，对比其他学科运用数据开展研究所取得的学术成果，我们认为，外国文学界既无需盲目追风，但也不能视而不见。尽管在数十年甚至更长的时间内，传统研究依然是其主流，但与此同时，我们应该正视大数据时代的技术发展给外国文学学科带来的机遇和挑战，开拓视野，学习和借鉴新方法。可以说，外国文学的计量研究，既是现代计算机技术发展的使然，又是深入挖掘文学问题，开拓文学文本研究范式的必然。

二、文学计量研究的发展现状

文学的计量研究之所以可行，是由于这种研究已有一百多年的发展史，并已形成了几个主要的研究方向。

国内对这部分的了解和介绍较少。《山东社会科学》发表的译文“查找与替换：约瑟芬·迈尔斯与远距离阅读的起源”认为，人文学科定量或计算方法的奠基学者起码应该往前追溯到迈尔斯。这位美国诗人兼加州大学伯克利分校的学者早在 20 世纪 30 年代的研究生期间就用人工方式“完成了自己的第一个远距离阅读项目：分析罗马诗人偏爱的形容词”（雷切尔 46）。50 年代，她与团队合作完成了计算词语检索项目。几十年中，她采用单词计数方法，对华兹华斯、怀亚特、奥登等诗人的语言及创作风格做了分析。这种研究无疑“有利于我们建立一个远距离阅读和定量文学史研究方法的多元化

的学科史谱系”（雷切尔 45）。也有学者提到 21 世纪英美学者基于语料库对文学文体所做的研究，如 “Tabata 研究了狄更斯小说中的文体风格变化；Clupeper 对莎翁作品《罗密欧与朱丽叶》中的对白做了关键词、词性及语义域的研究 Fischer-Starcke 则集中在对简 • 奥斯汀众多文学作品的研究等”（任艳 陈建生 丁峻 17）。

但这只是文学计量研究发展史上的一鳞片爪。事实上，文学计量研究有着百余年的发展，大致可分为三个阶段：1）计算机尚未出现之前的统计分析；2）计算机技术开始发展的 20 世纪中期；3）大数据时代。为了系统展示其总体发展过程并为我国的文学研究者提供参照，前两个阶段的发展侧重介绍俄罗斯学者的研究成果[1]，第三个阶段综述我国学者的研究成果。

计算机尚未出现之前，文学的计量研究主要采用人工统计方法。20 世纪初，随着形式主义诗学的兴起，俄罗斯的文学批评家和语言学家开始采用统计方法研究文学（诗歌）作品。波利万诺夫（Л. Поливанов）、维诺库尔（Г. Винокур）等文学批评家和语言学家，以及别雷（А. Белый）、丘特切夫（Ф. Тютчев）等诗人对诗歌的节奏、韵脚、诗节乃至情节、结构、主题、题材、文学流派等方面做了统计分析，在诗体语言的量化特征研究上取得了一定的成就，也为现代文体测量奠定了基础。莫罗佐夫（Н. Морозов）对普希金、果戈里、托尔斯泰等作家使用的前置词、语气词、代词等词类做了统计分析，以此辨别作家作品真伪，成为最早采用定量方法鉴别著作权的研究者之一。雅尔霍（Б. Ярхо）则将统计方法运用于斯拉夫文学、日耳曼文学、中世纪文学、古俄罗斯文学、民间文学的研究中，试图构建文学研究的计量理论。他认为“运用统计方法，可以解决大量同作品的修辞、作品的主题和谋篇、作品的总体思想和情感以及作品的题材有关的各种问题。文学的生成、演变和类型学问题，尤其是文学流派问题可以转换为完全对等的数字语言”（Ярхо xviii）。这种方法是一种“精密”研究方法，其本质在于“从分析到综合”，即“先从文学文本中提取重要特征进行分析，再对这些数据进行统计运算，然后对研究现象的发展及功能规律得出结论”（Ярхо xviii）。遗憾的是雅尔霍英年早逝，而他写于 30 年代，近 400 页的未竟之作《精密文学研究方法论》，也直到 2006 年才得以整理出版。

20 世纪 50-60 年代，随着计算机科学的发展，更多的学者在文学研究中尝试运用概率论、信息论、控制论等自然科学的理论与方法。如博布罗夫（С. Бобров）用常数和变量进行排列组合获取统计数据的方法，将普希金的诗歌与俄罗斯民间创作进行对比分析，最终证明普希金的长诗《西斯拉夫之歌》虽然以俄罗斯民间歌曲为基础，但绝“不是俄罗斯古代诗歌的翻版，而是俄罗斯诗歌史上史无前例的新诗典范”（Бобров 134）。列斯基斯（Г. Лесскис）则从 19 世纪 60 年代 7 位作家的 11 部心理小说中随机抽取了 70643 个

1 本文作者在这方面的研究仅限于俄罗斯，期待有学者能系统研究英美等国的文学计量研究发展。

句子进行研究，通过对描述性文字、对话说明、直接引语三大言语类型的统计分析，揭示出不同作家的写作风格。如陀思妥耶夫斯基的人物引语比例明显高于其他作家，显示出作家的“复调小说”特征。托尔斯泰使用描述性文字的比例高于其他作家，显示出其小说的哲理性特征。可以看到，这些研究已具备了当今数字人文的雏形。[1]

雅尔霍的精密方法论研究后继有人。伊万诺夫（В. Иванов）在诗歌节律研究的基础上对诗歌的统计分析做了理论思考。他认为，精密方法有助于发现文学创作的时代特征及个性特征。“诗行统计有助于清晰地显示某种创作手法在今天是否已变得寻常”，“在诗歌研究中采用精密方法的最终目的应该是清晰地揭示诗人在创作中贯彻的一些基本概念，这些概念模糊地存在于诗人对创作性质的直觉认识中”（Иванов 118）。

近几十年来，俄罗斯学者的文学计量研究偏向类型研究和规律性研究。比如文体测量侧重分析作家的个人风格，通过数据深层挖掘，揭示文本的内部结构及其构成规律，探究其不同维度和层面的相互关系。此类研究的结果可以用于作品甄别。安德列耶夫（С. Андреев）则在其专著《诗歌文本参数相互作用模型》（2014）中，集中分析了普希金、莱蒙托夫等俄罗斯诗人以及柯勒律治、济慈等英国诗人的作品，对诗歌文本的节律、句法、词法、词汇单位与主题的相关性进行多维度、多层面的分析研究，揭示出诗体文本各参数之间的相互关系与系统规律方面的特征，阐述了这些参数相互作用的机制，为诗歌文本的计量研究提供了理论与方法。[2]

总体而言，俄罗斯学者对文学开展的计量研究已有相当长的历史，且形成了文学计量研究的重点领域：诗歌格律研究、作家风格研究以及方法论研究。

我国文学研究界也较早就开始关注数学方法在研究中的运用。傅修延在《文学批评方法论基础》一书中介绍了系统论、控制论、信息论及数学方法的基本原则及其运用于文学批评所取得的成果，指出“文学批评要向更高的水平发展，要走向精确化与定量化，就不能不求助于数学方法”（傅修延 296）。数据的实际应用研究则可以追溯到 1987 年，《复旦学报》（社会科学版）发表了李贤平的论文《< 红楼梦 > 成书新说》。作者凭借其数学专业出身的背景，将数理统计方法及计算机技术运用于《红楼梦》的著作权研究。论文选定四十七虚字作为识别特征，对小说各回中这些虚字出现的频率做出统计，采用主成份分析、典型相关分析、类 x^2 距离与相关系数等聚类统计方法对各回进行分类，推翻了红学界盛行六十六年之久的胡适的观点。虽然该文统计方法中择取的数据点受到一些质疑，但毕竟“使红学研究开始有了‘量’

1　参见 Лесскис, Г. А. “О размерах предложений в русской научной и художественной прозе 60-х годов X1X в. ” Вопросы языкознания 2 (1962): 78-95.

2　俄罗斯文学的计量研究是与语言学的计量研究共同发展起来的。参见　王永、李昊天、刘海涛：“俄罗斯计量语言学发展述评”，《外国语》6（2017）：86-97；Андреев, С. Н. Модели взаимодействия элементов стихотворного текста. М.: ФЛИНТА, Наука, 2014.

的概念”（李贤平 15）。不过，此后的若干年内，数据统计方法未能得到更多同行的响应。直至 21 世纪以来，在大数据学术潮流的影响下，我国文学研究界才迈开了文学计量研究的步伐。

21 世纪初，中国古典文学研究者敏锐地意识到计算机技术对于文学研究的重要性，从研究目标出发开始建设相应的数据库。在此后的 20 年间，先后建成了国学数典、文渊阁四库全书、四部丛刊、中国基本古籍库、中国古代方志、中国古代金石等全文数据库。这些数据库具有强大的检索功能，包含海量的古籍善本电子版。使我国的“古典文学研究至少在文献的搜集、整理层面，取得了堪称革命性的突破”（刘成国 132）。利用这些数据库和分析系统以及自建的数据库，古典文学界在影响研究、版本鉴别、考据、用典等方面取得了显著的研究成果。

刘成国在《王安石年谱长编》的研究过程中利用《中国基本古籍库》的《宋会要辑稿》检索系统，“几分钟内，便将王安石长媳萧氏、过继孙王棣、次子王旁、孙王桐、曾孙王璹王珏的相关记载全部查出。然后首次利用阅读南宋文集时偶然发现的王珏墓志铭，以及常见的《至正金陵新志》，得以全面重建王安石身后四代后裔谱系，并澄清了宋代笔记、史书中关于王安石二子王雱、王旁的诸多错误记载，进而对这些记载产生讹误的原因一一分析，抉发出隐含其间的修辞策略及叙事意图”（刘成国 132）。刘京臣则于 2007-2010 年间借助自建的数据库，从字句、用典、意象、意境等角度，以唐朝六大诗人为中心，通过数据分析，考察盛唐中唐诗歌对于宋词的影响。所有结论均建立在数据分析挖掘基础之上（刘京臣 183）。

古典文学研究界讨论较多的还有数据库的信息标注问题，认为更完善的数据库应该包含诸多与作家相关的信息标注，如：作家的出生地、家族背景、科举、游历、仕宦、爱好、作品数量、作品创作时地、文体构成比例、作品选录情况。根据这些信息，可以对研究对象作可视化分析，可以构建地理知识图谱、作家关系图等。另一分析热点是文本情感分析（Text sentiment analysis）。有学者提出“将诗词文本经过语义概念分类，并将情感分为正面情感与负面情感，能使文学研究更趋细化和深化”（罗凤珠 141）。不过，迄今为止，情感分析主要用于互联网的舆情研判、大众点评分析以及各种媒体报道的倾向性分析上。

外国文学界的数据运用研究成果虽然较为单薄，但也有一定进展。相关论文，除了通过“数据”搜索到的 90 篇，加上通过关键词“计量”搜索到的世界文学类论文 18 篇，共计 108 篇。我们对此作了人工校对，剔除其中“敬告读者”“入选 CSSCI”“投稿须知”之类完全不属于文学研究的论文以及同数据运用无关的论文共 43 篇，与数据运用相关的论文为 65 篇。这些论文研究问题所属类别大致可分为六大类：1）大数据与外国文学研究关系的整体思考；2）大数据时代与文学教学；3）基于大数据的文学作品传播与接受度

研究；4）数据库建设构想；5）基于文献数据库的研究；6）文学文本的计量研究。可以看到，只有最后一类属于文学本体研究。此类论文不到十篇。从研究目标看，有的论文旨在阐释文学的创作特征，也有的试图检验计算机技术及其他学科的研究范式在文学研究领域的有效性；从研究角度看，主要依据的是语言学的相关理论，比如语料库语言学、文体学、语义学等；从具体研究路径看，既有根据研究问题采集数据再结合文本做定性研究的，也有利用数据挖掘、检索工具等技术手段获得相关数据再对数据做细致分析的。[1]

综上所述，文学的计量研究不仅有丰富的历史积累，且有可资借鉴的成果。正如研究者所指出，“数字人文的科学方法论和跨学科性质为建构外国文学研究新范式提供了可能性”（董洪川 潘琳琳 176）。

三、文学计量研究的路径与方法

文学的计量研究主要有三种范式：外部研究、文本内部的形式研究以及文本内部的内容研究。从近期的研究成果看，以外部研究成果居多，内部形式研究次之。这是由于外部研究及形式研究涉及的主要是客观知识，这些知识既容易实现数字化，又便于后期的计算机操作，挖掘统计数据。

作家谱系研究、作家关系网构建、文学地理知识图谱、文学发展的地区及时代研究，都属于外部研究。这类研究的重点在于构建数据库，将相关知识数字化，只要有数据库就可以得出统计分析的结果。内部研究的形式方面，如诗歌格律、作家创作风格、用典、版本等研究，其重点同样在于构建数据库。只要将相关的文学文本数字化，借助数据挖掘及统计分析工具，可以完成绝大部分任务。近来的热门话题之一“远读”[2] 也属于此类研究。这类研究对主要依赖文献的研究（如考据研究）冲击最大，传统研究范式需要若干年甚至几十年才能完成的任务，借助大数据和计算机技术，可能不到半天就可以完成。有学者甚至预言，“随着数据库技术从‘机械检索’到‘智能分析’的进步，古典文学研究中的考证范式将面临崩溃”（刘成国 133）。正因如此，这两类研究亦或将成为数字人文首先攻克的堡垒，最终实现数字人文研究者的理想：让机器替代人工阅读和创作，因为“依托数字环境的各种技术，通过加强对文本的批判性集展，版本迭代和文本流动将得以实现”（安妮·伯迪克 30）。当然，这些研究虽然基本能自动完成，但自动化处理的前提是要

1　相关研究可参阅以下文献：任艳、陈建生、丁峻：“英国哥特式小说中的词丛——基于语料库的文学文体学研究”，《解放军外国语学院学报》5（2013）：16-20+127；王永、李昊天：“俄语视觉诗的计量特征——以卡缅斯基诗集《与母牛跳探戈》为中心”，《外国文学研究》5（2015）：48-58；詹宏伟、黄四宏：“大数据时代的文学经典解读——《罗密欧与朱丽叶》计量文体分析”，《外语与翻译》2（2017）：63-68；毛文伟：“数据挖掘技术在文本特征分析中的应用研究——以夏目漱石中长篇小说为例”，《外语电化教学》6（2018）：（8-15）；韩诺等：“基于迁移学习的文学人物心理分析”，《心理技术与应用》7（2019）：63-68 等。

2　参见 Moretti, Franco. *Distant Reading*. London-New York: Verso, 2013.

有相关的数据库，而数据库的构建远非一朝一夕、一己之力所能完成。

第三类，内部研究的内容方面，需要借助数据统计分析，发现某些诗学特征，并进一步结合文本做深入的阐释。目前已发表的此类研究成果大多是技术操作过程的展示有余，诗学特征的阐释不足。借助数据分析发现的某些特征未能进一步运用到文本分析中，深入阐释文学问题，这在一定程度上削弱了计量分析对文学研究的价值。然而，创作主旨、意象、文化内涵、人物情感等内容方面的特征，对文学研究而言恰恰是最为重要的。因此，须重点介绍对文学内容开展计量研究的路径与方法。

此类研究大致有以下几个步骤：1）确定研究问题；2）从语料库（公共语料库；自建语料库）采集数据；3）对数据进行统计分析并得出结论；4）以数据统计分析结论为线索，结合文学文本做深入阐释。

第一个步骤，确定研究问题。这看起来不言自明，任何一种研究都始于问题。但基于数据的研究更须强调研究问题的重要性。鉴于数据有可为与不可为之处，研究者须了解数据统计分析的应用范围。

一般而言，创作特征在语言上有较为明显体现的问题是计量研究的首选。因为语言的多种特征较易标注，而且可以量化。比如未来派诗人致力于艺术实验，试图用诗歌来表现社会，构划未来。为了达到这个目的，他们在创作中大胆对诗歌语言开展实验，以表达某些“基本概念”。视觉诗是其中非常典型的示例，为了表达视觉形象，诗人尝试采用与此相匹配的语言形式。正因如此，未来派诗人卡缅斯基的视觉诗《与母牛跳探戈》是文学计量研究的理想对象。研究者通过数据统计分析，发现了诗人在词类分布、句法结构及语义搭配上的诸多特征，进而揭示出诗人以词语完成立体未来主义构图，践行其“诗画同行”理念的诗学特征。[1] 美国语言诗派的重要代表，查尔斯·伯恩斯坦的诗歌也非常适合计量研究。此外，文学作品中与人物的情感有关的问题也可以做计量研究，因为人类情感不仅有质性的区别，还有程度上的不同。通过对某些情感词进行统计分析，可以研究作品主人公的性格、行为、道德等问题。

第二个步骤是从语料库采集数据。这个步骤涉及两个问题，其一是语料库，其二是采集哪些数据。

语料库是文学计量研究的重要基础。语料库越完善，研究就可以越深入。但语料库的建设需要相关学科研究者与技术人员的合力，要有足够的财力且不说，而且耗时费力还不一定讨好。文学文本语料库的构建非常复杂。在书籍的电子版越来越多、各种软件功能越来越强大的今天，文本数据的倒入可以轻而易举地完成，但仅能搜索到文本的语料库无法用于内容研究。内容研究需要的语料库，需要有语言方面如词法、句法、语义等信息标注。此外，

1 参见 王永、李昊天：“俄语视觉诗的计量特征——以卡缅斯基诗集《与母牛跳探戈》为中心”，《外国文学研究》5 (2015)：48-58。

数据统计分析的准确度和深度，取决于语料库信息标注的准确度和丰富度。目前互联网上开放的各种语料库，大多从上世纪 90 年代开始建设，并且在向公众开放后仍在不断更新迭代。信息较为完善的语料库背后，都有强大的技术团队和众多语言学各分支学科专家的支持。仅有技术，自动标注的数据错误率非常高，需要相关学科的研究专家在对错误进行分析后告知原因，再由技术人员重新修改代码，如此反复无数次之后，再由人工最后校对完成。因此，只有两股力量通力合作，才能构建出能够为语言研究者及文学研究者使用的语料库。在我们自己尚未掌握技术手段之前，可以先借助现有的语料库，挖掘可以挖掘的数据开展相关研究，视需要再自建小型语料库。

与中国古典文学研究界不同的是，外国文学研究具有相对便利的条件。因为国外尚无较好的中国古典文学数据库，所以只能自己建设。外国文学研究则不同，几种主要语言的国家都已建成强大的语料库。比如英国的杨百翰大学语料库（BYU）（https://corpus.byu.edu/），英国国家语料库（BNC）（http://www.natcorp.ox.ac.uk/；https://corpus.byu.edu/bnc/）；美国国家语料库（ANC）（http://www.anc.org/）；俄罗斯国家语料库（ruscorpora.ru）等，都包含了文学作品库。

用于文学文本内容计量研究的数据采集，须从研究目标出发。这个过程需要研究者基于自己的知识储备做出大致判断，做出某种假设[1]，再利用功能较为完善的数据库，获取高频词、词类、句法、语义等方面的相关数据。比如作家研究，可以先选定作为研究对象的作家，即能得到该作家的所有文本；之后，通过输入相关条件，就能得到各种所需的数据。前文提到的诗人卡缅斯基视觉诗的计量特征研究，假设视觉诗在词类分布和句法构成上均有体现。因此，主要采集词类及句法数据[2]。数据提取路径为：进入语料库的诗歌子库；选定诗人卡缅斯基，获得所有文本及其总词数；在“语法特征”框内依次输入名词、形容词、动词等词类为统计条件，获得各个词类的词数；对所有采集的数据进行人工校对，修改错误的数据。而“曼德尔施塔姆诗集《石头》的‘世界文化’网络”研究，先期推测是：曼氏诗作中“世界文化”的构成在很大程度上可以通过对人名地名的分析得出。因此，主要提取这两个语义类别的词语，从诗歌子库中采集曼德尔施塔姆作品中带有“人名（t:hum）”“地名（t:topon）”语义标注的词语，并对获取的词语参照纸质版进行人工校对。

当然，从公众语料库中采集的数据通常不足以完成任何计量研究的既定目标。大多数都需要自建语料库，即在电子文本的基础上，利用文本分析工具采集相关数据，或者人工对所需统计的特征进行标注，再得出数据。比如要对作品的人物做情感分析，须编制情感词表。虽然社科领域的情感分析为

1　这种假设有可能通过研究被证实，也可能被推翻。其实这也是数据统计分析具有挑战的地方。研究可能成功，也可能失败。但这次的失败可以成为下一次成功的基础。

2　因语料库中没有句法标注，该数据经人工标注后再统计。

文学作品的情感分析提供了一定的参照，但其情感分析的指标类别较为简单。绝大部分仅分为积极（正面 / 肯定）评价、消极（负面 / 否定）评价及中性（无明显评价特征）三大类。而文学作品的人物情感非常丰富，仅“爱”就有“喜”“恋”“怜”“好（hào）”“中意”“爱慕”“迷恋”“稀罕”等近义词，以及“如胶似漆”“心心相印”“情意绵绵”“一见钟情”等词语。因此，要做相关研究，须先设定若干指标，根据词典列出同义词、近义词、相同意象词表，再借助计算机辅助的文本分析软件，测量情感词表中单词在样本文本中的词频，进而做出情感计算分析。

第三个步骤，对数据进行统计分析，得出分析结果。

数据提取后，要对其进行统计分析。这是文学计量研究中最为重要，但难度较大的步骤。不少研究者面对获取的数据，看不出任何端倪，不知数据背后隐藏着什么文学密码。解码的过程既需要研究者对研究对象非常熟悉，对文学文本有较深的理解，更需要有较为广博的知识。如果说在大数据分析条件下，词频分布是首要的统计参数，那么对于文学内容研究而言，词频退居次要位置，词语的语义分类更为重要。没有一定的知识积累，就无法对数据做出较为客观准确的分类，也就无法进一步做出数据分析。

以前文提到的“曼德尔施塔姆诗集《石头》的‘世界文化’网络”研究为例，数据采集完成后，人名和地名根据不同分类原则进行分类。人名涉及的因素较多，根据“世界文化”构成这一研究目标，以欧洲文化史、人物性质、文学艺术流派、人物所在国别为分类原则，将所有人名逐一分类；地名主要按地理行政区划来分。结合词频统计做出分析。最后，基于以上数据分析，得出结论：1）《石头》体现出阿克梅派的特征——“对世界文化的眷恋”。诗人笔下的“世界文化”网络，是一个上至古希腊罗马，下至诗人所处时代，以欧洲为中心，辐射到美洲、亚洲和非洲的时空域；2）在诗人的“世界文化”网络中，古希腊罗马文化占有独特地位，其中罗马构成了“世界文化”的核心；3）在“世界文化”网络中，文学艺术构成其最为重要的载体。这个结论构成了后续文本阐释（即第四个步骤）的基础，使得该研究最终得以揭示出曼德尔施塔姆诗歌创作中“世界文化”网络的特征，且所提出的观点和阐释都具有较高的科学性和精密性。[1]

第四个步骤是定性研究，文学研究者非常熟悉，在此不再赘述。

结语

综上所述，计量研究适用于诗歌、小说、戏剧等各种文学体裁，可以从外部研究、形式研究及内容研究等诸多层面进行数据统计分析。一方面，无论在哪个层面，数据都可以为文学研究提供科学的基础和准确的数据，带来

1 参见 王永：“曼德尔施塔姆诗集《石头》的‘世界文化’网络”，《文学跨学科研究》4(2017): 120-131。

定性研究可能无法获得的新发现，使文学研究具有精确性；另一方面，计算机技术也有其局限性，自动统计分析更多地适用于处理共性元素，在大数据分析基础上找寻文学创作的某些规律。然而，个性化永远是文学艺术创作的追求，也是研究者深入探讨的问题所在。因此，文学的计量研究不仅需要继续开展外部研究及文本的形式研究，更需要转向文本的内容研究。除了文本的文化内涵及语言特征，还可以通过情感分析对文学作品的爱情、婚姻、家庭以及道德伦理等问题进行计量研究。

当然，运用计量方法研究文本内容，是定性研究与定量研究的有机结合。这种研究需要研究者不仅具有传统研究所必备的文学知识、思维能力及问题意识，还要具备一定的数据库运用能力。但数据只是起点，而非终点。只有结合文学作品做深入阐释，数据的统计分析才能为文学研究带来既有深度又有精度的高质量成果。

外国文学研究有着深厚的学术传统，且不断推陈出新。在大数据时代，我们有理由相信，文学的计量研究，将使外国文学研究如虎添翼，在传统研究的深厚之上添加准确性、全面性、动态化的特质。那时，“定量研究与定性研究的联袂”，定能使“文学研究结出丰硕的成果”（王永 李昊天 刘海涛 95）。

Works Cited

约瑟夫·E·奥恩：《教育的未来：人工智能时代的教育变革》，李海燕 王秦辉译。北京：机械工业出版社，2018 年。

[Aoun, Joseph E. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. Trans. Li Haiyan and Wang Qinhui. Beijing: China Machine Press, 2018.]

Бобров, С. П. “К вопросу о подлинном стихотворном размере пушкинских «Песен западных славян». ” *Русская литература* 1(1964): 119-137.

[Bobrov, Sergei P. “On the Application of Exact Methods in Literary Studies.” *Russian Literature* 1(1964): 119-137.]

安妮·伯迪克等著：《数字人文：改变知识创新与分享的游戏规则》，马林青、韩若画译。北京：中国人民大学出版社，2018 年。

[Burdick, Anne et al. *Digital Humanities.* Trans. Ma Linqing and Han Ruohua. Beijing: China Renmin UP, 2018.]

雷切尔·萨格纳·布马、劳拉·赫弗曼文：“查找与替换：约瑟芬·迈尔斯与远距离阅读的起源”，汪蘅译，《山东社会科学》9（2018）：45-49。

[Buurma, Rachel Sagner and Laura Hefferman. “Search and Replace: Josephine Miles and the Origins of Distant Reading.” Trans. Wang Heng. *Shandong Social Science* 9 (2018): 45-49.]

董洪川、潘琳琳：“数字人文与外国文学研究范式转换”，《西南民族大学学报》（人文社会科学版）9（2018）：174-179。

[Dong Hongchuan and Pan Linlin. “Digital Humanities and the Paradigm Shift of Foreign Literature

Studies." *Journal of Southwest Minzu University (Humanities and Social Science)* 9 (2018): 174-179.]

傅修延：《文学批评方法论基础》。南昌：江西人民出版社，1986 年。

[Fu Xiuyan. *Basis of Literary Criticism Methodology*. Nanchang: Jiangxi People's Publishing House, 1986.]

Иванов, В. В. "О применении точных методов в литературоведении." *Вопросы литературы* 10(1967): 115-126.

[Ivanov, Vjaceslav V. "On the Application of Exact Methods in Literary Studies." *Topics in the Study of Literature* 10(1967): 115-126.]

Ярхо, Б. И. *Методология Точного Литературоведения. Избранные Труды по Теории Литературы*. М.: Языки слав. культур, 2006.

[Jarcho, Boris I. *Methodology of Exact Literary Studies. Selected Works on The Theory of Literature*. Moscow: Languages of Slavs. Cultures, 2006.]

李贤平："《红楼梦》成书新说"，《复旦学报》（社会科学版）5（1987）：3-15。

[Li Xianping. "A New Theory about Writing Process of A Dream of the Red Mansion." *Fudan Journal (Social Science Edition)* 5 (1987): 3-15.]

刘成国："机遇、挑战与回应——数据库时代古典文学研究中的考证：以宋代为例"，《浙江社会科学》2（2018）：131-137。

[Liu Chengguo. "Opportunity, Challenge and Response." *Zhejiang Social Science* 2(2018): 131-137.]

刘京臣："大数据时代的古典文学研究——以数据分析、数据挖掘与图像检索为中心"，《文学遗产》3（2015）：182-190。

[Liu Jingchen. "Classical Literary Study in Big Data Era: A Case Study of Data Analysis, Data Mining and Image Retrieval." *Literary Heritage* 3(2015): 182-190.]

罗凤珠："引信息的'术'入文学的'心'——谈情感计算和语义研究在文史领域的应用"，《文学遗产》1（2009）：138-141。

[Luo Fengzhu. "Introducing the 'Art' of Information into the 'Mind' of Literature—On the Application of Affective Computing and Semantic Research in the Field of Literature and History." *Literary Heritage* 1 (2009): 138-141.]

任艳、陈建生、丁峻："英国哥特式小说中的词丛——基于语料库的文学文体学研究"，《解放军外国语学院学报》5（2013）：16-20。

[Ren Yan, Chen Jiansheng and Ding Jun. "Word Clusters in British Gothic Novels: A Corpus-based Literary Stylistic Analysis." *Journal of PLA University of Foreign Languages* 5 (2013): 16-20.]

王永、李昊天、刘海涛："俄罗斯计量语言学发展述评"，《外国语》6（2017）：86-97。

[Wang Yong, Li Haotian and Liu Haitao. "Quantitative Linguistics in Russia." *Journal of Foreign Languages* 6 (2017): 86-97.]

郑永晓："加快'数字化'向'数据化'转变——'大数据'、'云计算'理论与古典文学研究"，《文学遗产》6（2014）：141-148。

[Zheng Yongxiao. "Accelerating the Transition from Digitalization to Datamation: Big Data Cloud Computing Theory and Classical Literature Studies." *Literary Heritage* 6 (2014): 141-148.]