

基于机器学习的译者风格研究

——以《红字》两中译本为例

徐颖楠

中国矿业大学外国语言文化学院, 江苏 徐州

收稿日期: 2025年4月10日; 录用日期: 2025年5月21日; 发布日期: 2025年5月31日

摘要

基于机器学习的方法, 采用卡方评估法提取有效特征集, 使用支持向量机(SVM)分类器对《红字》傅东华与潘庆龄译本进行分类。研究发现, SVM 能够有效区分两个译本, 分类正确率达到 98.0%。通过分析文本, 揭示了两译者风格的差异。傅东华在翻译中偏好使用复杂定语、显化原文时态, 词汇选择相对简洁, 并频繁运用中文特殊句式, 更多地使用连词和破折号作为衔接手段。潘庆龄的翻译风格更贴近现代汉语的表达习惯, 倾向于隐化助词, 偏好使用三字词语和四字成语来装点译文, 常以冒号作为衔接手段, 更加注重译文的流畅性和可读性。通过研究验证, 机器学习方法在译者风格研究中具有可行性。

关键词

译者风格, 《红字》, 机器学习

A ML-Based Study on Translator's Style

—A Case Study of Two Chinese Translations of *The Scarlet Letter*

Yingnan Xu

School of Foreign Studies, China University of Mining and Technology, Xuzhou Jiangsu

Received: Apr. 10th, 2025; accepted: May 21st, 2025; published: May 31st, 2025

Abstract

Based on the machine learning approach, the distinctions was revealed between the Chinese translations of *The Scarlet Letter* by Fu Donghua and Pan Qingling. Utilizing the Chi-squared Attribute Eval feature selection method to extract significant style features, a Support Vector Machine (SVM) classifier was trained, achieving an accuracy of 98.0%. The analysis reveals marked stylistic divergences between the two translators. Fu's translation exhibits a preference for complex attributive

structures, explicit temporal markers, and simplified lexical choices, alongside frequent use of Chinese-specific syntactic patterns and cohesive devices such as conjunctions and dashes. In contrast, Pan's translation aligns closely with modern Mandarin conventions, characterized by the omission of auxiliary particles, the incorporation of tri-syllabic expressions and quadri-syllabic idioms for stylistic embellishment, and a reliance on colons as cohesive markers, prioritizing textual fluency and readability. These findings underscore the efficacy of computational methods in identifying translator-specific stylistic fingerprints.

Keywords

Translator's Style, *The Scarlet Letter*, Machine Learning

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

译者风格是某一译者与其他译者相比,其在语言习惯上的个人特征[1]。自 Baker 引入语料库方法探究译者风格以来,研究多基于这一路径。黄立波(2018)认为译者风格研究应当借鉴语料库文体学、计量语言学、计算语言学等相邻领域的研究方法[2]。

近年来,机器学习算法结合语料库的方法开始被应用于译者风格研究当中。詹菊红,蒋跃(2017)采用信息增益算法和支持向量机(SVM)算法成功区分《傲慢与偏见》两个中译本,但该研究风格参数多聚焦于词汇层面[3]。Lynch & Vogel (2018)综合 SVM、逻辑回归、朴素贝叶斯和决策树多种算法验证了 unigram 和 bigram 区分译者的有效性,但并未进行描述性分析[4]。仲文明,姚梦妮(2023)通过结合降维法和 SVM 考察了 *Silent Spring* 五个中译本的风格差异和变迁,发现在区分同一译者的重译作品时表现不佳[5]。鉴于此,本研究以《红字》两中译本为例,尝试构建一个较为完善的风格参数集,使用机器学习算法对两译本进行分类,并试图回答以下问题:(1) 机器学习分类模型能否有效区分不同译者的译本?(2) 哪些风格参数最能解释两个译者的风格差异?(3) 两位译者的风格存在哪些差异?

2. 研究设计

2.1. 研究语料

《红字》(*The Scarlet Letter*)是美国作家纳撒尼尔·霍桑 1850 年出版的一部长篇小说,为 19 世纪美国文学的经典之作,至今仍是学术界广泛研究与探讨的对象。本研究以傅东华译本和潘庆龄译本作为语料,语料数据见表 1。傅东华翻译的《猩红文》是《红字》的最早译本,出版于 1937 年。潘庆龄翻译的《猩红 A 字母》于 2017 年出版,为该作品的最新译本。两译本跨越 80 年,在风格上可能存在较大的差异,因此值得探讨。

Table 1. The information of two Chinese translations of *The Scarlet Letter*

表 1. *The Scarlet Letter* 两中译本详情

语料	译者	出版时间	形符	类符	样本
《猩红文》	傅东华	1937	84636	7809	24
《猩红 A 字母》	潘庆龄	2017	82967	10378	24

在语料的处理上,首先使用 OCR 技术提取文本,然后对语料进行清洗。鉴于机器学习对数据量的需求,自编 Python 脚本将两部译作的 24 个章节分别切割成 48 个独立样本。随后,使用张华平,商建云[6]团队开发的 NLPIR-Parser 对 48 个样本进行分词和词性标注。

2.2. 风格参数

尽管目前研究从多个维度探讨了译者风格研究的风格参数[1][2][7]-[9],但目前尚未形成一个固定的译者风格参数框架,尤其是针对中译本的研究。本研究基于 Leech & Short [10]的小说文体学分析框架,从词汇、句子、修辞和语篇四个层面进行综合考量,并融入中文相关参数如四字词语、中文特殊句式“把”字句、“被”字句、“有”字句、“是”字句等[11][12]。在句子层面,还借鉴了吴思远等[13]的算法,在 Python 中调用 Stanford Core NLP 生成句法树,统计各类型短语的数量。风格参数如表 2 所示,共计 48 个风格参数。除 STTR、平均词长、平均句长使用 wordsmith7.0 提取外,其余数据均使用 Python 自编脚本批量提取。四字词语在 Python 使用自编正则表达式提取后,人工剔除人名、地名后重新统计数量。

Table 2. Style attributes
表 2. 风格参数

词汇层面	宏观特征	STTR、平均词长、平均词长标准差
	词性分布	实词、名词、动词、形容词、数词、量词、副词比例 虚词、介词、助词、语气词、叹词、象声词比例
句子层面	宏观特征	平均句长、句长标准差
	句型分布	陈述句、感叹句、疑问句比例
	短语分布	名词短语、动词短语、副词短语、形容词短语、介词短语、方位词短语、限定词短语、量词短语、由“的”构成的表示修饰关系的短语、由“的”构成的表示所属关系的短语数量
	中文特殊句式	“把”字句、“被”字句、“是”字句、“有”字句比例
修辞层面	四字词语	四字词语、ABAC 式词语、ABCC 式词语、AABB 式词语、ABCB 式词语、ABBC 式词语、ABAB 式词语、ABCA 式词语比例
语篇层面	衔接	连词、代词比例
		冒号、破折号、省略号比例

2.3. 研究方法

2.3.1. 风格参数选取

在机器学习算法中,特征选择是提升模型性能的重要步骤,常见方法包括主成分分析法[5][14]、信息增益法[3]以及卡方评估法[4][15]。本研究选择卡方评估法(Chi-squared Attribute Eval)作为特征的选取方法,该方法的优点在于它可以帮助识别在不同类别中具有显著分布差异的特征,且不受限于数据分布形态,能够用于非正态分布的数据集。根据卡方统计量值的高低,从词汇、句子、修辞、语篇四个层面各筛选 3 个风格参数,共计 12 个风格参数,如表 3 所示。

2.3.2. 分类器和算法

得到特征集后,采用支持向量机(Support Vector Machine, SVM)对两位译者的样本进行分类。SVM 是一种监督学习算法,主要用于分类和回归任务。本研究调用 Weka 平台中的软件包 LIBSVM,并通过 GridSearch 算法优化参数,避免模型过度拟合或欠拟合。此外,还需对数据进行归一化(normalize)处理,提高 SVM 算法的性能。

Table 3. Style attributes selected by Chi-squared attribute eval

表 3. 卡方评估选取的风格参数

	Average merit	Average rank	Attribute
	43.2 ± 0.4	1 ± 0	助词比例
词汇层面	36.33 ± 1.476	3.2 ± 0.6	平均词长
	30.262 ± 1.9	5.4 ± 0.66	STTR
	24.773 ± 1.9	6.8 ± 0.4	由“的”构成的表示所属关系的短语数量
句子层面	14.62 ± 2.557	9 ± 1.1	是字句比例
	14.62 ± 1.718	10.4 ± 1.2	把字句比例
	39.758 ± 1.445	2.1 ± 0.3	四字词语比例
修辞层面	28.324 ± 1.423	5.8 ± 0.6	ABAC 式词语比例
	13.917 ± 1.456	9.3 ± 1.35	AABB 式词语比例
	3.938 ± 6.088	10.2 ± 1.25	连词比例
语篇层面	4.874 ± 4.884	10.4 ± 0.66	破折号比例
	1.262 ± 3.787	11.7 ± 0.9	冒号比例

3. 分类结果

处理后的数据以.csv 格式储存并导入分类器,采用十折交叉验证法(Ten folds Cross-Validation)评估模型,即每次迭代以一子集测试,九子集训练,汇总十次结果计算平均性能指标。评估包括准确率、召回率及 AUC 值。准确率衡量预测正确比例;召回率衡量正样本预测准确率;AUC 反映分类能力,越接近于 1,模型的分类能力越佳。本次分类结果见表 4 和表 5。结果显示,48 个样本中 47 个样本被正确分类,仅一个样本被误分,平均分类准确率达 98.0%,误分率仅 2.0%,平均召回率和 AUC 值均趋近于 1,说明模型分类和预测能力较好。

Table 4. SVM classification result

表 4. SVM 分类结果

类别	准确率	召回率	AUC
傅译本	0.960	1.000	0.979
潘译本	1.000	0.958	0.979
Weighted Avg.	0.980	0.979	0.979

Table 5. Confusion matrix
表 5. 混淆矩阵

a	b	分类结果
24	0	a = 傅译本
1	23	b = 潘译本

4. 讨论

经过验证，根据卡方评估选取的 12 个风格参数，SVM 算法能够成功地区分 *The Scarlet Letter* 两个中译本。因此根据这些风格参数进一步讨论两位译者风格上的具体差异，风格参数数据如表 6 所示。

Table 6. Statistics of 12 selected style attributes of two translations
表 6. 两译本选取的 12 个风格参数数据

	风格参数	傅译本	潘译本
词汇层面	助词比例	12.46%	8.92%
	平均词长	1.46	1.66
	STTR	43.37	50.24
句子层面	由“的”构成的表示所属关系的短语数量	205	116
	是字句比例	39.25%	30.79%
	把字句比例	9.55%	6.0%
	四字词语比例	0.25%	1.60%
修辞层面	ABAC 式词语比例	0.03%	0.13%
	AABB 式词语比例	0.03%	0.09%
	连词比例	3.12%	2.60%
语篇层面	破折号比例	5.31%	3.97%
	冒号比例	0.58%	1.06%

4.1. 词汇层面

词汇层面主要探讨助词、平均词长和 STTR。数据表明，傅译本与潘译本在助词得使用上存在显著差异。助词主要包括结构助词，如的、地、得等与动态助词，如着、了、过等。傅译本助词使用频率均值为 12.46%，显著高于潘译本的 8.92%。如例 1 中，傅东华高频使用嵌套式修饰关系，构成了复杂定语。潘庆龄则简化“的”的使用，通过合并修饰成分，将“的”字使用频率降低，避免定语过多，这类虚词的减少也可使 STTR 更高，词汇丰富度更高。在结构助词的使用方面，傅东华也更加倾向于显化原文时态，如例 2，原文使用了过去时的被动语态，傅东华使用了两个动态助词“了”来体现时态，而潘庆龄则没有使用明显的时态标记。

例 1: But on one side of the portal, and rooted almost at the threshold, was a wild rose-bush, covered, in this month of June, with its delicate gems, which might be imagined to offer their fragrance and fragile beauty to the prisoner as he went in,

and to the condemned criminal as he came forth to his doom, in token that the deep heart of Nature could pity and be kind to him [16].

傅译: 我们希望它可以替这故事中也也许沿途会得发现的那种美丽的道德的花做一个象征, 或者替这关于人类脆弱或悲哀的故事的阴暗的结束做一种调剂[17].

潘译: 但愿这朵野玫瑰能用来象征在这个讲述人性脆弱和悲哀的故事中可以发现馥郁芳香的道德之花, 也可在读到故事令人黯然神伤的结局时备不住得到一些宽慰[18].

例 2: The very ideal of ignominy was embodied and made manifest in this contrivance of wood and iron [16].

傅译: 所谓耻辱的理想, 是在这种木头和铁做成的器械之中体现了, 显示了[17].

潘译: 这种铁木结构的刑具分明极尽让人含垢忍辱之能事[18].

就平均词长而言, 傅译本平均词长为 1.46 个字, 而潘译本为 1.66 个字, 傅译本的平均词长更贴近汉语翻译文学均值 1.44 个字[19], 这种差异源于两位译者翻译时对词汇的差异化选择。潘庆龄较傅东华更频繁使用四字格和三字词语, 如例 3 中, 潘庆龄在翻译“cold”时使用了 ABB 式三字词语“冷冰冰”的, 翻译“ridicule”和“stern”均使用了成语, 使得其平均词长更长。而傅东华多以二字形容词翻译, 使得其译文风格更为平实, 平均词长较短。

例 3: Meagre, indeed, and cold, was the sympathy that a transgressor might look for, from such bystanders at the scaffold. On the other hand, a penalty which, in our days, would infer a degree of mocking infamy and ridicule, might then be invested with almost as stern a dignity as the punishment of death itself [16].

傅译: 犯罪的人登上断头台, 所得期望于这种旁观人给予的同情, 实在是稀薄而又冷漠的。从他方面看, 则我们现在不免有些儿要算做不名誉和可笑的一种刑罚, 在当时却也许要给予差不多和死刑一般严肃的一种尊严[17].

潘译: 一个站在刑台的罪人, 从这些围观者中间所能见到的同情, 端的是少得可怜, 而又冷冰冰的。余外, 现如今或恐仅仅意味冷嘲热讽的一种惩罚, 可在当时却被赋予犹如死刑般令人发指的严酷色调[18].

4.2. 句子层面

前文已经指出, 傅东华以“的”为标记的复杂定语的使用倾向性更加明显。从句法层面上来看, 其译本中由“的”构成的表示所属关系的短语的平均使用量为 205 个, 远超潘译本的 115 个, 印证了词汇层面的发现。如例 4 中, 傅东华在翻译“fire in her and throughout her”时, 使用了两个由“的”构成的表示所属关系的短语。

例 4: There was fire in her and throughout her: she seemed the unpremeditated offshoot of a passionate moment [16].

傅译: 有火在她的身上, 贯澈她的全身, 她似乎是一个热情的俄顷的未曾预计的萌芽[17].

潘译: 瞧她浑身上下是一团火, 活像情欲激越瞬间不期然而来的旁系亲属[18].

“是”字句是中文的特殊句式之一, 通常用来强调主语或谓语, 表明一种存在。傅译本的“是”字句平均占比(39.25%)要高于潘译本(30.79%)。有研究认为, 翻译汉语中“是”字句增多与英语原文中 be 动词的高频使用和使用范围有关[20] [21]。然而, 傅东华对“是”字句结构的偏好还体现在原文没有 be 动词的情况下。如例 5, 从句子上来看, 傅东华采用了“是”字句的结构, 通过译为“是”字句将补语成分主位化, 从而使译文更直白, 而潘庆龄使用连动结构, 更加具有文学色彩。

例 5: The rose-bush, by a strange chance, has been kept alive in history [16].

傅译: 由于一种奇异的机缘, 这一棵蔷薇花丛是历史上一向有名的[17].

潘译: 由于某种奇妙的机缘, 这丛野玫瑰却历尽劫难而永生[18].

“把”字句也是常用的中文特殊句式之一。“把”字句的特点在于动词前后常带有其他成分，动词不能单独出现，尤其是单音节动词，其后通常跟有补语、宾语、动态助词或动词的重叠式[22]。傅译本中“把”字句的平均占比高达 9.55%，潘译本仅为 6.0%，傅东华更倾向于采用“把”字句结构。如例 6，傅东华将动词“show”译为“把……显显”，采用了“把 + 宾语 + 状语 + 动词 + 助词”的结构，保留了原文的指令性。潘庆龄则译为“秀一秀”，“秀”作为动词，表示展示、炫耀，来源于英文单词“show”的音译。潘庆龄的译法更贴近现代汉语的表达方式。

例 6: Come along, Madame Hester, and show your scarlet letter in the market place [16]!

傅译: 来罢, 赫丝脱小姐, 你可以把你那个猩红字母在市场上显显了[17]!

潘译: 来吧, 赫丝黛太太, 上集市去秀一秀你的猩红 A 字母[18]!

4.3. 修辞层面

在中文中，四字词语通常含有丰富的意义，同时四字词语还有修辞功能，例如常见的 AABB 式词语不仅有重复修辞的作用，还有增强文本韵律等作用。潘译本的四字词语使用量显著多于傅东华译本。若从数量上来看，经过统计发现，潘庆龄平均每章使用 56 个四字词语，而傅东华仅平均使用约 9 个。相对而言，潘庆龄尤其偏好使用 ABAC 式和 AABB 式词语，如例 7，潘庆龄连续使用了两个 AABB 式和一个 ABAC 式四字词语，文本节奏与韵律较强，文学韵味浓厚。而傅东华则没有使用任何四字词语，译文更为简洁。

例 7: ...these two men, so different in age, came gradually to spend much time together. (154)

傅译: 因此这两个人虽然年纪相差很远, 却渐渐的往来颇密了。(170)

潘译: 论年龄, 这两个人固然相差悬殊, 不知不觉地却在一起消磨了许许多多日日夜夜。(150)

4.4. 语篇层面

连词是词汇衔接的重要手段之一，从表 6 的数据来看，傅东华对于连词的使用更加频繁，例如“这一下看，使得她的心收缩而颤抖，因为那眼光是如此的熟悉，然而又是如此的陌生而冷酷的，——及至最后觉得他的诊察已经满意，就动手配起另外一服药来”[17]这句译文中便出现了“而”、“因为”、“然而”、“及至”、“就”多个连词。

标点符号是语篇衔接的其中一种手段[23]，源文本中使用了大量的破折号来衔接上下文。在中文中，冒号可以引出分说内容、破折号可以补充说明等。根据数据，傅东华更频繁使用破折号，如例 8 中源文本有两处破折号用于插入补充说明，傅东华在翻译时更注重形式上的对等，保留原文的标点和结构，导致破折号使用较多。而潘庆龄更注重译文的流畅，因此减少破折号，转而使用逗号连接。由于英语中常出现以逗号连接的插入语、同位语等，通常是对先行词进行补充、解释和说明，在汉译时译者通常需要将这种逻辑关系表达出来。潘庆龄在处理这类句子时时常使用冒号充当衔接手段，如例 9，“with the left shoulder a trifle higher than the right”是补充说明“slightly deformed”的具体表现，潘庆龄采用冒号，将这一逻辑关系译出。

例 8: Notwithstanding his high native gifts and scholar-like attainments, there was an air about this young minister—an apprehensive, a startled, a half-frightened look—as of a being who felt himself quite astray and at a loss in the pathway of human existence, and could only be at ease in some seclusion of his own [16].

傅译: 这位青年的牧师虽然具有那么高的天分和学问的造诣, 却总流露着一种神气, ——一种恐吓的, 惊愕的, 一半惶骇的神情, ——仿佛一个人觉得自己已经深深陷入了迷途, 在人类生活的径路上不知所往, 只有在独自幽隐

的地方才觉得舒服似的[17]。

潘译: 这位年轻牧师虽有极高的天赋和学者般的造诣, 但在他身上却凸现出一种忧心忡忡、诚惶诚恐的神色, 好像某一个人自己感悟到迷失了方向, 在人生的道路茫然不知所措, 唯有一个人独处时方才觉得轻松自在[18]。

例 9: This figure of the study and the cloister, as Hester Prynne's womanly fancy failed not to recall, was slightly deformed, with the left shoulder a trifle higher than the right[16]。

傅译: 这一位书斋和修道院中的人物, 照赫丝脱·普林的女性的幻想形容起来, 是稍稍有点畸形的, 左边的肩膀比右边的略微高出一点[17]。

潘译: 赫丝黛·普丽思的女性想象力欲罢还休, 依然回想起来, 那个学者兼修道士的身形不知怎的还是出现了: 他稍微有点儿畸形, 左肩比右肩略高些[18]。

5. 结语

研究表明, 支持向量机分类器可以根据卡方评估选取的特征, 有效地将不同译者翻译的同一作品区分开来。通过对比分析两个译本发现, 两位译者在词汇、句子、修辞、语篇各个不同层面均具有显著的差异。在当前数字人文的背景下, 机器学习算法作为一种数字技术, 为挖掘文本背后的译者风格特征提供了有力工具, 为语料库翻译学提供了新的研究方向。

参考文献

- [1] Baker, M. (2000) Towards a Methodology for Investigating the Style of a Literary Translator. *Target. International Journal of Translation Studies*, 12, 241-266. <https://doi.org/10.1075/target.12.2.04bak>
- [2] 黄立波. 语料库译者风格研究反思[J]. 外语教学, 2018, 39(1): 77-81.
- [3] 詹菊红, 蒋跃. 机器学习算法在翻译风格研究中的应用[J]. 外语教学, 2017, 38(5): 80-85.
- [4] Lynch, G. and Vogel, C. (2018) The Translator's Visibility: Detecting Translatorial Fingerprints in Contemporaneous Parallel Translations. *Computer Speech & Language*, 52, 79-104. <https://doi.org/10.1016/j.csl.2018.05.002>
- [5] 仲文明, 姚梦妮. 基于降维分类模型的译者风格研究[J]. 外语电化教学, 2023(4): 24-31.
- [6] 张华平, 商建云. NLPPIR-Parser: 大数据语义智能分析平台[J]. 语料库语言学, 2019(1): 87-104.
- [7] Olohan M. (2004) *Introducing Corpora in Translation Studies*, Routledge.
- [8] Hu, K.B. (2016) *New Frontiers in Translation Studies*. Springer, Berlin & Heidelberg.
- [9] 胡开宝, 黑黝. 数字人文视域下翻译研究: 特征、领域与意义[J]. 中国翻译, 2020, 41(2): 5-15.
- [10] Leech, G. and Short, M.H. (2007) *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2nd Edition, Pearson Education.
- [11] 胡开宝, 崔薇. 基于语料库的莎士比亚戏剧汉译本中“使”字句应用的研究[J]. 外语教学, 2016, 37(1): 102-106.
- [12] 刘艳春, 胡显耀. 基于语料库的翻译汉语变体多特征统计分析[J]. 语料库研究前沿, 2021(1): 25-36.
- [13] 吴思远, 于东, 江新. 汉语文本可读性特征体系构建和效度验证[J]. 世界汉语教学, 2020, 34(1): 81-97.
- [14] 董琇. 基于降维法的译者风格研究[J]. 外语教学与研究, 2014, 46(2): 282-293.
- [15] 孔德璐. 机器学习视域下《三国演义》三译本翻译风格对比研究[J]. 大连大学学报, 2023, 44(4): 38-47.
- [16] Hawthorn, N. (2002) *The Scarlet Letter*. Penguin Classics.
- [17] 霍桑. 猩红文[M]. 傅东华, 译. 北京: 商务印书馆, 1947.
- [18] 霍桑. 猩红 A 字母[M]. 潘庆龄, 译. 北京: 九州出版社, 2017.
- [19] 胡显耀. 基于语料库的汉语翻译语体特征多维分析[J]. 外语教学与研究, 2010, 42(6): 451-458.
- [20] 肖忠华. 英汉翻译中的汉语译文语料库研究[M]. 上海: 上海交通大学出版社, 2012.
- [21] 戴光荣. 译文源语透过效应研究[M]. 上海: 上海交通大学出版社, 2013.
- [22] 黄伯荣, 廖序东. 现代汉语(增订六版)上册[M]. 北京: 高等教育出版社, 2017.
- [23] 胡明亮. 语篇衔接与翻译[M]. 成都: 巴蜀书社, 2007.