arXiv:2304.00363v1 [cs.CL] 1 Apr 2023

蒂尔索·德·莫利纳作品中的自动作者归属

Miguel Cavadas Docampo[1] 和 Pablo Gamallo[2]

[1] 圣地亚哥-德孔波斯特拉大学，miguel.cavadas@rai.usc.es[2] 信息技术研究中心（CiTIUS）圣地亚哥-德孔波斯特拉大学，pablo.gamallo@usc.gal

摘要。 自动作者归属（AAA）是将数字人文领域的工具和技术应用于作者归属研究的成果。通过定量统计方法，该学科能够进一步探讨传统批评家数百年来一直研究的著名作者身份问题，为风格比较打开新的途径。本文旨在通过检验传统上归属于西班牙剧作家蒂尔索·德·莫利纳（1579–1648）的五部喜剧作品的作者身份，证明这些工具和技术的潜力：《*La ninfa del cielo*》、《塞维利亚的诱惑者》、《*Tan largo me lo fiáis*》、《被迫的女人》 和《因多疑而受罚者》。为实现这一目的，研究采用 R 语言中的 Stylo软件包进行聚类分析实验，并使用四种距离度量，对由蒂尔索、安德烈斯·德·克拉拉蒙特（约1560–1626）、安东尼奥·米拉·德·阿梅斯夸（1577–1644）和路易斯·贝莱斯·德·格瓦拉（1579–1644）的戏剧作品构成的语料库进行分析。所得结果表明，除《被迫的女人》外，所有归属于蒂尔索的作者归属判定均应被否定。

1 引言

作者身份问题贯穿文学史。无论是由于法律禁令、不当挪用，还是多位作者合作创作，世界文学经典作品中都有相当一部分作品的作者身份存疑。数字人文吸收了自然语言处理的方法与文学研究的理论，逐渐成为处理作者身份问题的有效框架，并催生了这一学科：非–传统作者归属研究，亦称自动作者归属（以下简称 AAA）。这一学科并非否定传统研究，而是对其加以补充，通过识别可量化的作者标记来实现这一点（Almeida，2014年，第10页）。AAA 基于以下假设：“通过比较文本中的文本测量指标与各位作者写作样本中相应的指标值，可以从一组可能的作者中选出文本的作者”（Grieve，2005年，第1页）。人们对 AAA 日益浓厚的兴趣催生了一项国际研讨会，该研讨会由一系列相关的共享任务组成，自 2011 年起作为 PAN实验室“揭示剽窃、作者身份与社会

软件滥用”项目的一部分举办（Argamon & Juola，2011年）。PAN 共享任务的主要目标是提出一个标准化的作者归属判定评估框架，以便比较参与该活动的各项策略。

本文的首要目标是强调，计算语言学领域的某些工具在作者身份研究中可能发挥的作用，并由此指出 AAA 作为学科间富有成效的交汇点所具有的价值。为通过实例阐明这一点，我们将具体聚焦于蒂尔索争论，旨在得出能够作为重要论据的结论，以强化一些最受支持的批评立场。更具体地说，我们旨在找出黄金时代五部戏剧作品最有可能的作者；这些作品传统上都被归于蒂尔索·德·莫利纳名下，其中包括著名剧作《塞维利亚的诱惑者与石像客人》（《塞维利亚的诱惑者与石像客人》），该剧正是标志性情人唐璜神话的来源。尤其在过去几十年间，这五部戏剧作品的所谓作者身份一直引发巨大争议。为开展此类实验，我们配置并应用了不同的 AAA 策略，研究语料库由十二部归于四位不同作者名下的戏剧作品组成，其中包括蒂尔索·德·莫利纳的作品，以及五部作者身份存在疑问的戏剧作品。

我们工作的主要贡献，概括而言，是将蒂尔索争论引向作者归属判定领域，并强化否定蒂尔索在若干传统上归于其名下的戏剧作品中作者身份的批评立场，进而将这些作品归还给其原作者。研究结果迫使我们强调安德烈斯·德·克拉拉蒙特这一人物，并指出他最有可能是蒂尔索著名戏剧作品《塞维利亚的诱惑者》的作者。这引发了两场争论：一场关于蒂尔索·德·莫利纳在西班牙黄金时代文学经典中的地位，另一场关于过去如何开展作者归属研究。鉴于我们所采用的方法，将非–文体计量计算策略应用于作者归属判定领域，同样是一项引人注目的贡献，因为我们取得了极具前景的结果。

本文结构如下：第2节简要比较过去开展作者归属研究的方式与数字人文正在完善的新技术。第3节详细介绍了我们为使用蒂尔索语料库设计实验而采用的手段和方法；我们在第4节对这些实验进行解释、执行并讨论。因此，第5节不仅呈现了有关蒂尔索·德·莫利纳戏剧作品作者身份的重要事实，也提出了未来研究应遵循的指导方针，以应对这场富有成果且令人振奋的争论。

2 背景与相关研究

涉足作者归属研究这一崎岖领域的研究者，已经开发并改进了大量方法，这些方法基于对各种文体标记的统计比较，据称能够揭示每位作者的个人语言变体。

这些标记包括度量指标、词长或句长、标点符号、词汇指数、错误、词类、搭配及其他形态句法模式。然而，对这些要素的分析通常未能得出明确结论，暴露出其方法论上的弱点。解决这一问题的主要障碍可能在于，许多情况下，所用方法的区分度不足以对归属判定作出可靠判断。正是在这一点上，AAA或许能够发挥作用，并凭借文本数学化实现更强的判定能力。AAA旨在与传统研究形成的主要区别是："只有定量技术才能接受实证评估并以机械方式应用"（Grieve，2005）；因此，其目标不是取代传统研究，而是以数值方法对其加以补充。

AAA是一门发展了约五十年的学科。即便如此，它仍不乏批评、争议和持续存在的问题，其中 Rudman 提到了："研究受权宜之计支配；缺乏合格的研究；统计技术存在缺陷；原始数据遭到篡改；缺乏相关领域的专业知识；研究方法业余；对错误的处理不充分"（Rudman，1998年，第351页）。这些弱点通常源于非传统作者归属研究（NTAAS）所需采用的科学视角，这与其他人文研究不同。Rudman 的以下论述清楚地说明了这一点："每一项非–传统作者归属研究都是一项实验，一项'科学'实验"（Rudman，2016年，第310页），因此必须严格遵守两项原则：可重复性和实验方案。Rudman 针对这一缺陷提出的解决方案是："构建正确而完整的实验设计；培训从业者；全面研究文体；识别并培训把关人；建立完整的理论框架；组建从业者协会"（Rudman，1998年，第351页）。

迄今为止，大多数AAA研究仅以英语为对象开展，其中大量研究旨在确定莎士比亚文本的作者身份（Brian Vickers于2011年评述了其中一些研究）。由于缺乏以可接受格式提供的文学文本，极少有论文聚焦于古西班牙文学；这迫使每位研究者都得负责从不同的HTML页面中获取质量合格的文本（Calvo Tello，2016）。尽管存在这些障碍，西班牙文学史上两个最著名的作者身份谜题已经在这一方法下得到探讨：Javier de la Rosa和Juan Luis Suárez（2016）获得的结果表明，法学家Juan Arce de Otálora很可能是《小癞子》的作者；而Javier Blasco（2016）则试图揭开Avellaneda的身份，并将《伪堂吉诃德》与Castillo Solórzano或蒂尔索·德·莫利纳等作家联系起来。最近，2017年，为研究西班牙黄金时代戏剧的文体计量学方面，EstilometríaTSO项目得以创建（Cuéllar González与García–Luengos，2017）。鉴于迄今该领域所开展的工作甚少，显然这项研究仍处于非常初步的阶段。

西班牙黄金时代的戏剧文本，由于大多经历了复杂的文本传承过程，在作者归属判定方面面临着额外的困难。蒂尔索·德·莫利纳的作品

（1579–1648）是 Fray Gabriel Téllez 为人所知的笔名，这正是这类悬而未决的难题的一个鲜明例子。许多评论家致力于研究蒂尔索作品中各种作者身份问题，但总体而言，研究结果远谈不上定论。

目前已定义了两种 AAA 方法：基于作者画像的方法和基于实例的方法–。基于作者画像的方法–不考虑同一作者所写文本之间的差异，因为对于每位作者，它都会将该作者可获得的所有文本串联起来，构建一个单一的大型文件。基于实例的方法–要求每位作者提供多个训练文本样本，以便建立准确的归属判定模型（Stamatos，2009）。Nektaria Potha 和 Efstathios Stamatos 关注 PAN–2013 作者身份验证竞赛的结果，就这两种方法中哪一种更有效这一问题得出结论：基于实例的方法–更适合这项任务。不过，他们承认，“当可获得的文档样本数量有限且篇幅较短时，基于作者画像的范式–更为有效”（Potha & Stamatos，2014年，第314页）。鉴于我们的语料库规模有限，我们采用基于作者画像的方法–；同时，由于基于实例的方法–取得了更好的成果，我们也采用该方法。

AAA 所采用的文体特征主要依赖功能词的使用频率。Eisen 等人提出了一种基于功能词邻接网络（WANs）的新技术，“以功能词为节点，边则包含有关两个功能词在一定距离范围内共同使用的信息”（Eisen 等人，2018年）。若要考察此类功能词邻接网络，可使用 R Stylo 的 Rolling Delta 和 Rolling Classify 功能；Hartmut Ilsemann 在其近期关于《*Parnassus Plays*》（Ilsemann，2018年）和托马斯·基德《*Cornelia*》（Ilsemann，2019年）作者身份的两项研究中也采用了这些功能。事实上，Ilsemann 熟练运用了该软件的强大功能，充分利用了其所有特性，包括多维尺度分析、自助法共识树和聚类分析（本文采用的方法）。他提醒说：“尽管 R Stylo 应用程序的各项功能会根据所选参数的数量提供大量结果，但总体评估可以视为作者身份的外部证据”（Ilsemann，2018年，第555页）。因此，他鼓励推广这些功能，以建立一种（从词源意义上说）可接受公众辩论的法证学科。

3 方法

AAA 依赖于不同的统计度量，这些度量会考虑功能词的分布。功能词不仅是词汇指标，还包含大量句法信息，因为它们参与构成许多句法结构。从文本中提取最高频词，将是对文档进行文体计量比较的第一步。在作者归属任务中，选择特征时最重要的标准是其频率。一般而言，某一特征出现得越频繁，它所反映的文体变化就越多（Stamatos，2009）。文本中最高频术语（功能词）的分布构成了作者的文体指纹。作者的文体指纹

并不是其使用最多的具体词语，而是其使用这些词语的频率。AAA 策略之间的主要差异在于：如何选择适当的度量，以基于词语分布（或作者的文体指纹）比较文档。最适当度量的选择取决于可用的语料库和研究目标。然而，根据 Grieve（2005）的观点，定量作者归属判定的最佳方法，是尽可能基于多种文本测量指标的数值。这是因为一般而言，配置中的微小变化可能会导致结果发生非常大的变化。因此，本文采用了五种不同的策略。这五种策略分为两种相互对立的方法：基于实例和基于文体概貌–。

3.1 基于实例的研究方法

单篇文本是这一流程的基本分析单位。按照这种方法，我们对单篇文本进行聚类分析，根据文体相似性将它们分组；如果调整得当，这就相当于按作者对文本进行分组。单篇文本分组后的结果以树状图呈现。聚类过程基于 Delta度量，该度量由 John Burrows 于 2002 年专为文体计量学创立。此方法由 Stylo 实现；Stylo 是一个灵活的 R语言软件包，用于对文本集合进行高层次文体分析（ Eder，2016）。Delta度量通过 z 分数对频率进行标准化，以减少高频词的影响。设 f(D) 为文档 D 中 n 元语法的频率， 为语料库中该 n 元语法的平均频率，为其标准差，则 z 分数定义如下：

$$z\left(f_i\left(D\right)\right) = \frac{\left(f_i\left(D\right) - \mu_i\right)}{\sigma_i} \tag{1}$$

同一作者所写的一组训练文本与一篇未知文本之间的差异，是两者z–分数的绝对差值的平均数（Stamatos，2009）。给定归一化文档向量，Burrows 的Delta就是使用z–分数计算归一化频率的曼哈顿距离。给定文档D_1 和D_2 ，，距离Delta Δ 的计算公式如下：

$$\Delta = \sum_{\mathrm{i}\,=\,1}^{n} \left|z\left(f_i\left(D_1\right)\right) - z\left(f_i\left(D_2\right)\right)\right| \tag{2}$$

Delta 值越低，所研究文本之间的相似度越高。近期，Maciej Eder（2016）对 Burrows 的经典方法进行了一些创新。在其实验中，José Calvo Tello（2016）证明，这些改动能够带来更好的结果。Stylo 允许用户选择 Burrows 和 Eder 的度量方法，甚至还可以选择基于其他相似度方法的度量方法，例如 Wuzburg距离。正如我们的测试所表明的，由于这些方法都能提供类似的结果，我们将仅采用 Burrows 和 Eder 的方法。

3.2 基于–特征的方法

在这种方法中，属于同一作者的已知文本会被合并为一个单独的文档（作者画像），然后计算该作者画像与未知文本画像之间的距离度量。研究者设计并实现了四种不同的距离度量：Kullback–Leibler 散度、困惑度、基于排名–的距离以及分布相似度。这些度量代表了四种不同的基于语料库–的文本比较策略。它们最初并非为 AAA 的目的而设计，但通常被用于其他计算任务，例如语言识别、语言距离、信息检索和数据挖掘。

Kullback–Leibler Kullback–Leibler散度用于比较两个分布，更准确地说，它衡量一个概率分布（例如，未见文档的特征）与第二个参考概率分布（作者的特征）之间的差异。在 Iriarte 等（2018）中，该方法用于衡量由不同社会群体的个体撰写的文本之间的距离：男性/女性、大学/非–大学人群等。给定一篇测试文本或未知文本（T）以及某位作者的已知文本（A），分布 T 和 A 的 Kullback–Leibler散度 KL 定义如下：

$$KL\left(A,T\right)=\sum A\left(ngr_i\right)\log\frac{A\left(ngr_i\right)}{T\left(ngr_i\right)} \tag{3}$$

公式 3 可计算分布 T 与分布 A 之间的差异，其中考虑了每个待比较文本语料库（T 或 A）中 n–元组（词或字符）的概率。该散度是一种非对称度量，通过计算两种互补比较的均值将其转换为对称度量（即距离）：X 相对于 Y 的散度，以及 Y 相对于 X 的散度。在我们的实验中，我们将 Kullback–Leibler散度应用于最高频词一元组的概率分布。

困惑度 困惑度常用于评估基于从文本语料库中提取的 n–元语法构建的语言模型的质量，也可用于衡量模型（例如作者的特征画像）与测试数据（即一篇未见文档的特征画像）的匹配程度。更正式地说，困惑度是输入测试文本的归一化逆概率。它可用于比较测试文本（T）与作者模型（M）。给定作者模型 M 时，T 的困惑度 PP 由下式定义：

$$PP\left(M,T\right)=2^{-\sum T(ngr_i)\log_2 M(ngr_i)} \tag{4}$$

其中 ngr_i 是 T 和 M 共有的一个 n–gram。公式 4 可用于设定测试集与作者模型之间的差异。T 在 M 条件下的

困惑度越低，两个比较对象之间的距离就越小。文本可以用词或字符的 n–gram 建模。在我们的实验中，我们对用 7–gram 字符建模的文本应用了 PP 距离。在其他研究中，PP 也用于比较 40 种欧洲语言之间的语言距离（Gamallo 等，2017），以及计算同一语言的历时变体之间的距离（Pichel 等，2019）。

排名–基础 两种语言之间基于排名的距离，是通过比较它们各自的排名列表得出的。它使用两个词语概要（作者的概要和未见文档的概要），并根据“离位”度量计算一个简单的排名次序统计量。该度量用于确定一个概要中的 n–元语法与其在另一个概要中的位置相差多远（Cavnar 和 Trenkle，1994）。这种度量常用于计算语言识别（Gamallo 等，2014）。更正式地说，给定测试文本（T）的排名列表 Rank_T 和某位作者文本（A）的排名列表 Rank_A，基于排名的距离 R 计算如下：

$$R\left(A,T\right)=\sum_{i=1}^{K}\left|Rank_{A}\left(ngr_{i}\right)-Rank_{T}\left(ngram_{i}\right)\right| \tag{5}$$

其中，K 表示出现频率最高的 n–元语法的数量，$\text{Rank}_\text{A}(\text{ngr}_\text{i})$ 是特定 n–元语法 ngr_i 在 A 中的排名，$\text{Rank}_\text{T}(\text{ngr}_\text{i})$ 是同一 n–元语法在 T 中的排名。在我们的实验中，我们将这种距离应用于词一元语法列表。

余弦相似度 作为第四种度量指标，我们采用向量空间模型。向量空间模型是文档词汇最常用的表示方法之一，用于计算文档之间的分布相似度。具体而言，该策略将从作者档案中提取的词向量与从未见文档中提取的词向量进行比较。由于这是一种相似度度量，其值越高，文本之间的相似度就越高，最大值为 1。由于它与其他三种距离度量的作用方向相反，因此我们通过用 1 减去相应值，将其转换为距离度量。在实验中，我们参考 Afzali 和 Kumar（2017）对不同度量指标（即余弦相似度、杰卡德和戴斯）所进行的比较研究，采用了余弦相似度，以评估这些指标在确定两个文本文档相似度方面的表现。余弦相似度明显优于其他度量指标。给定 A 和 T，余弦相似度（作为距离）定义如下：

$$Cosine\left(A,T\right)=1-\frac{\sum A\left(ngr_{i}\right)T\left(ngr_{i}\right)}{\sqrt{\sum A\left(ngr_{i}\right)^{2}}\sqrt{\sum T\left(ngr_{i}\right)^{2}}} \tag{6}$$

3.3 度量平均值

最后，将上述定义的四种度量指标通过取平均值进行合并。为了计算四种距离度量所得值的平均值，需要先对它们的最终得分进行归一化。作者已将这四种度量指标及其平均组合实现为一款使用 PERL 语言编写的可执行软件。该软件用于执行作者归属判定，并可免费获取。[3]

必须强调的是，除了上文定义的基于实例的策略之外，使用四种不同的基于特征–的策略，使我们能够对作者身份得出更加坚实可靠的结论。这些并不是单一策略的不同配置，而是涵盖同一问题不同方面的五种互补方法；在大多数基于 Stylo–的研究中，通常采用的是前一种做法。

4 实验

4.1 蒂尔索·德·莫利纳部分作品的作者身份问题

蒂尔索作品的编辑与传播，清楚地展现了黄金时代戏剧传承背后复杂的过程。喜剧作品在上演后，经过至少导演和出版商之手，以称为 卷 的合集（通常每卷收录十二部戏剧作品）或单独的小册子形式印刷出版，而诗人在这一过程的任何阶段都无法进行控制。因此，作品在传播过程中出现作者身份误判也就不足为奇了。在以蒂尔索之名出版的五卷作品中，评论家指出其中两卷，即于 1630 年代出版的第二卷和第三卷，可靠性存疑。关于第二卷，其中收录了十二部戏剧作品，蒂尔索本人揭露了出版商的欺诈手段：“在这十二部喜剧中，我以自己的名义献出四部属于我的作品，另外八部则以其真正作者的名义献出（我不知道他们究竟遭遇了什么不幸：这些作品明明是如此显赫的父亲所生，却被推到我的门下），剩下的作品则……”（Oteiza，2000 年，第 106 页）。[4] 遗憾的是，蒂尔索没有说明那八部并非由他创作的作品的实际作者身份。直到今天，评论家仍已能够厘清若干部戏剧作品的作者身份。当前关于第二卷作者归属判定的情况如表 1 所示。

关于作者身份，主要疑点集中在两部戏剧作品上：《*La mujer por la fuerza*》（《被迫的女人》）和《因多疑而受罚者》（《因不信任而被定罪者》）。关于后一部作品，人们提出了大量论据，否定其为Tirso所作，并将其归于其他作者。评论家之间的分歧极为激烈，甚至有人认为该剧是合作创作（Rodríguez López–Vázquez，2010年）。值得注意的是，最近，Alejandro

[3] https://github.com/gamallo/Autoria[4] “我将其余作品献给这八部戏剧作品的真正作者（他们并不知道，这些作品身为如此杰出父亲的女儿，却因不幸而来到我的门前）。在这十二部戏剧作品中，只有四部是以我的名义创作的，其余八部并非如此。”

喜剧	作者
La reina de los reyes	Hipólito de Vergara
Amor y celos hacen discretos	蒂尔索·德·莫利纳
Quién habló pagó	Rodrigo de Herrera
《真相总会相助》	此前：Luis Belmonte Bermúdez？Juan Ruiz de Alarcón？Rodrigo de Herrera？蒂尔索·德·莫利纳？
	自 García–Reidy（2019年）起：Lope de Vega
《特鲁埃尔的恋人》	胡安·佩雷斯·德·蒙塔尔班
《经由地窖与转盘》	蒂尔索·德·莫利纳
《以谨慎对谨慎》	米拉·德·阿梅斯夸
《被迫的女人》	蒂尔索·德·莫利纳？
《因多疑而受罚者》	米拉·德·阿梅斯夸？安德烈斯·德·克拉拉蒙特？贝莱斯·德格瓦拉？蒂尔索·德·莫利纳？合作创作？
《唐·阿尔瓦罗·德·卢纳第一部》	米拉·德·阿梅斯夸
《唐·阿尔瓦罗·德卢纳第二部》	米拉·德·阿梅斯夸
《这才叫谈判》	蒂尔索·德·莫利纳

表1。 Tirso《第二部 》中的喜剧及其对应作者或最有可能的人选（问号前）

García–Reidy（García–Reidy，2019年）似乎已证明《真相总会相助》（《真相总会相助》（真相总会相助））出自洛佩之手；这部作品传统上被认为是Tirso所作。他采用的方法包括使用R语言的Stylo软件，该软件也被用于开展我们的一些实验。

最重要的作者身份问题涉及以单行本小册子形式出版的作品；这正是《*La ninfa del cielo*》的传播媒介。阿尔弗雷多·罗德里格斯·洛佩斯–巴斯克斯为Cátedra编订的校勘版将该剧归还给路易斯·贝莱斯·德·格瓦拉（2008），迄今尚无人提出反驳。另一个重要问题是著名的《塞维利亚的诱惑者》（《塞维利亚的诱惑者》）及其名为《*Tan largo me lo fiáis*》的异本，两者有1433行诗句相同，占全文总量的60%以上（García Gómez，2005）。值得指出的是，罗德里格斯·洛佩斯–巴斯克斯（Rodríguez López–Vázquez，1990）认为，《*Tan largo*》（因此也包括《*El burlador*》）的作者是剧作家安德烈斯·德·克拉拉蒙特（c. 1560–1626）；尽管这个名字或许并不为人熟悉，他却是这场争论的核心人物。罗德里格斯·洛佩斯–巴斯克斯认为，这位作者的参考文献匮乏、在文学史中缺席，以及其作品中有很大一部分至今未刊行，原因都在于他一直被视为剽窃者。因此，解决《*El Burlador*》的作者身份问题，对于确定克拉拉蒙特在文学史中应占据的位置至关重要。

4.2 语料库

应用这些方法的理想语料库应包括所有作者身份存在争议的Tirso作品，以及所有可能作者的若干代表作。然而，数字化作品难以获取，极大地限制了语料库的创建，因为即便是最知名作者的作品，也严重缺乏数字化版本。能够

在网上免费查阅的黄金时代喜剧所占比例非常小。其中存储大多数此类作品的来源是米格尔·德·塞万提斯虚拟图书馆（BVC）网站，[5] 该网站提供了涵盖西班牙语文学完整历史的大量文学和批评文本。西班牙古典戏剧协会（AHCT）[6]的网站则更加专注于黄金时代喜剧，因此包含了这一时期众多作者的作品。需要指出的是，数字化不足这一缺陷，是另一个更为严重问题的直接后果：版本匮乏。这种稀缺现象在安德烈斯·德·克拉拉蒙特的作品中尤为严重；批评家很少关注他，这使得建立一个用于研究归属判定的广泛语料库变得极其困难。

综上所述，构建语料库以考察Tirso de Molina戏剧作品作者身份问题时，所采用的首要标准必然是文本的可获得性。表2列出了作者身份存疑的戏剧作品，其中我们能够将其转换为可供计算机处理的文本文件。

Text	可能的作者
《塞维利亚的诱惑者》	Claramonte / Tirso
《因多疑而受罚者》	Amescua / Claramonte / Guevara / Mira / Tirso / 合作创作
《被迫的女人》	Tirso / 其他作者（未提出具体姓名）
《*La ninfa del cielo*》	Guevara / Tirso
《*Tan largo me lo fiáis*》	Claramonte / Tirso

表 2。文本　本研究讨论的剧本及其对应的可能作者　　作者。

从表 2 可以看出，除 Tirso 外，参与本研究的作者还有安德烈斯·德·克拉拉蒙特、米拉·德·阿梅斯夸和路易斯·贝莱斯·德·格瓦拉。

黄金时代的喜剧彼此十分相似，因为它们遵循相同的风格和叙事模式。在这种情况下，鉴于作者之间的区分更加复杂，有必要汇编能够充分代表每位作者风格的文本。为此，我们遵循对文本的传统评价，从现有作品中选取最为人熟知的标题。在文本数量充足的情况下（Tirso 和 Mira），我们选择了创作日期尽可能接近作者身份存疑作品的戏剧作品，并力求遵循以下原则：“每位作者的所有文本都应创作于同一时期，以避免随时间推移而发生风格变化”（Stamatos，2009年，第558页）。BVC 被用作文本的主要来源，因此几乎所有文本都是依据相似标准编辑的。当某部喜剧无法在该网站获取时，我们便求助于 AHCT；若仍无法获取，则采用相关版本。最终，所构成的语料库如表3所示。

[5] http://www.cervantesvirtual.com/[6]
http://www.wordpress.comedias.org/

作者	喜剧	创作日期	来源	诗行数 诗行数
蒂尔索·德·莫利纳	《宫廷里的羞怯者》	1610–1625	BVC	3968
	《绿裤子唐希尔》	1615	BVC	3277
	言辞与羽毛	1627年之前	BVC	3979
				11224
米拉·德·阿梅斯夸	《以谨慎对谨慎》	1635年之前	BVC	2845
	魔鬼的奴隶	1612	BVC	3296
	希罗与勒安得洛斯	—	AHCT	3310
				9451
Andrés de Claramonte	《我不喝这水》	1617年之前	BVC	2743
	《佛兰德斯的勇敢黑人》	1638年之前	Bubok	2973
	《太阳落在我身上》	—	BVC	3206
				8922
贝莱斯·德·格瓦拉	《魔鬼在坎蒂利亚纳》	1622	BVC	2621
	维拉山脉	1613	BVC	3306
	侮辱也是毒药 （幕第一）	—	BVC	1196
				7123
未知	《塞维利亚的诱惑者》	1612–1617	BVC	2900
	《因多疑而受罚者》	1635年之前	BVC	2995
	《被迫的女人》	1635年以前	Clásicos Hispánicos	2886
	《*La ninfa del cielo*》	1610–1620	AHCT	3505
	《*Tan largo me lo fiáis*》	1612–1617	BVC	2760

表3. 语料库中的戏剧作品。尽可能注明创作日期。

文本获取并转换为纯文本后，便会进行预–处理。对文本所作的主要修改如下：采用 UTF–8编码，将 Windows 的换行符替换为 Linux 的换行符（后者不包含回车符），以及删除诗句编号。戏剧角色的姓名未被排除，因为姓名学一直是研究者在其提出的各种归属判定方案中所依据的论据之一。只有完成这一系列处理后，才有可能开始对文本进行实验。

4.3 实验

我们进行了两种不同类型的实验：戏剧聚类和测量戏剧作品之间的距离。

聚类分析 为正确配置 Stylo 工具（0.6.9版），我们使用精简版语料库（剔除作者不详的作品）进行了一系列初步测试。通过这种方式，我们检验了哪些参数值能得出最一致的结果。我们得出的结论是，最高频词参数的最合适取值为 250，即最大值和最小值均设为 250，以生成一棵树状图。这是实验的关键，因为该数值的变化可能会显著影响结果。我们决定从文本中提取的特征是词一元组，即词元，并认为没有必要进行任何词项筛除。随后，我们

将最佳配置应用于上文表3中列出的戏剧作品。实验得出的树状图如图1所示。

正如预期的那样，距离最小的是《塞维利亚的诱惑者》与《*Tan largo me lo fiáis*》，因为在很大程度上，它们是同一文本。毫无疑问，它们与 Claramonte 的作品归为一组。由于 Mira 的三部作者身份确定的喜剧之间距离较小，Mira 似乎是风格最为一致的作者，而 Guevara 的风格则最不稳定。Guevara–Tirso 和 Claramonte–Mira 这两组作者之间的文体相似性最为显著。至于作者不详的戏剧作品归属判定，所得结果十分明确：《因多疑而受罚者》与 Mira 的作品相关联，《被迫的女人》与 Tirso 的作品归为一组，而《*La ninfa del cielo*》则归入 Guevara 一组；最后这一项是所有归属判定中最为明确的。

如前所述，经 Eder 的实现方式修正后的新 Delta度量（2016）在一些实验中取得了更好的结果。为了与经典的 Burrows Delta 所得结果进行比较，适合采用 Eder 的方法开展另一项类似分析。这带来了图 2 聚类分析中展示的新结果。

与之前的树状图相比，变化虽小，却十分显著。Claramonte–Mira这一对仍然保持着高度的相似性，但Tirso与Guevara组成的这一对则不再如此；Guevara如今与另外三位作者的差异更大。主要差异体现在喜剧《因多疑而受罚者》上，它在风格上与《被迫的女人》非常接近，因此与Tirso的作品被归为一组。其他归属判定仍然保持着明确性，尤其是《La ninfa del cielo》，其归属判定甚至得到了进一步强化。

距离度量 我们实现了一个AAA免费软件工具，用于应用上一节定义的四种距离度量：困惑度、Kullback–Leibler散度、基于排名–的度量以及分布相似度。按作者和度量分别列出的结果见表4。

对于所研究的所有作品，困惑度都排除了Tirso的作者身份；不过，在其中大多数作品中，它都将Tirso列为非常接近第一选项的第二选项。在其他度量中，Tirso的排名往往不太有利。困惑度几乎在所有情况下都将Guevara列为距离最远的作者，其中还包括《La ninfa del cielo》这一颇为出人意料的情况，因为按预期该作品应归属于Guevara。

Kullback–Leibler 与 Stylo 的结论一致，唯独《La ninfa del cielo》例外：该剧被归于 Mira 名下，尽管这项度量将 Guevara 列为这部剧最有可能的第二作者。

基于排名的度量认为，Claramonte 最有可能是《El condenado》的作者，不过他与 Mira 和 Tirso 之间的距离较小，而这两位作者彼此之间的距离更小；这与困惑度在上述喜剧中显示出的两者之间显著差异形成对比。然而，排名与困惑度之间的主要差异在于《La mujer》

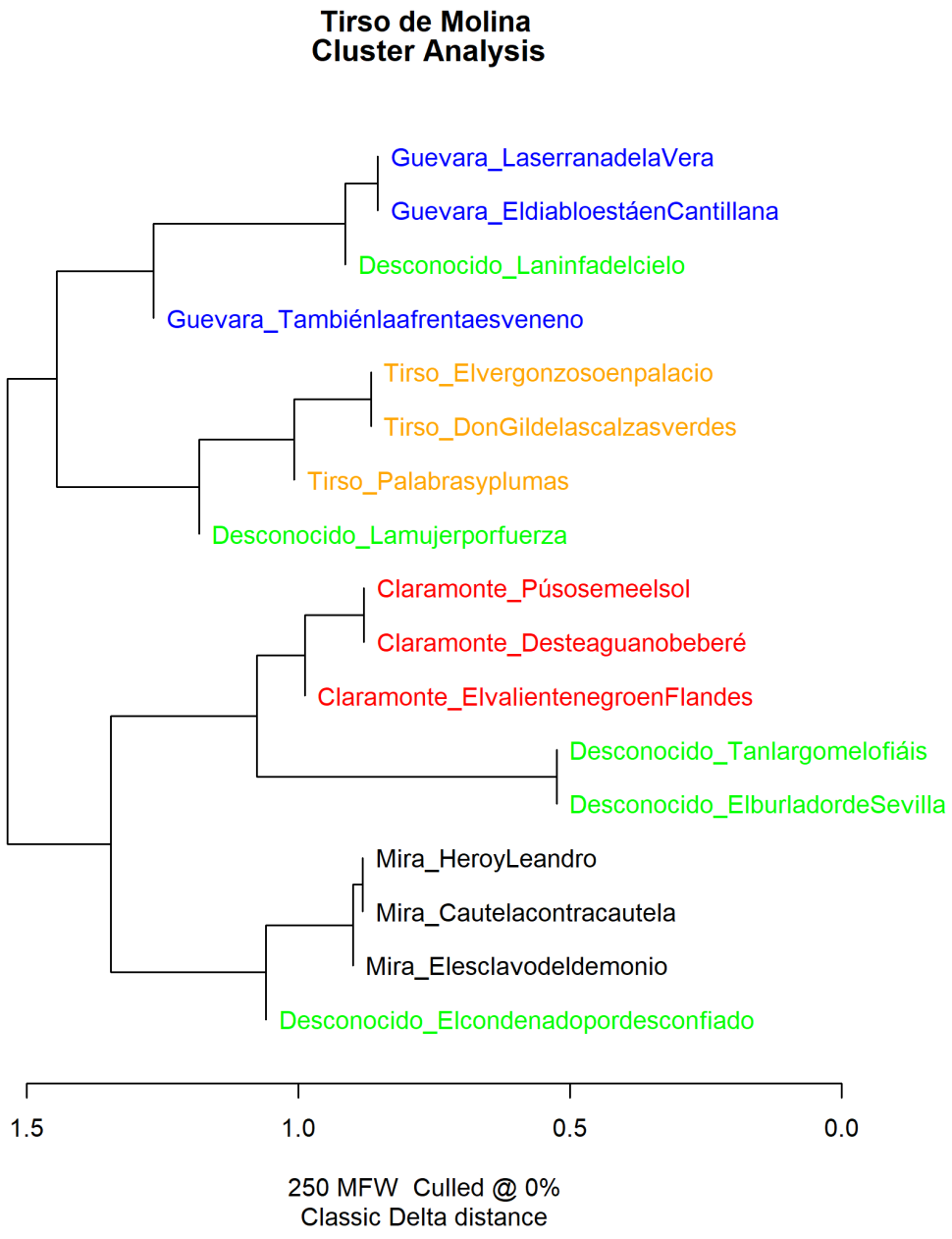

图 1. 使用布罗斯距离进行聚类分析所得的树状图。

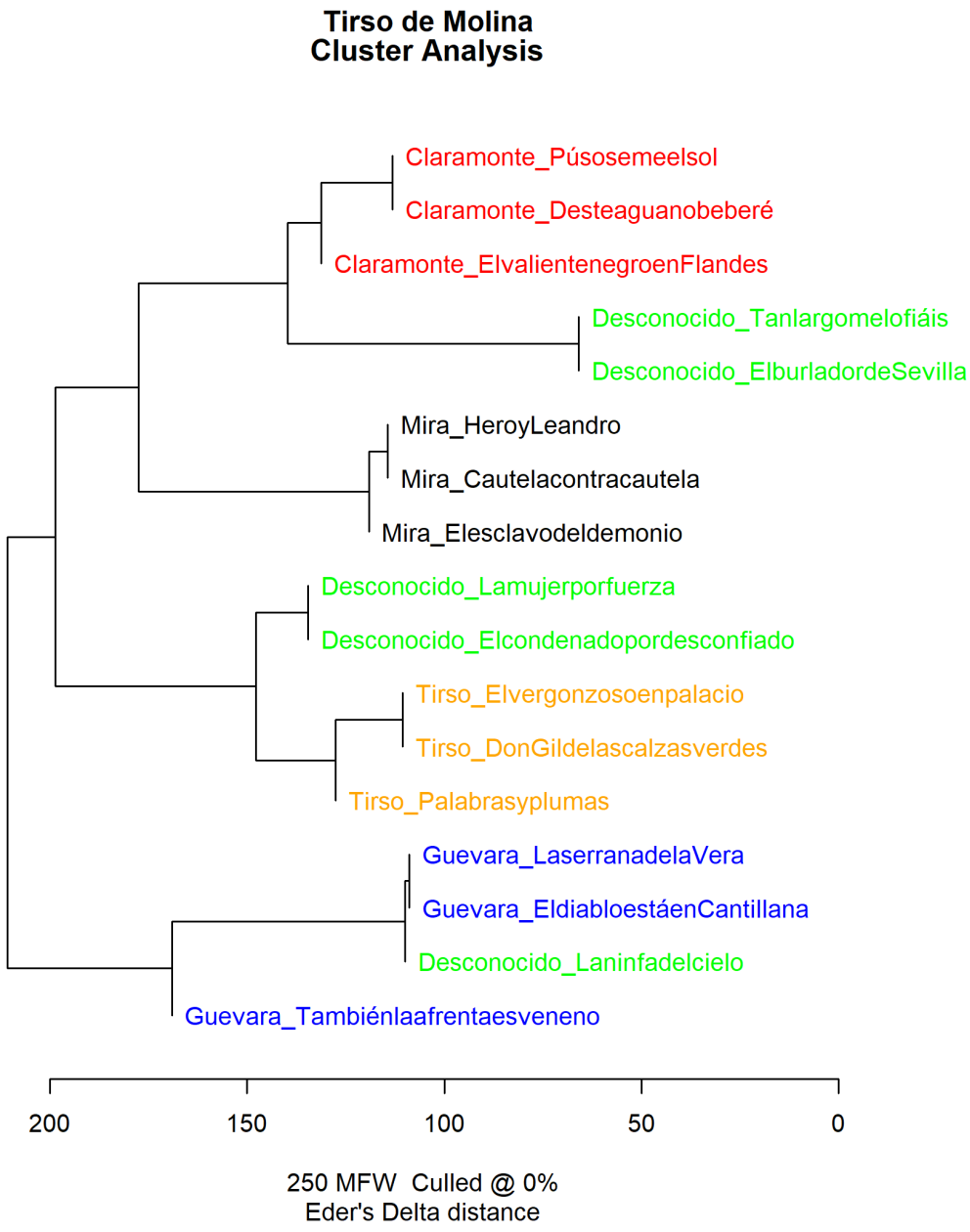

图2. 使用 Delta Eder距离进行聚类分析的树状图。

	《*La ninfa del cielo*》	***El burlador de*** 塞维利亚	《*Tan largo me lo fiáis*》	《被迫的女人》力量	《*El condenado*》*por* 不信任的
困惑度	0.0000 Mira	0.0000 Clar	0.0000 Clar	0.0000 Mira	0.0000 Clar
	0.0483 Tirso	0.1525 Tirso	0.5612 Tirso	0.1739 Tirso	0.2502 Mira
	0.4193 Clar	0.6602 Mira	0.7078 Guev	0.7770 Clar	0.6701 Tirso
	1.0000 Guev	1.0000 Guev	1.0000 Mira	1.0000 Guev	1.0000 Guev
Kullback Leibler	0.0000 Mira	0.0000 Clar	0.0000 Clar	0.0000 Tirso	0.0000 Mira
	0.5770 Clar	0.5322 Guev	0.1294 Tirso	0.1626 Mira	0.5719 Clar
	0.6636 Guev	0.6288 Tirso	0.8314 Guev	0.6457 Guev	0.8423 Guev
	10000 Tirso	1.0000 Mira	1.0000 Mira	1.0000 Clar	1.0000 Tirso
排名	0.0000 Mira	0.0000 Clar	0.0000 Clar	0.0000 Tirso	0.0000 Clar
	0.4000 Guev	0.7895 Guev	0.0238 Mira	0.2149 Clar	0.2410 Mira
	0.8828 Clar	0.8684 Tirso	0.3333 Guev	0.6033 Mira	0.2530 Tirso
	1.0000 Tirso	1.0000 Mira	1.0000 Tirso	1.0000 Guev	1.0000 Guev
分布式	0.0000 Mira	0.0000 Clar	0.0000 Clar	0.0000 Tirso	0.0000 Mira
	0.3204 Guev	0.0805 Mira	0.4545 Mira	0.0058 Mira	0.1944 Clar
	0.4757 Tirso	0.6356 Tirso	0.6098 Tirso	0.1930 Clar	0.4259 Tirso
	1.0000 Clar	1.0000 Guev	1.0000 Guev	1.0000 Guev	1.0000 Guev

表4. 将作者归属存疑的戏剧与正在讨论的作者进行比较所得的归一化结果。为统一单元格大小，Claramonte 和 Guevara 的名字经过了缩写。

por fuerza，与其他度量方法不同，困惑度将其归属判定为 Mira。

分布相似度首先给出的作者与 Kullback-Leibler 相同，但对于其他选项，它更倾向于给 Mira 更多机会。

在*El burlador* 和*Tan Largo*之间，四种度量方法对*El burlador*的判断更为明确，但对于这两个版本均归属判定为 Claramonte，它们相当明确且一致。四种度量方法达成的共识值得注意，尽管困惑度的数值与其他度量方法略有差异。这种轻微差异的原因在于各策略从文本中提取的 n–元语法类型不同：其他方法都使用词一元语法，而困惑度使用 7–元字符语法。表5展示了四种度量方法的归一化平均分。

《*La ninfa del cielo*》	***El burlador de*** 塞维利亚	《*Tan largo me lo fiáis*》	《*La mujer*》*por* 力度	《*El condenado*》*por* 多疑者
0.000 Mira	0.000 Cla	0.000 Clar	0.043 Tirso	0.123 Mira
0.596 Guev	0.571 Tirso	0.575 Tirso	0.193 Mira	0.192 Clar
0.631 Tirso	0.685 Mira	0.620 Mira	0.546 Clar	0.587 Tirso
0.720 Clar	0.830 Guev	0.718 Guev	0.911 Guev	0.961 Guev

表5. 将作者归属存疑的戏剧作品与讨论中的作者进行比较所得结果的平均值。

《*La ninfa del cielo*》，专家认为其作者身份属于Guevara已相当确定，但它的风格似乎与这位作者不太相近，反而与米拉·德·阿梅斯夸有更多相似之处。对于《塞维利亚的诱惑者》和《*Tan largo me lo fiáis*》，各项指标的结果都足够明确且相似，足以否定其据称属于蒂尔索·德·莫利纳的作者身份。所有

这些指标都旨在支持捍卫Claramonte作者身份的评论家的观点。《被迫的女人》和《因多疑而受罚者》是最令人存疑、也最容许其他假设的戏剧作品，因为四种距离的数值彼此相差不大。事实上，在所有情况下，排名第一和第二的作者之间的距离都小于或等于0.15。米拉和Tirso最有可能是这两部喜剧的作者。

4.4 讨论

现在，我们需要综合比较所有策略得出的各种结果。这样，我们便能得出结论，以确认或否定这些文本的作者身份。不同方法提出的作者身份判断，包括传统语文学研究的结论，均列于表6。

	语文学研究	聚类分析	距离度量
《*La ninfa del cielo*》	Tirso / Guevara	Guevara	Mira
《塞维利亚的诱惑者》	Tirso / Claramonte	Claramonte	Claramonte
《*Tan largo me lo fiáis*》	Tirso / Claramonte	Claramonte	Claramonte
《被迫的女人》	Tirso / 另一位	Tirso	Tirso
《*El condenado*》*por* 多疑者	Tirso / Claramonte / Guevara / Mira / 合作	Mira	Mira

表6. 各项策略对所讨论戏剧作品提出的作者归属判定。

《*La ninfa del cielo*》被重新归属于Guevara，这一结论似乎已成定论；然而，距离度量并未得出预期结果，尽管Stylo的树状图确实将它与Guevara的其他作品归为一组，而且这是所有结果中最明确的归属判定。尽管不能排除米拉·德·阿梅斯夸的可能性——她在黄金时代戏剧的批评讨论中常常被忽视——但她创作这部喜剧的可能性似乎不大。因此，我们的研究进一步强化了评论家支持Guevara的论据。

El burlador 和*Tan largo*是我们分析的最重要作品，所得结果极具启发性。Rodríguez López–Vázquez多年来坚持的假说很可能是正确的。所有度量都指向同一位作者：安德烈斯·德·克拉拉蒙特。这位人物一直被评论家忽视和贬低；鉴于这些结果，他的作品理应得到重新评价，其在西班牙文学经典中的地位也应重新审视。我们不能据此断言克拉拉蒙特确实是第一部唐璜戏剧的作者，但可以肯定，蒂尔索·德·莫利纳最著名的作品并非由蒂尔索·德·莫利纳创作，因为纳入语料库的其他作者（Guevara和Mira）与这部著名喜剧的风格都同样接近蒂尔索。

《被迫的女人》和《因多疑而受罚者》必须结合起来考察。如前所述（见第4.1节），其中一部是Tirso第二组喜剧中的第四部作品，因此两部不可能都属于Tirso。

Tirso的第二组喜剧作品。因此，它们不可能都出自Tirso。鉴于此前无人提出《被迫的女人》出自其他作者，且该作品所得结果也如此，最有可能的是，这第四部作品确实属于Tirso。关于《*El condenado*》，已有许多不同的归属说法，因此贸然押注其中一种风险很大。尽管本研究认为Mira是最可能的人选，并基于此前的归属判定排除了Tirso，但也有必要考虑合作创作这一可行假设；以本研究采用的策略无法验证这一假设。

简而言之，许多作品被归于蒂尔索名下，依据的只是缺乏扎实论据和文献证据的猜测与批评性论证。迄今为止，传统上归于蒂尔索名下的戏剧作品中，有很大一部分实际上并非出自这位作家之手。在 17th 世纪，喜剧常以名家的名义出版，以提高销量；后来的文学评论家似乎也受这种个人崇拜倾向影响，偏向于将作品归属判定给知名作家。蒂尔索·德·莫利纳的案例尤其令人惊讶：这些错误的归属判定层层叠加，以至于质疑其中一项，就意味着要质疑全部。值得指出的是，正是这些存在争议的戏剧作品奠定了蒂尔索在公众中的声誉，也构成了对他极高的批评评价。因此，或许应当重新审视这位剧作家在西班牙文学史上的地位。然而，更迫切的是，人们应当关注传统的作者归属判定研究，因为这些研究往往没有遵守任何人文研究都应遵循的科学严谨性基本原则。

5 结论与未来工作

将传统研究与非–传统研究结合起来，以正确解读结果后，这些结果的确凿性促使我们支持这样的理论：黄金时代戏剧领域中一些不太知名的作家长期以来被置于次要地位。在这些不太知名的作家中，尤其应当提及米拉·德·阿梅斯夸、贝莱斯·德·格瓦拉，以及安德烈斯·德·克拉拉蒙特。种种确凿证据表明，克拉拉蒙特是首部唐璜剧作的作者，这一结论似乎比以往任何时候都更近了。为其正名的一种公正做法，便是从整理出版他的作品着手。然而，尽管本研究展示了相关证据，为了进一步确认研究结果，仍有必要通过新的研究来探讨这一问题，纳入更多文本和作者，并采用更先进的自然语言处理策略，例如基于分布语义学和上下文化词嵌入的方法（Gamallo 2019）。

在未来的研究中，喜剧《因多疑而受罚者》值得单独研究，因为关于其作者身份的假说过于多样且令人困惑，无法纳入我们目前的研究。为此，有必要使用能够分别研究文本片段的工具，以确定该作品是否为合作创作，或所提出的作者中是否确有一位是真正的作者。开发和改进自动作者归属工具，实际上是

继续深入开展AAA研究所需迈出的另一项步伐。我们工作的这一方向最有前景，因为我们探索了适用于这项任务、实用且高效的计算策略，尽管这些策略最初的功能并非用于作者归属判定的风格量化。我们认为，未来有望找到其他适用于AAA研究的技术和策略。

无论如何，最亟待推动的举措是编辑并数字化黄金时代戏剧作品，以便能够大规模复现我们针对Tirso作品所做的实验。

参考文献

Almeida, D. C. (2014年). 法证用途的作者身份归属：概述与分析方案。*ReVEL*：语言研究虚拟期刊，*12(23)*,148−186。取自 http://www.revel.inf.br/files/539b2f0878d56 cb6604363c111dfe116.pdf

Argamon, S.& Juola, P. (2011年). *PAN*国际作者身份识别竞赛概述−*2011*。CLEF 2011实验室与研讨会：阿姆斯特丹。

Azfali, M. & Kumar, S. (2017年).用于寻找两篇文档相似性的各种相似度度量的比较分析。《国际数据库理论与应用期刊》，*10(2)*, 23−30。

Blasco, J. (2016). 从文体计量学视角看 Avellaneda。载于 P. Ruiz（编），《塞万提斯：旅程与日子》（第 97−116 页），马德里：Sial出版社。

Burrows, J. (2002). “Delta”：一种衡量文体差异并为判断可能的作者身份提供指引的方法。《文学与语言计算》，*17(3)*，267−287。获取自 https://academic.oup.com/dsh/article/17/3/267/929277

Calvo Tello, J. (2016). 从人文学科角度理解 Delta。《数字领域文化与批评研究》，*5(1)*，140−176。

Cavnar, W. B. & Trenkle, J. M. (1994). 基于 N−gram− 的文本分类。第三届文档分析与信息检索研讨会论文集，拉斯维加斯：内华达大学。

Cuéllar González, Á. & García−Luengos, G. V.（2017年）。EstilometríaTSO：应用于黄金时代戏剧的文体计量学http://estilometriatso.com/

Eder, M., Rybicki, J. & Kestemont, M.（2016）。使用 R语言进行文体计量学：用于计算文本分析的软件包。《*R*期刊》，*8(1)*，107-121。取自 https: // journal.r-project.org/archive/2016/RJ-2016-007/index.html

Eisen, M., Ribeiro, A., Segarra, S. & Egan, G.（2018年）。早期现代英语戏剧的文体计量分析。《人文数字学术》，*33(3)*，500−528。doi:10.1093/llc/fqx059

Gamallo, P., Garcia, M., Sotelo, S. & Pichel, J. R.（2014年）。比较基于排名的方法与朴素贝叶斯方法在推文语言检测中的表现。西班牙自然语言处理学会第三十届大会论文集，赫罗纳：SEPLN。

Gamallo, P., Pichel, J. R. & Alegria, I. (2017年)。从语言识别到语言距离。《*Physica A*》，484，162−172。

Gamallo, P., Sotelo, S., Pichel, J. R. & Artexte, M. (2019年)。利用分布式组合语义学和单语语料库实现短语动词的语境化翻译。《计算语言学杂志》，*45(3)*，395–421。

García Gómez, Á. M.（2005）。关于《塞维利亚的诱惑者》与《*Tan largo me lo fiáis*》优先权之争的文献贡献：Jerónimo Sánchez 的喜剧手稿集。载于 A.J. Close 和 S.M. Fernández Valez（编），《剑桥西班牙黄金时代》：国际黄金时代协会第七届黄金时代西班牙语研究者大会论文集，马德里：国际黄金时代协会（第 281–286页）。

García–Reidy, A.（2019）。解构《真相总会相助》的作者身份：一部 Lope de Vega 的戏剧? 《新语文学家》，1–18。取自 https: //link.springer.com/content/pdf/10.1007%2Fs11061-019-09607-8.pdf

Grieve, J.（2005）。定量作者归属判定：技术的历史与评估。伯纳比：西蒙弗雷泽大学。

Ilsemann, H.（2018年）。文体计量学进军帕纳索斯山。《人文数字学术》，*33(3)*, 548–556。doi:10.1093/llc/fqx058

Ilsemann, H.（2019年）。法证文体计量学。《人文数字学术》，*34(2)*, 335–349。doi:10.1093/llc/fqy023

Iriarte, Á., Gamallo, P. & Simões, A.（2018年）。《检测文本特异性的词汇计量策略》。《语言数学》，*10(1)*, 19–26。检索自 https://linguamatica.com/index.php/linguamatica/article/view/263

La Rosa, J. & Suárez, J. L.（2016）。《托尔梅斯河边的小拉撒路》的一生及其机器学习困境：在《小癞子》语境下的非–传统作者归属判定技术。《莱米尔：西班牙中世纪与文艺复兴文学杂志》，*20*, 373–438。

Oteiza, B. (2000). 我们了解蒂尔索·德·莫利纳的真实文本吗? 载于 I. Arellano 和 B. Oteiza（编），《蒂尔索·德·莫利纳研究杂论》（西班牙古典作品编辑中心第八届研讨会论文集）（第 99–128页）。潘普洛纳：蒂尔索研究所。

Pichel, J., Gamallo, P. & Alegría, I. (2019). 使用困惑度测量历时语言距离：应用于英语、葡萄牙语和西班牙语。自然语言工程，1–22。doi:10.1017/S1351324919000378

Potha, N. & Stamatos, E. (2014). 一种–基于特征的作者身份验证方法。载于 A. Likas、K. Blekas 和 D. Kalles（编），人工智能：方法与应用（*SETN 2014*，计算机科学讲义）（第 313–326页）。取自 https://link.springer.com/chapter/10.100 7%2F978-3-319-07064-3_25

罗德里格斯·洛佩斯–巴斯克斯，A.（1990年）。关于Claramonte与《塞维利亚的诱惑者》的争论现状。《卡斯蒂利亚：文学研究》，*Murgetana*，*82*，5–22。

罗德里格斯·洛佩斯–巴斯克斯，A.（2010年）。《被迫的女人》、《因多疑而受罚者》和《塞维利亚的诱惑者》：三部归于蒂尔索·德·莫利纳名下的喜剧。《卡斯蒂利亚：文学研究》，*1*，131–153。

Rudman, J.（1998年）。作者归属研究的现状：一些问题与解决方案。《计算机与人文科学》，*31(4)*，351–365。

Rudman, J.（2016）。威廉·莎士比亚经典作品的非传统作者归属研究：一些注意事项。早期现代研究期刊，*5*， 307−328。取自 http://www.fupress.net/index.php/bsfm-jems/article/vi ew/18094/16848

Stamatos, E. (2009). 现代作者归属判定方法综述。《美国信息科学与技术学会会刊》，*60(3)*，538−556。取自 http://citeseerx.ist.psu.edu/viewdoc/downlo ad?doi=10.1.1.440.1634&rep=rep1&type=pdf

Vázquez, L. (1995).《塞维利亚的诱惑者》：显然出自 Tirso，而非 Clara- monte（简短的批判性注释）。喜剧演员公报，*47(2)*， 183−190。Vickers, B. (2011). 书评：二十一世纪的莎士比亚与作者归属研究。《莎士比亚季刊》，*62(1)*， 106−142。