第一章 语料库文体统计学简史

本章将向读者简明扼要地呈现语料库文体统计学的发展概貌。我们将从语料库文体统计学的几个基本概念开始，分两个阶段梳理现代文体统计学的发展简史——语料库介入前文体统计学的发展和语料库介入后文体统计学的发展趋势。

1.1 什么是语料库文体统计学？

本书中“语料库文体统计学”（Corpus Stylo-statistics）是指一种基于语料库对特定类型文本的语言特征进行分析的量化研究方法。显而易见，这种研究方法涉及三个不同学科——文体学、统计学和语料库语言学的概念、理论、工具和方法。本节将简述这三门学科的联系、区别及语料库文体统计学的主要应用领域。

文体学是一门研究文体的语言学分支学科。文体是指不同情境下语言的使用，如书面体、口语体、文学文体、非文学文体等。文体分析则是指对文体的语言特征和功能进行分析的方法和过程。早期的文体学主要关注文学文体和作家的写作风格，与文学研究和文学批评联系紧密。现代文体学则在日趋系统化的同时，越来越注重量化分析方法。对语言现象的量化

和统计分析随着计算机和网络信息技术的进步而更加便捷和精确。对文体特征的量化研究催生了“统计文体学”（Statistical Stylistics）或“计量文体学”（Quantitative Stylistics）。随着计算机和语料库的产生和应用，“计算文体学”（Computational Stylistics）或“语料库文体学”（Corpus Stylistics）得以发展。

随着统计学方法在文体分析中的应用，一种专门对特定文体类型的特征进行统计分析的研究方法产生了，即“文体统计学”或“文体测量学”（Stylo-statistics或Stylometrics）。这种方法是统计文体学或计算文体学的基本研究方法。统计学方法在语言学和文体学中的应用始于19世纪下半叶，以逻辑代数创始人伦敦大学的奥古斯特·德·摩根（Augustus de Morgan）在1851年的一封书信中提出以“单词的平均长度”来判别《新约》中保罗书信的真伪为标志。160多年来，文体统计分析在作者辨别、文本体裁分类、文本观点和态度分析、刑侦语言学、数据挖掘等领域取得了长足的发展，并越来越多地应用到社会学、心理学、人类学、历史学、翻译学等其他人文社会科学领域。这些人文社会科学学科的一个共同点是都需要对语言信息进行分析。这些学科可以从文体学中汲取两个方面的知识：一是关于描述特定文本所需要的语言特征和层次；二是对这些特征进行分析的具体方法，尤其是文体统计学的方法。

文体统计学与语料库语言学（Corpus Linguistics）具有自然的联系，二者都是基于对真实自然语料的语言特征进行的量化分析，都依赖于严格的统计学方法。两者的区别在于：文体统计学主要用于分析某一特定类型的文体的具体特征或个别作家的写作风格，通常使用较少的语料。语料库语言学则主要利用语料统计数据研究语言总体趋势、验证特定的语言学假设或建立语言学理论，通常使用较大规模的语料库。前者在具体特征的统计分析方面比后者更为成熟，而后者可以弥补前者在语料规模和语言特征识别两个方面的局限。因此，本书用“语料库文体统计学方法与应用”来命名这种基于具有代表性的语料库对语言特征及其功能进行统计分析的研究方法。

1.2 现代文体统计学的发展

美国统计学家大卫·霍姆斯（David Holmes）[1]在其论文《文体统计学在人文学科中的演变》（“The evolution of stylometry in humanities scholarship”）中梳理了统计方法用于文体分析的历史，勾勒了文体统计学发展的脉络。霍姆斯将文体统计学的发展划分为起源时期、突破时期、挣扎时期、多变量时期、累积和时期以及人工智能时期等六个时期。需要指出的是，这六个时期只是一个粗略的划分。实际上，各个时期之间的界限比较模糊且有相互重合之处。以下按这六个时期的顺序回顾统计学方法在文体分析中的应用情况。

1.2.1 起源时期

现代文体统计学起源于19世纪中叶。英国逻辑学家德·摩根1851年在致友人的一封信中提出，可以根据作品的某些统计特征来区别不同作家的文体，特别是词语的平均长度可作为作家文体风格的标志特征。同一时期，还有一些学者关注文体特征的统计分析，并将这种方法称为“文体测量学”（Stylometry或Stylometrics）。他们用平均值或百分比来计算特定词语反复使用的次数、诗行中格律的变化等。这些早期的文体统计学研究大都围绕莎士比亚的作品展开，以1874年“新莎士比亚学会”的成立为高峰，如Fleay（1874）“莎士比亚戏剧诗的格律实验”和Ingram（1874）“莎士比亚诗歌的非重音结尾”等。这些研究的主要成果是：人们发现莎士比亚从1589年26岁时起至1612年48岁时止，在22年中所创作的36部剧作的文体风格发生了缓慢但持续不断的变化。“文体测量学”这个名称也为德国学者Dittenberger（1881）所用，他试图用词频，主要是功能词词频，来确定《柏拉图对话录》的作者和年代。

1 本书中外国人名、作品名或其他专名的翻译采取“遵从定名”的原则，采用通行的既有译名。对尚未翻译的人名、文献出处的作者名和作品名，为便于读者查询原始文献，采取保留原文的方式。

德·摩根的文体统计学思想稍后也得到了美国地球物理学家门登霍尔（T. C. Mendenhall）的响应。门登霍尔认为，相较于德·摩根用词长算数平均数，不同长度的词语的分布更能说明文体的变化。Mendenhall（1887）的论文《作品的特征曲线》（"The characteristic curves of composition"）是朝向现代文体统计学迈进的重要一步。这篇论文发表于1887年的《科学》期刊，从词长分布的角度比较了狄更斯和萨克雷的写作风格，并使用了统计图示的方法。1901年，门登霍尔在其另一篇论文《对一个文学问题的机械解答》中，回答了一个文学史上的疑问：以莎士比亚为名发表的《哈姆雷特》等戏剧是否是同一时代的散文家和哲学家培根所作？门登霍尔通过对莎士比亚戏剧和培根的作品的词语长度统计分析发现，莎剧的每部作品四字母词总是多于三字母词，而培根的作品总是三字母词多于四字母词，且更偏好于使用更长的词，因此培根不是莎剧的作者。

真正有重要影响的现代文体统计学研究直到20世纪30年代后才产生。首先是美国语言学家乔治·齐普夫（George K. Zipf）于1932年提出的著名的"齐普夫律"——在自然语言语料库中，一个单词出现的次数与它在词频表里的排名成反比。具体来说，一个词在词频表中出现的序号（rank，简称R）与该词的出现频次（frequency，简称F）的乘积几乎是一个常数（constant，简称C），即R × F=C。根据这个定律，自然语言语料库的词频表中排名第一的词（如"*the*"或"的"）的词频是排名第二的词的两倍，是排名第三的词的三倍，以此类推。

其次，英国统计学家犹勒（G. U. Yule）1944年提出了一种新的衡量词汇多样性的统计方法，即"犹勒特征常量K"（Yule 1944）。特征常量K是一个文体参数，由于K值不受样本大小的影响，因此常常被用于测量文本的词汇密度。Yule（1939）还提出用句长来辨别作者身份，他对《效仿基督》（*De Imitatione Christi*）进行了研究，但该论文的结论是依据句长判断作者身份并不完全可靠。Williams（1940）发现如果将每个句子的平均词数取对数排列，每位作者的句长对数值近似于正态分布。这一结论在

Wake（1957）对希腊作家的研究中也得到了佐证。

另一个使用文体统计学方法解决作品年代问题的例子是Cox & Brandwood（1959）。这两位学者试图确定柏拉图作品的先后次序，他们研究了这些作品中每句话末尾五个音节的分布情况，将音节划分为长音节和短音节两类。结果发现《理想国》和《律法》的长短音节分布明显不同，因而可以根据与《理想国》的相似程度来排列其他作品的先后次序。

1.2.2 突破时期

文体统计学再次取得令人信服的突破是在20世纪60年代初。两位美国统计学家Mosteller & Wallace（1964）研究了著名的《联邦党人文集》（*The Federalist Papers*）的作者身份问题。这些通信于1787—1788年间匿名发表，旨在号召纽约州人民通过美国宪法草案。有三位作者后来公开承认写作了这些信件，分别是Alexamder Hamilton、John Jay和James Madison，但全部85篇通信中有12篇Hamilton和Madison都宣称是自己所作，因而作者身份存在争议。Mosteller和Wallace采用了介词、连词和冠词等功能词作为判别作者的标志，例如，upon一词在Hamilton的其他文章中使用的频次是每千词3.24次，而在Madison的文章中仅0.23次。两位研究者用概率来表示对作者身份（如Hamilton是第52篇通信的作者）的确信程度，并根据已有证据用贝叶斯定理（Bayes' theorem）来调整概率值。研究结论是所有有争议的12篇通信全为Madison所作，这与历史学家的看法大体一致，不过Mosteller和Wallace也申明这些通信的政治背景与Hamilton和Madison写作风格近似，判别作者十分困难，他们的研究着重统计方法的应用而不是彻底解决争议。但无论如何，这一研究所引起的反响极大地提高了人们对文体统计学方法的信心，它不仅是作者身份辨别领域的突破性贡献之一，而且为文体统计学打开了一扇现代计算机时代的大门（案例详见本书6.3.2节）。也正因如此，《联邦党人文集》后来常常被用来当作文体统计学新方法、新技术的“实验场”，不少研究以此为蓝

本来进行实验(如Holmes & Forsyth 1995; Martindale & McKenzie 1995; Tweedie *et al.* 1996等)。

1.2.3 挣扎时期

20世纪60至80年代可以看作文体统计学的“挣扎时期”。这一时期，文体统计学方法不断争取获得传统人文学科的认同，但其方法和结论却总是遭到传统学科的质疑和攻击。这一时期，文体统计学家们开始利用电子语料库提供的丰富数据进行词汇层面的分析。自齐普夫律以来，各种各样的数学模型建立起来，但有影响的实证研究却变得很稀少，文体统计学的主要内容似乎变成了研究方法的争论。

历史上最激烈的文体统计学之争莫过于关于“莫顿法”(Morton's method)的争论(Holmes 1998：113)。莫顿法通过对英语作品中词汇位置、搭配和词对进行检验以辨别作者身份。在词汇位置检验中，研究者对比存疑作品中前置词的出现次数以及存疑作品中前置词的出现次数。词汇搭配检验主要考察两个规定单词之间的先后顺序，而词对检验则比较文本样本中某两个规定单词之间的用法关系。莫顿法被Merriam(1979，1980，1982)运用于三项关于莎士比亚的研究中。与此同时，对莫顿法的批评也一直不绝于耳。Smith(1985a，1985b)认为莫顿法的数据收集存在问题，样本数量有限，缺乏严谨性。Smith与Morton之争的另一个焦点集中在对文本中一次性词(*hapax legomena*)的研究上。Morton在其1986年发表的文章中提出，研究作家作品中只出现过一次的词语有助于辨别作家身份。此类词语出现在句子结尾处的概率明显高于其出现在句子开头处的概率。针对Morton的观点，Smith于1987年撰文回应。他认为Morton的研究缺乏有力证据，无法自圆其说。Smith与Morton之争在一定程度上说明，学界对能否将统计学方法运用到人文学科的研究之中尚存疑虑，所以这段时期也被认为是文体统计学的“黑暗时期”。事实上，Smith与Morton之争持续数年之久，可谓硝烟弥漫、热度不减。

另一场文体统计学之争源于英国著名统计学家罗纳德·费希尔（Ronald Fisher）1943年提出的概率计算方法。费希尔将该方法用于预测新发现的蝴蝶种类及蝴蝶标本的数目。受此启发，20世纪70年代中期美国统计学家Bradley Efron和Ronald Thisted用费希尔的概率计算法研究了莎士比亚的作品。他们认为，与预测新发现蝴蝶种类的过程相似，莎士比亚作品中既有已发现词汇，亦有待发现的词汇。“已发现词汇”是指出现在作者作品中的词汇，而“待发现词汇”则是指作者知晓但未用于其作品中的词汇。此后，Thisted & Efron（1987）研究了匿名诗作“Shall I die”的作者归属问题。经过计算，他们认为该诗作中至少包含7个之前从未出现过的词汇。研究结果表明，“Shall I die”一诗中出现了9个这样的词汇，数量超过之前的预期（至少7个），据此，他们得出了该诗作出自莎士比亚之手的结论。但此项研究遭到学者Robert Valenza的质疑。Valenza于1991年撰文指出，Thisted和Efron对“Shall I die”一诗的作者归属问题研究证据不足，效度不够。至此，文体统计学再次遭到学界质疑。

1.2.4 多变量时期

20世纪80年代末90年代初，澳大利亚学者Burrows（1987a，1987b，1989，1992a，1992b）发表了一系列影响深远的论文，确立了文体统计学方法在作者身份辨别中的重要地位。尤其是1987年出版的《文学批评中的计算：奥斯丁小说研究和方法实验》（*Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*）发现了功能词在揭示语言模式中的重要作用，首次将主成分分析法（Principal Components Analysis，简称PCA）用于文体分析，从而奠定了计算文体学的基础。Burows的研究采用了多个语言特征和多元统计分析方法，因而这一时期被称为“多变量时期”。

Burrows（1987a）对大量（超过75个）常用高频功能词进行了研究，计算出这些高频功能词的每千词词频，采用主成分分析法来分析这些高频

词所构成的多变量数据集。主成分分析法是一种多元统计分析方法，其主要目的是将一组数量众多的可能存在相关性的变量转换为一组数量较少的线性不相关的变量。转换后的变量叫主成分，主成分按重要性降序排列，只需取排在前面的少数几个主成分就能反映出原来众多变量才能反映的信息，由此可以简化问题的维度。最常见的做法是，只取前两个主成分，就可以绘制出全部数据的二维图。主成分二维图由许多点构成，每个点代表一个文本。聚在一起的点阵表示这些文本具有共同的趋势，而点阵外的点表示不具备这种趋势。

Burrows（1987b）用这种方法分析了很多作家和不同体裁的作品，其中包括奥斯丁、勃兰特姐妹、司各特和拜伦的不同体裁的作品。他的研究结果令人振奋：很多作家在许多常见的功能词（如by、the、from、to等）的使用上具有鲜明的特征。功能词的使用往往是下意识的，Burrrows的研究似乎找到了文体统计学家们一直在努力寻找的有效的量化方法。这也使得主成分分析法现在几乎成了文体统计学方法进行作者辨别的“排头堡”。不少文体学家将这种方法运用到自己的研究中，如Holmes & Forsyth（1995）用主成分分析法再次研究了《联邦党人文集》的作者归属问题；Baayen *et al.*（1996）将主成分分析法用于研究语料库中的句法标注。

Burrows还提出了一种具有代表性的作者身份鉴定的文体统计学方法，即Delta分析法。这种方法的主要理据是：选择某位作家的作品中出现频率最高的80—150个文体标记，多为边缘性词汇，包括功能词、非主题名词或动词等，计算出每个高频词的Z值（z-score），然后对要鉴别的文章进行匹配性计算，各项Z值达到一定的总值，就可判定该文章的作者就是这位作家。Burrows用他的方法对简·奥斯丁（Jane Austen）小说进行了研究，主要利用对比小说人物对话和叙述中的情态动词的频次，揭示其所表现的意义和价值观。

Burrows的主成分分析法开启了采用多元统计方法进行文本量化研究之门。继此之后，多元统计方法几乎成为了文体统计学研究的必用方法。

除上文提到的主成分分析法外，其他常用的多元统计分析方法还有聚类分析法（cluster analysis）、判别分析法（discriminant analysis）和对应分析法（correspondence analysis）等。

聚类分析是一种归纳式多变量统计手段，通常根据高频词频次的异同将风格类似的样本进行归类。该方法常常被用于作家风格辨别、作家身份甄别以及作家风格历时变化的研究之中。Hoover（2007）采用聚类分析法研究了作家亨利·詹姆斯（Henry James）写作风格的历时变化，研究结果表明詹姆斯的前期作品与后期作品呈现出两种不同风格（该研究案例详见本书4.3.2节）。有的研究者将聚类分析法与主成分分析法结合起来，此类研究包括：Holmes（1992）的摩门经文研究、Frischer *et al.*（1996）的拉丁古典文本研究、Mannion & Dixon（1997）对十八世纪文本的研究等。

使用判别分析法的研究有Ledger & Merriam（1994）对莎士比亚作品的研究、Mealand（1995）对圣卢克福音的研究等。使用对应分析法的研究有Dixon & Mannion（1993）对Goldsmith的研究、Mealand（1995，1997）对新约福音书的研究、Tabata（2002）对狄更斯风格变化的研究等。研究中多变量方法的使用具有十分重要的意义。首先，多变量方法使研究者分析自己研究结果时能够更加得心应手、游刃有余。其次，随着多变量方法在研究中的广泛运用，文体统计学的研究前沿向前推进了一大步，有利于文体统计学在人文学科中地位的确立。

1.2.5 累积和时期

“累积和图”（cumulative sum charts或cusum charts）在文体统计学中的应用始于20世纪90年代初。累积和图原是一种用于工业生产过程中质量控制与监控的统计技术。Morton & Michaelson（1990）及Morton（1991）提出了将累积和图用于作者身份识别的设想。Morton认为任何人在书面表达或口语表达过程中都会养成一套独特的习惯。这些习惯体现为

句子中某些特有的成分，而这些成分是可以被量化分析的。在Morton的研究中，这些特有的成分主要包括短词（由2—3个字母组成的词）、元音词（由元音开头的词）和短词与元音词组合。每个人的习惯成分的出现频次大都具有一致性，一旦频次发生显著变化，就有理由怀疑出现频次变化的句子出自不同人之手。与此类似，对于一个作者身份存疑的文本，一旦该文本中某些特有词汇的出现频率与疑似作者作品中的词汇频率具有显著差异时，就可以断定该文本并非出自疑似作者之手。累积和图方法首先要求生成两张图，第一张图代表句子长度，第二张图代表每个句子习惯词汇的出现次数。然后将这两张图叠加在一起。对同一作者而言，两张图中的两个数值（即句子长度和每个句子习惯词汇的出现次数）应当彼此平行。一旦两个数值出现显著差异则表明文本出自不同作者之手。辩护律师们对Morton所提倡的累积和技术表现出极大的兴趣，并将该技术用于判别当事人供词。事实上，当时几个备受关注的法庭案件也采用了累积和技术。

与此同时，人们也开始质疑累积和技术作为一种司法技术的有效性。当时英国广播公司（BBC）的《明日世界》栏目以及第四频道的《街头法律》节目还对此进行了报道。一时之间，Morton的累积和技术走到了舆论的风口浪尖，面临遭受公众嘲笑的危险。几个研究团队就此课题开展了一系列独立研究，研究结果均表明累积和技术缺乏可靠性（Canter 1992；de Haan & Schils 1993；Hardcastle 1993，1997；Hilton & Holmes 1993；Holmes & Tweedie 1995；Sanford *et al.* 1994）。甚至有人将论证累积和技术缺乏科学依据的研究报告寄到了英国皇家检察署。反对者的声音集中于以下两点：其一是累积和图本身的解释具有较强的主观性；其二是关于一个人习惯词汇的一致性假设并不成立。然而，该技术的支持者们一方面对该技术信心满满，另一方面也在为该技术的有效性寻找证据。随着技术的不断进步，累积和技术统计方法更加严谨，统计结果也更加可靠。尽管如此，人们还是认为将该技术用于鉴别作品、文件及法律供词作者的做法不能令人信服。

即使在一片质疑声中，累积和分析法也并未销声匿迹，相反，它依然时常出没于世界各地的法庭之中。法官在辨别作者身份时常常采用这一颇受争议的方法，这也是该方法备受关注的原因之一。事实上，法庭面临采用新方法提取可靠证据的问题。正如Hardcastle（1997）所说，“累积和法的应用并未达到合格司法技术的标准，因此，以往试图通过语言工具判别作者身份的司法科学家还需要另觅他径”。

1.2.6 人工智能时期

关于累积和的争论并未阻止文体统计学发展的脚步。二十一世纪以来，随着计算机技术与人工智能的蓬勃发展，文体统计学研究也呈现出一派新景象。“模式识别”（pattern recognition）是文体统计学的核心问题之一。但在辨别有争议的作者身份过程中，辨别模式往往难以确定。这时，具有潜在数据组织识别能力的神经网络就有了用武之地。Tweedie等学者在1996年发表的论文中介绍了神经网络。神经网络以神经生理学为基础，由“连接”（connection）和“节点”（node）组成。在使用神经网络对未知文本进行分类之前，需要完成对神经网络的模式训练，以使其具备将两位候选作家区别开来的能力。在训练过程中，输入材料的选择显得尤为重要，通常可能包括单个功能词频次或单词出现的组合比。

Matthews & Merriam（1993，1994）的两篇论文堪称将神经网络用于文体统计学研究的开山之作。第一篇文章研究了莎士比亚与弗莱彻之争，重点关注了与两位作家相关的戏剧。在研究中，他们首先以两组辨识词输入材料，训练神经网络系统从两位作家的经典作品中辨别出作家身份。然后将经过训练的神经网络用于存疑作品《两个贵亲戚》（*The Two Noble Kinsmen*）的作者身份辨别。神经网络提供的量化数据显示《两个贵亲戚》为两位作家合作完成，此研究结果与之前由传统方法得出的研究结论完全一致。第二篇文章以莎士比亚与马洛之争为研究对象。受训后的神经网络具备了辨认两位剧作家作品的能力，之后被用来判定匿名剧本《爱德华三

世》(*Edward III*)的作者身份。研究结果表明《爱德华三世》系莎士比亚所作，但其创作可能深受马洛影响(该案例详见本书7.3.1节)。

1.3 语料库文体统计学发展简史

随着计算机技术的发展，语料库越来越多地被应用到文体统计研究中。事实上，语料库为包括文体学在内的语言学分支和其他人文社会科学提供了重要的研究基础和实证方法。本节将简述语料库与文体统计学的联系及语料库文体统计学的发展。

1.3.1 语料库与文体统计学

语料库(corpus或corpora)是指假定能代表一门语言、方言或某种语言现象的真实电子文本集合(Francis 1992)。建立语料库需要遵循严格的标准和规范，需要考虑语料库的规模、体裁平衡性、取样随机性、标注的层次和精度等问题，以保证语料库的代表性。在语料库基础上发展起来的语料库语言学则是指通过对大量真实语料进行系统分析发现语言规律的一种语言学理论，也称为语料库驱动的(corpus-driven)理论，或指一种基于语料库的(corpus-based)实证语言研究方法。语料库语言学家明确反对完全依赖语言学家个人语感的内省式研究，主张“系统地对大量的文本语料进行审视，从而发现一些以前从未有机会发现的语言事实”(Sinclair 1991)。

语料库语言学与文体统计学既有密切的联系，也存在显著的区别。两者的联系主要体现于：首先，二者都以真实的自然语料为研究对象；其次，都注重对语料语言特点的描写与分析；然后，都依赖系统而严格的统计学方法；最后，都强调研究方法的客观性。因此，语料库语言学与文体统计学在研究对象和方法上的相通之处为二者的结合奠定了基础。

同时，文体统计学与语料库语言学在语料使用和统计分析两个方面也存在明显的区别。首先，在语料使用方面，文体统计学主要用于分析某一特定类型文体的具体特征或某一作家的写作风格，通常使用较少的语料；而语料库语言学则主要利用语料统计数据研究语言总体趋势，验证特定的语言学假设或建立语言学理论，通常使用较大规模的语料库。其次，在统计分析方面，文体统计学在对具体特征的统计分析方面比语料库语言学更为成熟；而语料库语言学可以弥补文体统计学在语料规模和语言特征识别两个方面的局限。实际上，二者结合的结果——基于语料库的文体统计学已经有不短的历史，在研究的广度和深度、方法的应用和创新等方面都取得了长足的进步。

1.3.2 语料库文体统计学的发展

语料库在文体统计学中的应用可以追溯到20世纪40年代。1949年，意大利耶稣会牧师罗伯托·布萨（Roberto Busa）为圣托马斯·阿奎那（St. Thomas Aquinas）等人的作品编制词汇检索目录，共需要录入约1,100万中世纪拉丁词。为了提高工作效率，布萨开始探索计算机在这项工作中的应用，最终采用了打孔词汇卡和计算机检索程序结合的方式（参见Hockey 2004）。布萨的研究涉及人文计算领域的一个重要方面，即基于计算机的数据管理问题，揭开了语料库在文体研究领域应用的序幕（Mahlberg 2014）。不过，限于当时的计算机技术，语料库在文体分析中的应用仅仅是昙花一现。真正的语料库文体统计学研究必须要等到大量具有代表性的语料库出现后才能得以发展。语料库介入后文体统计学的发展也大致经历了两个阶段，即20世纪60至90年代的初步应用时期和21世纪以来的迅速发展时期。

1.3.2.1 初步应用时期

这一时期的研究主要集中在现代语料库建立、统计学方法探索及初步应用三个方面。毫无疑问，语料库的建立是对大规模语料进行文体统计分

析的前提和基础。20世纪60年代以前，第一代语料库的建设和研究主要采用人工手段，这使语料库的规模和代表性有很大的局限，加之研究者对语言现象的观察与分析存在很大的个体差异，这影响了对语言现象描述和分析的精确性。

随着计算机技术的飞速发展，第二代大型计算机语料库相继建成，为现代意义上的计算机语料库文体统计分析奠定了基础。1964年，弗朗西斯（Nelson Francis）和库赛拉（Henry Kucera）在美国布朗大学建成了“布朗语料库”（Brown Corpus）。布朗语料库是世界上第一个计算机机读平行语料库，通过随机采样的方法收入了15种文体的文本，容量为100万词左右。继布朗语料库之后，又陆续建成了一批大型语料库，这些语料库都遵循了布朗语料库的架构（采用相同的文本取样方法和相同的规模），与布朗语料库结构相似但时代不同，它们与布朗语料库一起构成了一个巨大的历时语料库，称为“布朗家族语料库”。自20世纪60年代以来，布朗语料库家族成员一直在增加，其中包括“Frown语料库”“LOB语料库”“F-LOB语料库”“LCMC语料库”和“COTE语料库”等（参见本书2.2.2节）。

目前最具影响力的大型平行语料库当数“英国国家语料库”（The British National Corpus，简称BNC）和“当代美国英语语料库”（The Corpus of Contemporary American English，简称COCA）。BNC由牛津大学出版社在20世纪80至90年代之间陆续建设完成，容量约为10亿词，覆盖口语、小说、杂志、报纸、学术等不同体裁的文本，是研究英国英语的权威语料库。COCA则是目前容量最大且完全免费的美国英语平行语料库。该语料库自1990年以来以每年收词2,000万的速度迅速发展，目前其总文本数已经超过5.6亿词，文本类型覆盖口语、小说、流行杂志、报刊、学术文本等体裁。COCA语料库可能是迄今为止使用最为广泛的英语语料库之一。

对文体统计分析方法的探索

各种大规模计算机语料库的建立为统计分析方法在文体学和其他学科中的应用提供了数据基础和广阔的应用空间。从20世纪80年代起，语料库文体统计分析方法的探索就在不断进行，其中较有代表性的学者有英国著名语言学家杰弗里·里奇（Jeffrey Leech）、文体学家Mick Short和Elena Semino。

1981年Leech & Short的《小说中的文体》（*Style in Fiction*）一书可以说是语料库文体学的奠基之作。两人在书中首次提出了“言语表征”和“思想表征”（speech and thought presentation）模式，并将其用于文体分析。他们在广泛搜集文学语料的基础上，依据语篇中叙述者干预程度的不同，对言语表征和思想表征进行了描述，验证了这种分析模式对小说文体分析的适用性。Leech & Short在书中列举了大量文体标记，供文体研究者根据需要从中选择使用，帮助研究者用更精确的数据来证明文体差异的存在。

Semino与Short继续发展了Leech & Short的分析模式，提出了第三种表征形式——“书写表征”（writing presentation），并尝试将这种分析模式应用于小说以外的其他文体。2004年，Semino和Short合作出版了《语料库文体学》（*Corpus Stylistics*）一书。该书的研究基础是一个约26万词的语料库，囊括了叙事、新闻和传记三种文体，每一种文体又被细分为严肃类文体和通俗类文体。由此，文体学研究便不再局限于文学文本。两人对语料库中出现的言语、思想和书写三种表征形式进行了人工分类和手工标注，并在此基础上对三种表征模式在叙事、新闻和传记三种文体中的出现频率、分布状态、所具功能等方面作了定量和定性分析。量化分析结果显示，各种表征模式在整个语料库中的出现频率以言语表征最高，思想表征次之，书写表征最低。此外，两人还专门讨论了三种表征中出现的特殊现象，包括：引用、内嵌式表达、歧义等。他们还选取了语料库中的两个文本做了个案分析研究，再次证明通过语料库标注和计算手段对长篇幅

文本进行文体分析可在某种程度上超越人的直感，得到其他研究方法很难或无法达到的研究结果（参见卢卫中、夏云 2010）。《语料库文体学》是文体分析与语料库语言学结合在研究方法上取得的最重要成果之一。

早期语料库文体统计学研究

20世纪60至90年代，采用语料库文体统计学方法的研究主要集中在作家风格量化研究和个别文本风格量化研究两个方面。

作家风格量化研究基于一个普遍认可的假设，即一个作家总有某些稳定的、自己所独有的语言特征，这些特征如同指纹专家对比的不同指纹一样，是作家在文本中留下的“写作指纹”，而作家的写作风格往往体现在这些可识别的写作指纹之中。因此作家风格的研究过程从某种意义上来说就是对其写作指纹的识别过程。此类研究常常被用于解决匿名作品的著作权归属问题。历史上某些时期的某些作品由于缺乏外部证据而难以确定其作者，这时其作者身份的确定往往就需要从文本内部寻找有力证据。

因此，如何有效识别作家在作品中留下的写作指纹就成了研究者必须面对和解决的问题。在早期的语料库文体风格研究中，为了识别写作指纹，研究者往往考察语料库文本的词频、搭配、句子长度、语法现象等语言特征。

有关语料库词频的研究，最著名的当属Milic（1967）对斯威夫特（Jonathan Swift）作品风格的考察。Milic的《乔纳森·斯威夫特风格的量化研究》（*A Quantitative Approach to the Style of Jonathan Swift*）一书被认为是该领域的先驱之作。Milic比较了斯威夫特作品中连接词的使用频次和同时代作家Addison、Johnson和Macaulay作品中连接词的使用频次。研究发现，斯威夫特作品中连接词的使用频率最高，这可能说明其写作的逻辑性较强。Milic还详细对比了斯威夫特作品中的语法结构与《给一位年轻诗人的忠告信》中的语法结构，判定该作品出自斯威夫特之手。澳大利亚学者John Burrows的研究也将关注点放在了语料库的词频上。

Burrows（1987a）使用主成分分析法分析了简·奥斯丁作品语料库中的30个常用词，其研究结果显示，词频的差异能够辨别小说中人物的差异。

搭配也是作家风格研究的参数之一。当具有相同语义特点的词项与节点词高频共现形成搭配时，往往产生一种持续的语义氛围，引起读者正面或负面的联想，这就是所谓的“语义韵”（semantic prosody）。Louw（1993，1997）认为语义韵这一概念可以帮助解释词汇偏离常规的创造性用法，为文体分析提供一个有效的研究工具。某一词汇在特定文本中的用法可以与语料库中该词的语义韵结合起来研究，Louw（1993）称这一方法为“对照语料库研究文本”。Louw以Larkin的诗作“Days”为例进行说明。Louw特别关注了诗作中days are这一搭配的语义韵。研究发现该短语往往与gone、over、past等高频词共现，逐渐产生了忧郁否定的含义，引起人们负面的联想。Louw（1997）认为词汇存在语义韵的现象常常不是人的直觉所能洞悉的，往往需要借助计算机语料库手段方能揭示。研究者可以使用语料库进行搭配检索来获取词汇的语义韵，从而验证文学鉴赏中的直觉感受。例如，Hori（2004）以460万字（共23个文本）的狄更斯作品语料库为基础研究了狄更斯作品中的搭配模式。他以220万字的“19世纪小说语料库”为参照语料库，详细探讨了狄更斯作品中特有词汇搭配模式及语义韵现象。Hori的研究揭示了狄更斯小说中语言搭配的创造性使用情况及其取得的文体效果，从而再次证明了搭配研究对于文体分析的重要意义。同样关注搭配研究的还有Hardy。Hardy的研究特别关注了Flannery O’Connor作品中单词eye的搭配情况。其研究表明词语搭配研究比单个词汇研究更易于发现文体模式。

句子长度也是学者们研究作家风格的维度之一。Milic（1991）的另外一项研究统计了威廉·福克纳（William Faulkner）和格特鲁德·斯泰因（Gertrude Stein）两位作家作品语料库中的句子长度。研究发现，斯泰因作品的平均句长为30个词，而福克纳作品的平均句长是38个词，由此可以看出两人的风格差异。此外，学者Mannion & Dixon（2004）也从句子长度

角度来研究作家风格。研究中，他们统计了奥利弗·戈德史密斯（Oliver Goldsmith）16篇文章的平均句长，并以此为依据鉴定10篇疑似文章。

作家风格有时也通过某些语法现象体现出来。Mah（1991）用语料库方法统计了法国小说《包法利夫人》（*Madame Bovary*）中的某些语法特征。研究表明，读者在阅读该小说过程中产生厌烦情绪的原因并非在于作者使用了乏味无聊的词汇，而在于作者使用了大量的未完成体动词。文本中未完成体的使用会让读者有时间延长之感，从而导致读者厌烦无聊情绪的产生。

对特定文体的风格研究也是文体统计学研究的焦点之一。语料库方法的介入意味着此类研究可以在海量文本数据分析的基础上进行，从而极大地提升了研究结论的客观性和可靠性。Cluett（1976）以Milic的研究为基础，利用计算机语料库研究了散文文体的总体特征。Gijsel & Vogel（2003）通过对语料库中一元字母单位和二元字母单位的统计，研究了右翼党派的政论文体特点及其发展变化。Argamon *et al.*（2005）三位学者通过对语料库中评论性文章（取自12本杂志，涉及6个学科）里的546个功能词汇的研究，揭示了实验科学文体与历史科学文体的不同特点，即实验科学类文章倾向于可能性的判断，而历史科学类文章则倾向于必然性的结论。Heylighen & Dewaele（1999）研究了语料库中正式文体与非正式文体里各类词汇的使用情况。他们采用F记分法统计了语料库中名词、形容词、冠词和介词的使用频次，研究结果表明正式文体中名词和形容词的使用频次高，而非正式文体中代词、副词、动词和感叹词的使用频次高。瑞典计算语言学家Karlgren & Cutting（1994）致力于英语文体自动分类研究。该项研究的关键在于文本分类模型的建立。在模型建立过程中，Karlgren主要依靠布朗语料库并使用了判别分析的方法。

1.3.2.2 快速发展时期

2006年是语料库文体统计学发展具有里程碑意义的一年。这一年，

牛津大学举办了以“文学语言语料库研究范式”为主题的诗学与语言学协会（PALA）论坛。论坛期间，英国伯明翰大学的Michaela Mahlberg做了一场题为“语料库文体学方法论、理论及模式”的主题报告，该报告被与会学者视为“语料库文体学正式得名的开山之作”。同年，劳德里奇出版社正式出版了美国纽约州立大学的David Hoover教授的专著《语料库文体学研究方法》（*Approaches to Corpus Stylistics*）一书。次年，著名学者Leech和Short正式提出了文体学研究语料库转向的观点。自此语料库正式成为文体研究的重要研究方法。与此同时，语料库文体统计学研究也进入了快速发展的黄金时期。

这一时期的研究一方面承袭了第一阶段的研究传统，另一方面也有突破性的尝试。与第一阶段研究相比，这一时期的研究呈现出以下四个特征：

第一，语料库方法论研究进一步深入。2004年Semino与Short的专著《语料库文体学》出版发行后，“如何以计算的方法研究言语与思想表达模式”一时成为研究的热点，学者们纷纷以此为题撰文立说。2006年《语言与文学》（*Language and Literature*）专刊收录了这一时期的研究成果。Busse（2010）以Semino & Short（2004）提出的语言、思想与书写表征模式为基础，对一个19世纪的小说语料库进行了研究。Busse认为严格的语料库标注有赖于明确的分类定义，而文本中的重复模式或许可以运用到语料库标注的自动计算中。

第二，语料库研究参数更加多元化。以往的研究主要集中在词频、搭配、句子长度、语法现象等方面，而这时期对关键词（key words）、关键语义域（key semantic domains）等方面的研究逐渐增多，使研究维度更加多元化。

在语料库语言学中，“关键词”是指与某一标准相比，频率明显偏高或偏低的词。Toolan（2009）以短篇故事为研究对象，考察了语料库中的短篇故事是如何通过关键词方法进行缩写的。研究发现故事文本中关键词以及关键词出现的句子与故事本身相关性较高，因此其衔接性和连贯性明显优

于文本的其他部分。Toolan的研究证明文本中关键词出现的句子是文本情节推进的标志性因素之一。Mahlberg & Smith（2010）认为语料库文体学与文学和语言研究紧密相联，故关键词分类可以从文学批评中获得指导。在此理念指导下，Mahlberg与Smith研究了奥斯丁小说《傲慢与偏见》（*Pride and Prejudice*）中关键词与主题之间的联系。次年，Mahlberg与McIntyre合作研究了英国作家Ian Fleming首部小说《皇家赌场》（*Casino Royale*）中的关键词。研究将关键词分为两类，即小说世界关键词和主题标志词。其中，小说世界关键词发挥着建构小说世界的作用，比主题标志词更为具体，通常用来表示小说中人物、具体事物或具体地点。主题标志词常常具有评价和隐喻意义，为研究者提供了广阔的阐释空间，因此可被视为“以读者为中心”的关键词。

有的研究将关键词分析与关键语义域研究有机结合起来。其中，关键语义域的识别可以通过一个叫做WMatrix的在线语料库分析工具实现（参见本书2.2.3节），该工具会自动为文本中的每一个词进行语义标注，并与参照语料库对比列出观察语料库中语义域频率相对较高的词汇。Culpeper（2009）将关键语义域信息与关键词分析结合起来对莎士比亚作品《罗密欧与朱丽叶》（*Romeo and Juliet*）中的人物话语进行了比较研究。此外，与关键词以及关键语义域相关的研究还包括Archer & Bousfield（2010）对《李尔王》（*The King Lear*）中人物话语的研究、McIntyre（2010）对电影《落水狗》（*Reservoir Dogs*）中对话及人物的研究等。

第三，语料库文本体裁更加丰富。第一阶段的研究大多集中于文学文本，尤其以小说文本为主，其他体裁文本相对较少。与第一阶段相比，第二阶的文本体裁可谓丰富多样。例如，2009年学者Swann和Allington就以读书小组的讨论文本作为研究对象。研究者使用一种名为Atlas.ti的数据分析软件，对30万字文本语料进行主题分析，并将文本分段进行标注。主题分析结果显示，涉及主观答复时，参与者往往会给出比较温和的评价。此项研究为后续研究者打开了一扇门，向他们展现了怎样在文

学文本之外进行文体分析研究。事实上，Swann和Allington认为研究者在研究对象的选择上应该享有更多的自由，可以包括广告、电影和演讲等文学文本之外的不同形式。

第四，语料库研究规模更加宏大。计算机语料库的迅猛发展使大规模文体研究成为可能。研究对象也从某位或某几位作者的风格研究扩展到特定时期特定主题下作家群体的整体风格研究。例如，2012年Hughes等学者利用Gutenberg数字图书馆语料库进行了第一次大规模的文体统计研究。研究显示，同一时代同一文学主题下不同作家的风格具有一致性（stylistic coherence）。该项研究以定量的方法阐释了“时代文体风格”（a literary style of a time）这一概念。

最后，简要介绍一下语料库文体统计学在国内的研究情况。受国外研究的启发，国内学者也开始关注语料库文体统计学的研究方法，并尝试将其应用到自己的研究中。据不完全统计，截至2018年6月，中国知网上以“语料库、文体、统计”为主题的期刊论文共计80余篇。这些论文中，绝大多数是对西方学界语料库文体统计研究成果的介绍和评论。如：隋桂岚等（2003）介绍了语料库、语言研究中的统计学以及文体分析的主要内容，并讨论了三者之间的关系；李涛、王菊丽（2009）梳理了语料库语言学与文学文体学交叉融合的历史，并介绍了相关的分析方法和分析软件；卢卫中、夏云（2010）撰文介绍了语料库文体研究的主要研究领域及研究成果，并在此基础上探索了该领域研究的不足之处及其成因。

与综述介绍类成果相比，原创性研究则显得比较薄弱。李德超、唐芳（2015）对英语旅游文本文体特征的研究可谓原创研究中比较有代表性的一项。该研究基于自建双语旅游语料库，比较了翻译旅游英语与原创旅游英语在类符/形符比、词汇密度、平均句长、平均词长等参数上的区别，得出的结论是：翻译旅游英语符合“简化”和“显化”趋势，时态更多样，文体更正式。

此外，值得一提的是：西方学界语料库文体统计的方法已经被广泛用

于考证佚名作品的作者身份，并取得了令人信服的研究成果。然而，遗憾的是，此类研究在国内似乎并不多见。李贤平（1987）曾以《红楼梦》中的47个文言功能词为识别特征，采用主成分分析、相关分析、聚类分析等方法对小说进行了统计分析，并依据统计结果得出了《红楼梦》为多位作者所作的研究结论。陈大康（1988a）发表论文《〈红楼梦〉"成书新说"难以成立——与李贤平同志商榷》，对李贤平的结论提出了质疑。施建军（2016）认为，自李、陈之辩以后，国内作者身份识别方面的研究似乎陷入了停顿，鲜有研究成果发表。事实上，中国古典文学作品作者存疑的情况比比皆是，这方面的研究可谓前景广阔。

推荐阅读

Fialho, O. & S. Zyngier. 2014. Quantitative methodological approaches to stylistics. In M. Burke (ed.). *The Routledge Handbook of Stylistics*. Oxon/New York: Routledge. 329-345.

Holmes, D. I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing 13* (3): 111-117.

Kenny, A. 1982. *The Compution of Style: An Introduction to Statistics for Students of Literature and Humanities*. Oxford/New York: Pergamon Press.

Mahlberg, M. 2014. Corpus stylistics. In M. Burke (ed.). *The Routledge Handbook of Stylistics*. Oxon/New York: Routledge. 378-392.

Semino, E. & M. H. Short. 2004. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. New York: Routledge.