北京市社会科学理论著作出版基金资助

JILIANG WENTIXUE DAOLUN

计量文体学导论

施建军 /著

北京大学出版社
PEKING UNIVERSITY PRESS

计量文体学导论

施建军 /著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目 (CIP) 数据

计量文体学导论 / 施建军著 . —北京 : 北京大学出版社，2016.12
ISBN 978-7-301-27872-7

Ⅰ. ① 计… Ⅱ. ① 施… Ⅲ. ① 计量方法—应用—文体论 Ⅳ. ① H052

中国版本图书馆 CIP 数据核字（2016）第 320759 号

书　　名	计量文体学导论 JI LIANG WENTIXUE DAOLUN
著作责任者	施建军　著
责任编辑	兰　婷
标准书号	ISBN 978-7-301-27872-7
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http://www.pup.cn　　新浪微博：@ 北京大学出版社
电子信箱	zpup@pup.cn
电　　话	邮购部 62752015　发行部 62750672　编辑部 62759634
印 刷 者	三河市博文印刷有限公司
经 销 者	新华书店
	650 毫米 ×980 毫米　16 开本　17 印张　280 千字
	2016 年 12 月第 1 版　2016 年 12 月第 1 次印刷
定　　价	56.00 元

前　言

大概在三十年前，还是在上大学的时候，从一本日语语言学的文献中读到有人尝试使用统计学的方法研究有关莎士比亚及其作品争论的课题。这是第一次听说莎士比亚是否确有其人居然还存在争论。联想到中国大量的古典文献也存在类似问题，特别是《红楼梦》的作者问题，不但一直是红学界争论不休的热点，甚至因电视剧《红楼梦》的热播，也成了中国社会关注的对象，于是就想，难道没有一个科学的方法能够解决此类问题吗？恰好当时数学课程正在讲“概率论和数理统计”，便对用统计学方法研究佚名作品的作者问题产生了兴趣。当然，当时并不知道什么是文体学，更不知道还有计量文体学这门学问。我对文体学有系统认识是在硕士研究生时代。当时洛阳外国语学院的张云多教授开设了“文章论・文体论”，这门课系统介绍了文体学这门学问，同时也介绍了日本学者关于文体学研究方面的成果和方法。张云多教授也是我硕士时候的授业恩师。由于计量文体学研究需要进行大量数据的统计分析，而 20 世纪八九十年代获取文本数据比较困难，虽然具备从事这项研究所需的基本数学知识和计算机技术，但是，终因时代和数据条件的限制，这项研究暂时被搁置起来了。但是，我对计量文体学研究的兴趣始终未减，而且一直关注着日本在这方面研究的进展。

进入 21 世纪后，随着信息技术的进步和互联网的普及，数据的获得

比较容易，文本数据的分析和挖掘研究受到广泛关注。世界上计量文体学领域的研究也有了长足的进步，日本就出版了一系列这方面的著作，而且出现了多位这方面研究的专家，比如同志社大学的村上征胜教授、金明哲教授就是这些专家学者的杰出代表。国内虽然也有一些学者在开展这方面的研究，但还是相对比较薄弱，我们甚至看不到一本系统介绍利用计量的方法研究中文文体习惯的专著。

文体计量研究有一个非常相似的研究领域，那就是文章的剽窃研究。国内因学术评价的需要有很多学者在研究学术论文的剽窃问题，这方面的成果非常丰富。学术剽窃问题研究也是研究文章的相似性问题，这和文体研究密切相关但又有严格区别。文章的相似性实际上包括两个方面，一是文章内容和观点的相似性，二是文章写作风格的相似性。通常学术剽窃主要是在自己的文章中抄袭别人文章的内容和观点，为了掩盖其抄袭行为通常会将别人的观点用自己的语言描述出来，说成是自己的。这种情况下，虽然内容观点是别人的，但是由于是用自己的语言表述的，所以存在学术剽窃嫌疑的文章通常是内容观点同别人的相似，但是文章所体现出来的写作风格却与别人不同。当然，如果是不加掩饰的全文抄袭，则不但内容观点相同，而且写作习惯也相同，这种情况是彻头彻尾的剽窃。与剽窃研究不同，文体研究的一个主要目标是要鉴别作品的真伪问题。模仿别人的习惯和风格写作，古来有之，有的是善意的，有的是恶意。如《红楼梦》的续写，作者为了能够让这部不朽之作有一个完整的结局以满足读者欣赏的需要，这个出发点不能说是不好的。而如今充斥网络的匿名文章、匿名信，却没有这样的初衷，这些东西往往会模拟别人的口吻和风格，进行造谣、污蔑和对他人进行人身攻击。这些行为有很多是恶意的。无论初衷是善意的还是恶意的，这些文字产品都会给社会留下困惑，有的需要对其作者进行鉴别。这就需要分析内容不同的文章所体现出来的写作习惯和写作风格的相似性。

研究学术剽窃和研究模拟别人写作风格的作品其实存在实质性不同。学术剽窃主要研究文章内容和观点的相似性，需要考察的对象是文章中反映文章内容和作者观点的语言表达形式以及利用这些语言表达形式来判断论文相似度。而计量文体学研究的对象是文章中能够反映文章作者写作风格和写作习惯的语言表达形式以及以此来判断不同文章是否具有相同的写作习惯、是否出自同一人之手。这两种研究都有非常高的实用价值，前者可以用以鉴别学术不端，而后者可以用以鉴别伪作。

人们普遍使用计算机写作的今天，甄别电子文本的真伪已经不能够仅依靠笔迹这种传统的证据，作者写作习惯的分析将是电子文本真伪分析的重要手段。相信随着大数据理念的提出和数据分析技术的进步，这种用计量的方法进行文体研究的学问将会越来越受到人们的关注，同时计量文体学的方法手段将会在很多领域得到应用。基于以上想法，我觉得自己有责任尽自己的能力将有关计量文体学研究的基本知识和理论梳理出来奉献给国内读者，尽管我在这方面的研究和认识是很肤浅的。2011 年初我入选教育部“新世纪人才支持计划”，作为本人在该计划支持下的重要研究内容，我真正开始了利用计量方法研究文体的工作。经过 3 年多的努力，终于完成了拙著《计量文体学导论》。从统计的角度讲，有很多统计学方法可以在文体计量研究中得到应用，特别是多变量分析的方法层出不穷，本书中所涉及的是最基本的，目的是让读者对计量文体学有一基本认识。关于一些复杂方法的应用读者可在自己的研究中进行深入探讨。文体的计量研究至少涉及语言学、文学、数学、计算机信息处理技术等领域，属典型的跨学科交叉研究领域，限于本人能力和知识的限制，书中难免存在诸多疏漏、不足，希望能够得到广大读者的批评指正。同时，也希望拙著能够起到抛砖引玉的作用，能够吸引更多的学者投入到计量文体学研究领域中来。

2016 年初，承蒙彭广陆教授的厚爱和努力，北京大学出版社接受了拙著的出版申请。在北京大学出版社兰婷老师的鼓励和帮助下，又承蒙彭广

陆教授、陈小明教授的推荐，本书通过北京大学出版社申请了北京市社会科学理论著作出版基金资助并获得了成功。在此向在拙著出版过程中给予帮助的专家、学者和朋友们表示衷心的感谢！本书的出版还与父母、家人的理解、支持是分不开的。特别是妻子和孩子，正是因为有她们在后面默默的付出和努力，我才得以专心致力于此项研究，顺利地完成书稿的写作。值此书出版之际也向亲人们表示由衷的感谢。

施建军

2016 年 6 月 19 日于北京

目　录

第五章 文体计量分析中的参数估计问题

第六章 文体特征差异的假设检验

第七章 文体风格个体性差异的方差分析

第八章 文体特征的多变量分析

第九章 支持向量机技术和文学作品作者鉴别

第一章　绪论

第一节　什么是计量文体学

计量文体学（stylometrics, computational stylistics）是研究如何用统计学的方法分析文学作品的文体特征的学问。计量文体学是文体学的一个重要分支，横跨文学、语言学、数学、计算机科学等众多学科，为文体研究提供精确、科学的测量方法。计量文体学的任务主要是解决以下几个方面的问题：

1. 特定作家风格的精密计算和描述。

2. 宏观文体（包括新闻报道、广告、科技说明文、学术论文等功能文体；小说、诗歌、散文等文学文体；口语、书面语、网络语言等语体）的特征分析和归类研究。

3. 佚名作家作品的作者鉴定。

4. 作品年代的测定。

5. 作家文体的变化及同一作家作品先后顺序的测定。

目前，计量文体学在国内学界有如下多种提法：计算文体学、计算风格学、统计文体学。笔者还是认为用计量文体学更为精确一些。计量文体学不仅要对作家作品的文体特征进行统计，而且要在此基础上用统计学的原理对作家、作品的文体特征进行分析，甚至要对有效利用文体特征进行

分析的统计理论以及统计工具进行研究和开发。另外，计量文体学和计量经济学的情况基本相似。计量文体学和计量经济学研究中所使用的统计学理论、方法、工具大多是共通的。计量经济学的概念已经成为家喻户晓、耳熟能详的术语，所以使用计量文体学更容易为大家所接受。

第二节　国内外计量文体学发展的历史和现状

用统计学的理论方法研究作家的文体在国外可追溯到19世纪。《新约圣经》中有“罗马书、哥林多前书、哥林多后书、加拉太书、以弗所书、腓立比书、帖撒罗尼加前书、帖撒罗尼加后书、提摩太前书、提摩太后书、提多书、腓利门书、希伯来书”等14封保罗写给各地教主的书信。这些书信是否均出自保罗之手，历史上一直存在争议（村上，1994）。尤其是最后一封“希伯来书”，由于现存《新约》的“希伯来书”中没有“保罗致……”字样，有人认为这封书信很有可能不是保罗的作品。因此，保罗书信作者的鉴定一度成为学界的热点问题。最初提出用数学方法证明此问题的是英国著名数学家、理论代数奠基人德·摩根（Augustus de Morgan，1806—1871）。

1851年，德·摩根在给剑桥牧师W. Heald的一封信中提出，每个人的文章都有自己的个性，即便是思维相近的两个作家，其作品或文章中单词的平均词长总是或多或少地存在着差别，同一个人的不同作品的平均词长的差别总是要比不同人所做的内容相同的作品的平均词长的差别要小得多。因此，德·摩根认为用这种办法就可以进行作品真伪的鉴定。

1887年美国地球物理学家门登荷尔（T. C. Mendenhall）受到德·摩根思想的启发，认为词长能够反映作家的写作习惯，就像光谱能够反映各种颜色的光的特点一样。如果能够获取这种“词谱”就能够确定某一部作品的作家。并认为“词谱”能够给作家考证提供科学的解决办法。他利用这

种方法对比研究了莎士比亚 40 万词、培根 20 万词的作品，获得了反映这两位作家写作习惯的不同的“特征曲线”，从而解决了当时有关莎士比亚和培根是否是一个人的争论，并且在《科学》杂志上发表了论文。同一时期欧洲也有许多学者在从事着同样的研究。由于这种研究需要进行大量的统计分析，受到研究手段的限制，Mendenhall 时代的统计文体学研究是一项艰苦的工作。

第二次世界大战以后，随着计算机的出现和统计学理论的发展，文体的统计研究也有了较大的发展。这一时期比较有名的研究成果是瑞典文史学家 A. Ellegard 关于《Junius 投稿集》的研究。《Junius 投稿集》是 1769 年至 1772 年英国报纸上发表的笔名为 Junius 的人所写的攻击英国政府和王室的一系列文章。这些文章的作者到底是谁一直是英国文学史上的谜。1962 年 A. Ellegard 发表了《作者考证的统计方法》一书，书中 A. Ellegard 统计了 Junius 比同时期作家使用得更多的词汇和不怎么使用的词汇以及 Junius 对同义词的选择倾向，然后同当时被怀疑为 Junius 的 40 名作家一一进行对比。最后发现 Junius 的写作习惯和 Philip Francis 的习惯惊人一致，因此 A. Ellegard 认为他的统计证据有 99% 的把握可以证明 Junius 和 Philip Francis 是同一个人。

20 世纪中后期，随着计算机的普及，统计文体学的研究特别是利用统计文体学方法进行西方语言文本的研究已经不像此前那样高深莫测。开始有人用统计文体学的方法研究文学作品的伪作问题。在英国，计量文体学考证作者的方法甚至被警察用来判别自首书的真伪。70 年代中期，英国剑桥大学的两位师生曾经运用统计文体方法和计算机技术侦破了出版商伪造莎士比亚作品的案子从而震动西方文学界（贾洪卫等，1991）。80 年代，在日本，华岛忠夫、寿岳章子两位学者利用统计学的方法研究了 100 多名日本作家的写作风格，并出版了《文体的科学》一书。90 年代，日本学者村上征胜运用多种统计手段对被誉为世界上最早的小说《源氏物语》的作者

存疑问题进行了研究，于1994年出版了专著《真赝的科学》。

进入21世纪后，随着信息技术的进步，特别是自然语言处理技术在汉语、日语自动分词等方面取得了突破性的进展，国外有学者开始利用新的信息技术研究中国古典文献。如日本的石井公成（2002）、师茂树（2002）、山田崇仁（2004）等。山田崇仁利用自然语言处理中的N-GRAM和文本挖掘技术中的聚类方法对我国先秦时期诸子百家留下的历史文献的成书年代进行了探索。石井公成、师茂树等学者用同样的方法对佛教经典的真伪进行了研究。

受到西方研究方法的影响，我国学者真正开始用统计文体学方法研究中国古典文学作者问题始于20世纪80年代初。由于计量文体学涉及数学方法，加上计算机对中文处理能力的限制，尽管中国古典文学作品作者问题存在许多奇案，但是利用计量文体学方法研究中国文学作品作者问题的学者并不太多，成果数量也有限，且主要集中在《红楼梦》的研究上。

根据前文论述可以知道，使用统计方法进行文学作品作者的考证在西方取得了令人信服的成果。而使用同样的方法对《红楼梦》的研究却得出了截然相反的结论。这一方面说明《红楼梦》这部作品的复杂性，同时也让人怀疑在中国古典文学作者的考证研究中计量文体学的方法是否使用得当。自1987年陈大康先生发表《红楼梦“成书新说”难以成立》一文，提出与李贤平商榷以后，至今已经有20多年。这二十多年似乎这方面的研究陷入了停顿，很少能够看到这方面文章的发表。

可以说我国在计量文体学研究方面和世界先进水平还是有一定差距的。这种差距表现在以下三个方面。

一是我国计量文体学研究的现状和社会现实需求存在着很大的距离。我国古典文献的作者问题一直是困扰学界的热点问题，至今没有得到科学的解决。在现实生活中，随着计算机和互联网的普及，计算机输入已经取代了用笔写作的习惯，这又给我们提出了如何科学鉴定电子作品作者的课题。

二是计量文体学研究成果的数量存在很大差距。欧美这方面的研究起始于19世纪，而我国20世纪80年代之前基本没有这方面研究成果。即便是现在，针对中文文献进行文体计量研究的原创性论文也非常少。而根据日本学者金明哲、村上征胜在『言語と心理の統計』中提供的资料，截至2002年欧美有关文体计量学和作家鉴定方面的英语论文（著作）有100多篇（部），日本约50余篇（部）。

三是尚未找到汉语文体的有效测量方法。文体的测量方法和指标，根据语言的不同呈现出其独特性。词长分布在进行英语文献的计量分析时能够收到很好的效果；日语助词和标点的组合情况能够有效地反映日语文献的文体特征。但是这些特征指标很难在汉语文体测量上发挥有效的作用。我们必须要找到古代汉语和现代汉语的文体特征指标。

第三节 文体的计量特征

计量文体学作为完整的体系，其研究包括理论和应用两个层面。从应用层面讲，计量文体学主要解决文献和文学作品的那些与文体相关的实际问题，如：文学作品的风格差异分析、佚名作者的考证、作品剽窃的鉴定等等。我们所说的计量文体学理论层面的研究主要是指文体特征的把握研究和利用文体特征进行分析的统计学方法研究。这里的统计方法研究是指，如何利用已知的文体特征载体更加精确、更加快速、更加简便地计算分析文体的差别之所在，也就是找到更加合理的数学方法和理论，使得依靠这种数学方法和理论所开展的文体分析更加可靠和简便。这种理论研究主要突破点在数学方面，不属于人文研究的领域，因此，这里不对此做过多涉及。

但是，作家或者作品的特定风格或者是文体的主要载体是什么？这是文体学研究的最基本的问题，也是计量文体学的出发点。计量文体学的所有的统计分析必须建立在能够充分反映作家或者作品的写作风格的文体特

征上。因此，文体特征的把握和分析，是我们必须要重视和深入探讨的计量文体学重要研究领域。根据『文章の計量』，有学者认为能够用于文体测量的文体特征项多达 500 多种（アンソニーケニイ，1996:13）。但是在文体测量中经常被采用，被认为是有效的文体特征项却很少，而且根据语种的不同，能够反映文章作者写作风格的语言特征既有共性，也有具有与语种相对应的独特特性。这里介绍几种学界已经归纳出来的文体特征。

1.3.1 文体的词长特征

计量文体学启蒙阶段，德・摩根认为作品的平均词长能够反映作家的写作特点，同一作家的不同作品其平均词长十分接近，而不同作家的作品的平均词长相差却很大。德・摩根以两位古希腊历史学家希罗多德（Herodotus，约公元前 485—约公元前 425）和修昔底德（Thucydides，约公元前 460—公元前 400）的著作为统计对象，对这两个作家用词的平均词长进行了统计。希罗多德著作第一卷的平均词长为 5.624 个字符，第二卷的平均词长为 5.619; 而修昔底德著作的第一卷和第二卷的平均词长分别是 5.713 和 5.728。可见同一个作家的作品的平均词长是非常接近的，而不同作家作品的平均词长的差距要比同一作家作品间的平均词长的差距大得多。德・摩根对《新约》圣经圣保罗的前 13 封书信的统计结果是，其平均词长为 5.428，而书信《至希伯来人》的平均词长为 5.516。由于平均词长差距比较大，所以德・摩根认为，根据这个结果可以认为《致希伯来人》出自另外一个人之手。德・摩根的思想比较朴素、简单，但是，现在看来用这种差别来衡量作家的写作风格或者是文体特征的差别还是十分粗糙的，特别是当研究对象涉及多个作家的作品时，仅以平均词长恐怕很难区分出不同作家。

门登荷尔也认为作家所使用词汇的词长能够反映作家的写作特征。但是，门登荷尔所利用的词长特征不是简单取作家的平均词长，而是使用作家词长的分布特征来衡量作家的文体特征的。1887 年门登荷尔在《科学》

杂志上发表论文指出，可以根据词长及其出现的频率描绘特定作品的词的分布图，就像用光谱可以描述光的特征一样，用这种词长的分布——词谱可以分析文章的文体特征。门登荷尔在对莎士比亚的作品进行研究时发现，莎士比亚的作品无论是诗还是散文，其词长分布曲线是一致的，均呈现出莎士比亚独特的文体特征，莎士比亚作品中词长为 4 的单词出现频率最高，这与莎士比亚同时代的作家有明显的差别。此外，门登荷尔还对狄更斯、萨克雷、丹尼尔·笛福等多个作家的多部作品进行了统计分析，结果均表明词长的分布特征可以反映作家的文体特征。

但是，1975 年威廉姆斯（Williams）在对门登荷尔的结论进行验证研究时发现，同一作家不同体裁的作品，如诗歌和散文，词长的分布也有可能不一样。威廉姆斯以莎士比亚、培根、锡德尼（Philip Sidney，1554—1586）为例，调查了莎士比亚的诗歌、培根的散文、锡德尼的诗歌和散文的词长分布。下图为其词长分布曲线图。

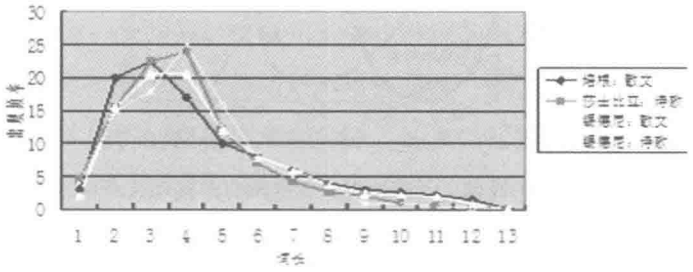

图 1.1 培根、莎士比亚、锡德尼三位作家散文、诗歌词长分布

英语等西方语言，单词长度的取值范围比较大，其分布的多样性足以区分不同作者。同时，由于西方语言单词之间存在明显的界限，这也为利用词长分布作为文体特征进行文体的计量分析提供了很大的方便。但是，词长的分布能否有效区分汉语和日语这样的东方语言作家的文体是一个值得研究的问题。

由于日语书面语的连续书写特性以及计算机分词处理技术的限制，日本学界很少利用日语词长分布进行文体研究。但是，为了验证日语词长分布在日文文体特征的区分上到底是否有效，日本学者金明哲等还是在这方面做了一些尝试。

根据金明哲等著『言語と心理の統計』，金明哲选取了井上靖、中岛敦、三岛由纪夫等三位日本作家的作品为对象，用主成分分析的方法对这三位作家作品中的所有单词的词长信息进行了分析。以第一主成分的得分作为横轴，第二主成分的得分作为纵轴，绘制了三位作家作品的散点图。结果三岛由纪夫的作品和井上靖的作品没有能够有效地区分开来。

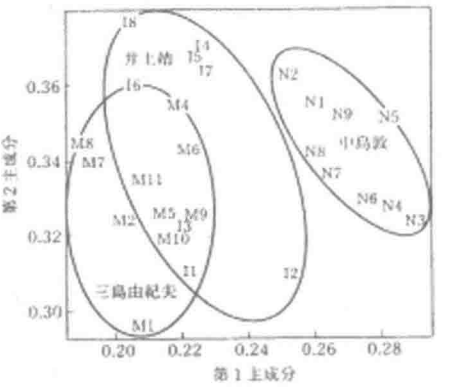

图 1.2　井上靖、中岛敦、三岛由纪夫作品所有单词词长主成分分析图

由于有些词汇和文章内容存在密切的关系，如果用词长作为文体特征时，采用较多的与文章内容关系紧密的词汇信息，则不能很好地区分作品文体风格。这也是以文章中出现的所有单词词长为依据不能够很好区分日语文体风格的重要原因。为了克服这个问题，金明哲等利用与文章内容关系比较弱的动词的词长为依据，用同样的手法对上述三位作家的作品进行了主成分分析，结果发现日语文章中动词的词长能够有效地区分不同作家的写作风格。