机器学习视域下《三国演义》三译本翻译风格对比研究

孔德璐

（同济大学 外国语学院，上海 200092）

摘　要：本文使用机器学习中的分类和聚类方法，对中国四大名著之一的《三国演义》三译本进行翻译风格考察，从特征集中筛选出20个显著特征，并结合语篇进行阐释和总结。研究结果表明，筛选后的特征能够有效区分三译本的风格差异，分类器和k-means聚类分析能够实现的平均准确率均达到95%以上。在篇章层面，各译本均在词汇、符号上显示出不同的风格特征；在关键词特征方面，使用频次差异足以显示出译者的个人偏好。本研究以期对翻译风格创新对比研究提供一定实用方法。

关键词：机器学习；翻译风格；《三国演义》

中图分类号：H059　　**文献标识码：**A　　**文章编号：**1008-2395（2023）04-0038-10

收稿日期：2022-11-23

作者简介：孔德璐（1996- ），男，博士研究生，主要从事机器翻译、语料库翻译学研究。

《三国演义》被誉为是中国“四大名著”之一，因其卓越的艺术成就和复杂丰富的思想内涵，受到历代中外学者的普遍关注。19世纪，随着西方知识传入中国，译者开始将中国一些古典作品外译，引起西方对中国的关注[1]。《三国演义》的译介也是从那个时代开始，1820年英国人汤姆斯（Peter Perring Thomas）第一次将《三国演义》的节选“The Death of the Celebrated Minister Tung-Cho”（《丞相董卓之死》）翻译成英文，并发表在《亚洲杂志》上，距今已经刚好有200年的英译历史[2]。

《三国演义》因其雄伟庞大的历史叙事、细致入微的人物描写，以及偏文言文的古代白话文风，致使该书的外译极为棘手。据统计，截至目前，《三国演义》英文节译本众多，但全译本只有三部，分别是1925年上海别发洋行出版的邓罗译本（C. H. Brewitt-Taylor）、1992年加利福尼亚大学出版社和外文出版社联合出版的罗慕士译本（Moss Robert），以及2014年由新加坡塔托出版公司出版的虞苏美译本。了解典籍翻译过程中的风格因素有助于对中国文化经典文学向外传播的再思考，“中国文化经典的有效传播是中国文化走出去的重要一环”[3]，在“一带一路”的大背景下，中国既需要了解世界，也需要让世界更深入地了解我们，基于此，本文将针对三个不同译者的《三国演义》中文译本，运用机器学习算法，对中国典籍文学的翻译风格特征进行量化对比研究，以期为中国典籍文学的外译提供参考。

一、文献回顾

翻译风格研究是语料库语言学和基于语料库的翻译研究中的热门话题之一，它主要涉及翻译语言模式的发现和特征的阐释过程，此类研究随着语料库语言学的发展成熟而变得更加高效。译者风格，又称译者的翻译风格，是指译者在翻译文本选择、翻译策略与方法的应用以及翻译文本的语言应用等方面所表现的个性化特征[4]。翻译传统常常要求译作要忠实于原文原作，译文风格被视为是源语文本作者风格的再现。

基于语料库的译者风格研究开展于21世纪初，这主要依赖于语料库技术的发展和成熟。Hermans认为，译本中隐藏着“译者的声音（Translator's Voice）”[5]。Baker提出译文中会不可避免地展现“译者的指纹（Thumbprint）”[6]，并提出文学翻译风格考察的框架，且已被广泛用于基于语料库的翻译研究中，其中一些典型参数（例如形符比、平均句长和特殊动词等）常被用来检验某一译者或译作的风格。然而，基于语料库的翻译风格研究范式还是存

在着有待完善的空间。Mikhailov 和 Villikka 使用语料库的方法，通过研究多个俄语和芬兰语对应译本，发现在作者判定研究中常用的参数并不能有效判断译者风格 [7]。另外，大多数的基于语料库的翻译风格研究中，研究者会预先设定并过于强调某些特征来反映译作风格的作用，却忽略掉了一些更能突显风格的隐性特征，削弱了总体概括能力；同时这种预先设定研究特征的做法已经打上了研究者主观判断的“烙印”，也被批评为“总是基于学者的自觉……显然这种方法是‘人工设计’的”[8]。

机器学习方法可以有效弥补基于语料库翻译风格研究框架的局限，通过建立特征集并利用算法提取出最突显的译作风格特征。该方法缘起于计量风格学（Stylometry），主要包括采用机器学习方法对作者的创作风格和作品风格进行比较研究。Lynch 和 Vogel 利用支持向量机模型，证明英语中的 N 元语法可以有效识别译者风格 [9]。近年来国内学者也开始将机器学习方法灵活运用到定量翻译研究中，如詹菊红、蒋跃使用支持向量机模型区分《傲慢与偏见》的两个中文译本，并对部分显著特征进行阐释 [10]。

前文述及，《三国演义》作为中国经典名著之一，目前学界对于三国的研究多为针对单一译本的译作介绍 [11,12]，或从某个角度来对译本进行评价分析，如副文本角度 [13,14]、词汇中的骂词 [15] 等，抑或是从翻译策略、翻译范式来分析某一译本 [16-18]。

整体来看，学界对三国多译本对比分析的讨论并不多见，运用语料库等量化手段来研究三国多译本则更是阙如。其中，董琇在词汇、句法、篇章层面，通过对比邓罗译本，突显出罗慕士译本的翻译特征 [19]。冉明志通过对比邓、罗两本译本中的军事术语，得出邓译偏归化、罗译偏异化的结论 [20]。目前尚未发现有学者运用计量方法对三国多译本的翻译风格进行研究，这也为本文的研究提供了空间和前提。

本研究采用机器学习的方法，通过自建《国富论》平行语料库，对三位译者的文本进行分类和聚类，用十折交叉验证法（10-folds Cross Validation）对分类器效果进行检验，在特征解释部分，我们将运用语料的统计数据，对差异性语言形式特征进行定性探究。本研究拟回答以下问题：

1. 机器学习方法能否区分不同译本的风格？
2. 哪些特征最能解释不同译本的风格差异？
3. 如何解释这些差异特征并进行分类和归纳？

二、研究设计

（一）语料选择

此次研究选择《三国演义》原文以及三个不同译本作为研究语料。中文原本选择华夏出版社的《三国演义》重印本；三译本分别选择 1925 年上海别发洋行出版的邓罗译本（C. H. Brewitt-Taylor，以下简称邓译本）、1992 年加利福尼亚大学出版社和外文出版社联合出版的罗慕士译本（Moss Robert，以下简称罗译本），以及 2014 年由新加坡塔托出版公司出版的虞苏美译本（以下简称虞译本）。

首先，对 OCR 后的语料进行除噪，只保留正文部分作为研究语料，并用 TreeTagger 对英语语料进行分词和标注。随后，按照 UTF-8 编码，并依照全书共 120 回章节，对三个英文语译文样本切分。由于标注系统和编码类型的不同，较难实现跨语言特征集的建立，因此《三国演义》中文源本不参与译本风格分类，未进行切分处理。最终建成的英汉双语平行语料详情如表 1 所示，最终实验所用样本共计 360 个。

表 1　语料详情

作品	样本	形符数	类符数	句子数
邓译本	120	598321	13793	36025
罗译本	120	551789	15526	37848
虞译本	120	601886	14970	27241

（二）特征集建立

参照以往研究中对于特征的筛选和使用 [9,10]，本文建立的特征集主要分为两个大类，一类为篇章级特征（Document-level），包括从词汇、句法到篇章的三个层次共计 45 个具体特征；第二类为关键词特征（Keywords），选取每个词表中关键值最高的 20 个主题词，总词表去重后得到 55 个关键词，最

终选取45个篇章级特征、55个关键词特征，共计100个实验特征（如表2所示）。篇章级特征主要针对语篇的宏观层面，从具体的词类分布、句子类型和标点使用等因素来考察各译本间的风格差异；而通过对比所提取的关键词特征，则能够更好地从微观角度来分析各译本风格差别，从而体现译者的下意识个人用词偏好。

表2 机器学习实验中所使用的特征集

篇章特征			关键词特征
形符比	代词比例	叹号比例	men, prefect, advisor, he, don, officers, to, his, mans, bei, advisors, that, so, was, liang, zhuge, him, officer, m, also, and, shall, miles, o, wherefore, bade, they, of, changan, till, latter, mangs, leader, yiazhou, upon, xuande, kongming, s, riverlands, zilong, lord, commanders, southland, nt, responded, excellency, answered, honour, field, realm, troops, sovereign, raced, honoured, cao
标准形符比	形容词比例	逗号比例	
平均词长	副词比例	分号比例	
平均段长	数词比例	括号比例	
平均句长	介词比例	引号比例	
名词比例	介词 to 比例	冒号比例	
动词比例	介词 in 比例	连接符比例	
Be 动词比例	介词 into 比例	缩写符比例	
Do 动词比例	介词 on 比例	实词比例	
Have 动词比例	介词 of 比例	动词 say 比例	
动词过去时比例	连词比例	动词 saying 比例	
动词现在分词比例	存现句比例	动词 says 比例	
动词过去分词比例	叹词比例	动词 said 比例	
动词现在时比例	句号比例	词组比例	
情态动词比例	问号比例	wh- 词比例	

注：（1）词类比例 = 每种词类频次 / 总词数，标点符号和空白字符不算入总词数；（2）句型比例 = 每种句型频次 / 总句数，以句号、问号、叹号、分号、冒号作为判断句子的标准；（3）符号比例 = 每种符号频次 / 总符号数；（4）实词比例 = 总实词数 / 总词数，实词密度 = 总实词数 / 总虚词数；（5）关键词特征均为比例，即每个关键词频次 / 总词数。

各译本经过切分后，总样本数量达到300余例。借助自编Python代码和语料库检索工具WordSmith 6.0，可以批量提取出形符比、标准形符比、总词数、总句数、词类频次、标点符号的频次和总数、总成语数等，节省人工成本。最终经过统计，将形成的360个文本样本转换为对应的基于100个特征的数学表达模型，使用支持向量机（Support Vector Machine，简称为SVM）、简单逻辑回归（Simple Logistic Regression）、C4.5决策树作为分类器。通过机器学习平台Weka，验证分类器效果，找到能够区分三个译本的主要特征，结合实际语料进行规律总结和阐释。

三、研究方法和结果

（一）分类器和算法

机器学习过程涉及多种分类器和算法，在本次研究中，我们使用SVM、朴素贝叶斯和决策树C4.5分类器。SVM分类器在1963年由著名的苏联数学家弗拉基米尔·瓦普尼克等人设计，并在随后的广泛应用中展示出良好的性能。SVM最初发展于线性可分情况，其本质是找到基于支持向量的最优的分类面，分类线方程表示为W·X+b=0，其中W和b分别为超平面的法向量和截距[21]。

本研究使用台湾大学林智仁教授团队开发的SVM模式识别与回归的软件包LIBSVM，旨在帮

助用户将 SVM 便捷地应用于实践当中。在 Weka 平台中，可从扩展库直接下载该软件包。通过 GridSearch 算法调参，最终 LIBSVM 的参数设置如图 1 所示，参考分类器均使用默认设置。

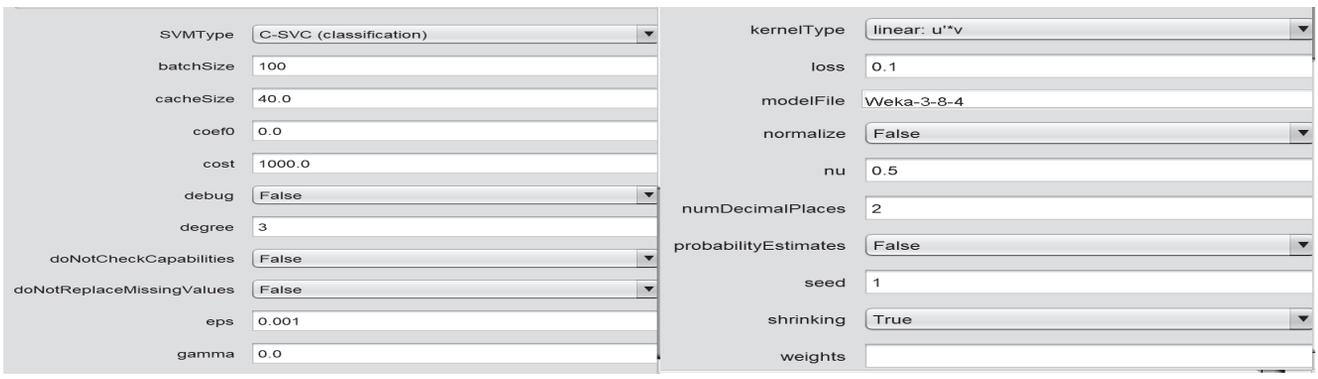


图 1　LIBSVM 参数设置

（二）实验结果

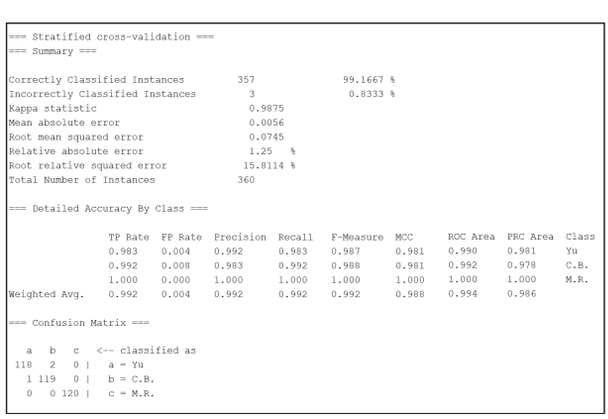
将预处理后的数据导入分类器，使用十折交叉验证法进行分类模型精度评估。该方法主要是将总数据集随机切分成 K 份，每次运行时都使用其中 1 份作为测试集，剩下 K-1 份作为训练集，并重复验证 K 次，这种方法就叫做 K 折交叉验证。将 K 取值为 10，就成为常见的十折交叉验证法，最后的精度验证结果是十次重复验证结果的平均值。为选取分类过程中最显著的特征，本研究利用 Weka 中内置的特征选择分类器（Attribute Selection Classifier），使用卡方评估法（Chi-squared Attribute Eval），通过计算各类别中样本例的卡方统计量的值来评估特征价值，从结果中挑选出 20 个卡方值最高的参数，并利用挑选后的特征再次进行实验，以检验其分类效果。最终的实验结果如表 3 所示。

表 3　特征选择前后分类结果

分类器	特征选择前（全部 100 特征）			特征选择后（显著 20 特征）		
	准确率	召回率	AUC*	准确率	召回率	AUC
SVM	68.6111 %	0.686	0.765	99.1667 %	0.992	0.994
朴素贝叶斯	100 %	1.000	1.000	99.1667 %	0.992	0.998
C4.5 决策树	97.5 %	0.975	0.981	97.7778 %	0.978	0.981

注：AUC（Area Under Curve）为 ROC 曲线下与坐标轴围成的面积。

表 3 中每组呈现的结果主要包含三个指标：准确率，表示分类样本在样本数中的占比；召回率，表示样本中的正例有多少被预测正确；以及 AUC，用来反映模型预测能力。其中，召回率和 AUC 越接近 1，模型分类结果和预测能力越好。整体而言，使用特征选择前全部 100 个特征构建的分类器模型中，除了 SVM 分类器，其余两个分类器已经达到了理想的结果，其中基于概率统计的朴素贝叶斯方法取得最优分类结果。通过对比，经过特征选择后提取的 20 个显著特征同样在三个分类器中得到了最优秀的结果，平均准确率达到 98.7%，平均召回率和 AUC 值均达到 0.9 之上，证明其分类结果和预测能力较好。使用特征选择方法可以在确保准确率的同时，减少实验的复杂度，节省实验时间，剔除噪声特征，同时可以提取出在文本分类过程中贡献较大的特征，以便进行聚类检验和特征阐释 [22]。



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         357               99.1667 %
Incorrectly Classified Instances         3                0.8333 %
Kappa statistic                          0.9875
Mean absolute error                      0.0056
Root mean squared error                  0.0745
Relative absolute error                  1.25   %
Root relative squared error             15.8114 %
Total Number of Instances              360

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.983     0.004     0.992      0.983     0.987      0.981    0.990     0.981     Yu
               0.992     0.008     0.983      0.992     0.988      0.981    0.992     0.978     C.B.
               1.000     0.000     1.000      1.000     1.000      1.000    1.000     1.000     M.R.
Weighted Avg.  0.992     0.004     0.992      0.992     0.992      0.988    0.994     0.986

=== Confusion Matrix ===

   a   b   c   <-- classified as
 118   2   0 |   a = Yu
   1 119   0 |   b = C.B.
   0   0 120 |   c = M.R.
```

图 2　基于显著特征的 SVM 分类器结果

图 2 显示的是基于 20 个显著特征的 SVM 分类器结果，在 360 个《三国演义》的样本中，共有 357 个样本被正确分类。图底部的混淆矩阵（Confusion Matrix）反映出样本的具体分类情况，其中，罗译本表现出较好的整体性，120 个样本全部分类正确，且无其他译者的错误分类样本。误分样本主要出现在虞译本和邓译本中，前者有 2 个样本被分入邓译本中，而后者有 1 个分入了虞译本。纵使出现了 3 个错误样本，较高的准确率（99%）也足以证明经过特征筛选后的 20 个显著特征可以满足区分译本风格的需求。表 4 列出 20 个经过特征筛选后对应的特征，其中斜体的属于篇章级，反映出宏观词汇、句法层面的特征，共 7 个；其他为关键词特征，体现不同译本当中较为微观的用词差异，共 13 个。

表 4　经过特征选择的前 20 个显著特征

卡方范围值	平均排名	特征	卡方范围值	平均排名	特征
411.98 -11.253	1	*缩写符比例*	238.939 +- 6.27	11	shall
338.851 +- 7.308	2	men	236.673 +- 4.961	12	so
290.449 +- 4.246	3	responded	225.747 +- 5.889	13	wherefore
266.617 +- 5.212	4	*实词比例*	220.541 +- 5.632	14	commanders
258.939 +- 9.949	5	’s	214.19 +- 5.08	15	miles
254.282 +- 7.284	6	*分号比例*	212.497 +- 8.247	16	*连词比例*
251.462 +- 8.198	7	*Be 动词比例*	206.666 +- 5.455	17	O
245.312 +- 4.298	8	n’t	205.5 +- 4.082	18	*平均词长*
240.877 +- 3.824	9	*名词比例*	192.343 +- 9.777	19	was
237.906 +- 4.588	10	answered	182.19 +- 3.103	20	southland

（三）聚类分析

聚类分析（Cluster Analysis）可以将数据对象划分成多个类或“簇”，使同一类或簇中的对象具有较高的相似度，而不同类的对象间差异尽可能大[23]123。作为一种无监督的机器学习方法，聚类分析能够较为直观地展现可视化结果，在各领域运用范围广泛；在文本分析研究中，如风格判定、作者识别、韵律分析等方面发挥较大的作用。

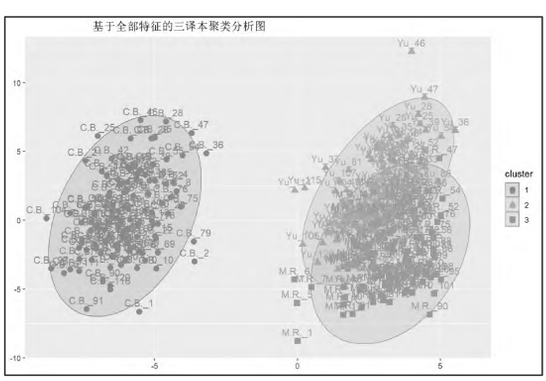


图 3　三译本基于全部特征的 K-means 聚类结果

本研究首先使用全部 100 个特征对三个译本进行聚类分析，使用 R 语言中内置的 k-means 函数，生成三个译本共 360 个样本的聚类图（图 3），置信椭圆为 95%。如图 3 所示，圆点表示的是邓罗译本的样本，三角表示的大多是罗慕士译本的样本，方块表示的大多是虞苏美译文的样本。虽然邓译本风格较为统一，可以与虞译本和罗译本的风格相区分开，但罗译本和虞译本的聚类效果不佳，两个簇的样本大多重叠在了一起，说明二者在这全部 100 个特征的向量化表征上来看，并不能将这两个译本的风格很好地区分开。

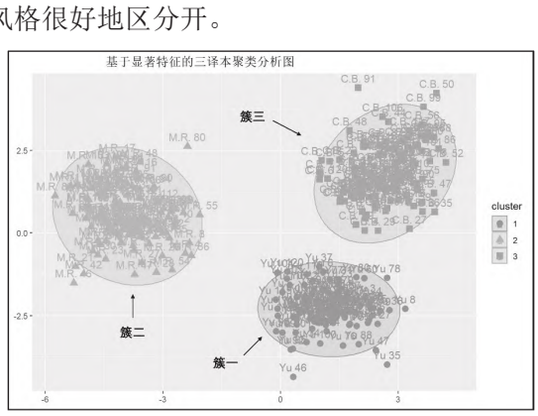


图 4　三译本基于显著特征的 K-means 聚类结果

随后我们借助特征选取后的 20 个显著特征，生成三个译本的聚类图（图 4），置信椭圆为 95%。如图 4 所示，所有样本明显可见共区分出三个簇，簇一中的样本表示为圆形，皆为虞苏美的译文样本；簇二样本表示为三角形，为罗慕士的译文样本；簇三样本表示为方块，为邓罗的译文样本。三个簇之间的距离较为清晰明显，说明三译本可通过 20 个显著特征来进一步地区分。换言之，基于这 20 个显著特征，就可以揭示不同译者翻译《三国演义》的风格差异，相反使用全部特征则无法很好地对三译本的特征作出区分和解释，这主要是由于部分特征可能包含噪声，抑或是有些特征在译本区分能力上较弱。从显著特征的表征上来讲，三个译文整体风格趋于一贯而终，即译文中各部分的风格活跃度较小，没有明显的风格偏移。

至此，机器学习算法中的分类和聚类分析结果，均足以证明提取出的 20 个显著特征可以有效区分《三国演义》三个译本的翻译风格，也能够说明机器学习算法可以完成较为困难的典籍文学翻译风格识别。下一部分将对这些特征在不同译文中的实例进行归纳总结。

四、特征分析与讨论

（一）篇章级特征

篇章级特征所反映的是不同译本在宏观层面表现出的差异，具体体现在符号比例、词汇比例等方面。

1. 符号比例

从表 4 的排名可以看出，缩写符比例与分号比例可以显著区分出三个译本，特别是缩写符属于是最能凸显三译本风格的显著特征。通常，研究者更多会着眼于词汇句法层面的特征，较少关注标点符号所起到的作用。在此次研究中，相较于邓译本和虞译本来说，罗译本在缩写符比例上有着显著的差异。由于关键词特征中的 's 和 n't 同样包含缩写符，所以我们也将这两个关键词特征一并纳入缩写符特征的考察。结合表 5 我们可以发现，在缩写符的使用上，虞译本中出现 4 784 个，邓译本中出现了 4 022 个，罗译本出现最多，共 7 638 个。同样体现在关键词中，表示物主关系成分（Cao Cao's army）或是第三人称单数完成时（He's gone）缩写的 's 形式与总体缩写符的分布类似，罗译本中大量出现。而 not 的缩写 n't, 在虞、罗两译本中出现数量相当，在邓译本中出现的非常少。

表 5　缩写符、's、n't 在译本中的数量和占比

	缩写符		’s		n’t	
	数量	比例	数量	每万字	数量	每万字
虞译本	4 784	0.79%	2 982	49.54	494	8.21
邓译本	4 022	0.67%	2 507	41.90	8	0.13
罗译本	7 638	1.38%	5 345	96.87	484	8.77

例（1）为平行语料库中抽取的实际语料，这段话是《三国演义》中比较经典的故事，体现出了曹操的残酷和野心，后面曹操便道出那句万古流传的话“宁教天下人负我，不教我负天下人”。从缩略符上来说，例（1）中短短两句话，罗译本中出现了八次。他善于应用缩略符起到表达语气的作用，如实例中使用缩略符来表现说话者讲的不是“官话”，符合说话者乡下人的身份，英语读者看到这种“不规范”的英语行文之后会自然而然地感受到这一点。然而此处实际上是译者根据自己的理解进行的补充，因为原文中并没有体现出由于说话人是乡下人，所以他们说不好“官话”。相反，邓译本和虞译本中出现较少，特别是邓译本的缩略符的使用尤为有限。

例（1）：

a. 操曰：“吕伯奢非吾至亲，此去可疑，当窃听之。”二人潜步入草堂后，但闻人语曰：“缚而杀之，何如？”操曰：“是矣！今若不先下手，必遭擒获。”（原文）

b. "There**'s** something suspicious about his leaving. Le**t's** look into this." The two men stole behind the cottage and overheard someone mumble, "Le**t's** ti**e'm** up a**n'** kil**l'm**." "I thought so," Cao Cao whispered. "If we do**n't** strike first, w**e'll** be caught."（罗译本）

c. "He is not my real uncle; I am beginning to doubt the meaning of his going off. Let us listen."

So they silently stepped out into a straw hut at the back. Presently someone said, "Bind before killing, eh?"

"As I thought;" said Cao Cao, "now unless we strike first, we shall be taken."（邓译本）

d. Cao Cao said, "H**e's** not my real uncle. I begin to doubt his reasons for going off. Le**t's** go inside and listen."

So they quietly stepped into the back of the cottage. Soon they heard someone inside saying, "Bind first, then kill, eh?"

"As I thought," said Cao Cao, "unless we strike first, we will be taken." （虞译本）

缩略符如 's 和 n't 一般使用在日常对话、交流访谈等偏口语化的非正式文体当中，相较其他两个译本，罗译本中富含较多的缩略符，似乎更偏向非正式的行文风格，从一方面来说，似乎有违三国原本偏文言正体的行文风格，且译者能动程度高，添加了更多自己的理解；但在另一方面，三国中同样富含较多的对话内容，人物纷繁复杂、特点性格各异。从这点来讲，罗译本能够灵活转达原文内容。与其说使用言风正式、文学性高的英文长难句来翻译，较口语的风格往往能够准确传达原文的对话内容。以往学界认为罗译本多偏异化[19]，但本研究的结果却表明，在具体的词汇和符号的使用上，罗译本译法似乎更偏归化。

在符号的使用上，分号比例被筛选为有力的显著特征比较出乎意料，鲜有文章会关注翻译风格中标点符号的运用。在三个译本当中，虞译本使用分号数量最少，共出现 551 次，在使用的所有符号当中占比 1.7%，罗译本出现 2 362 次，占比 3.77%，邓译本出现 2980 次，占比 4.92%。分号的使用差异究竟意味着什么？《三国演义》原文文体偏古风，分号较为鲜见，而英文当中分号使用比较频繁，起到了并列、停顿和分清层次的作用。虞苏美作为母语是汉语的译者，在一方面有意识或下意识地避免运用英语当中分号的表达习惯，转而运用逗号等其他符号替代了分号的作用；另一方面，较少使用分号也和原文的文体相似，表现出一定的异化特征。

2. 词汇比例

表 4 中包含三个词汇类型的显著特征，即实词比例、名词比例和 Be 动词比例，值得一提的是，本次研究中的实词以名词、动词、形容词、副词和数词为标准，具体数值如图 5 所示。

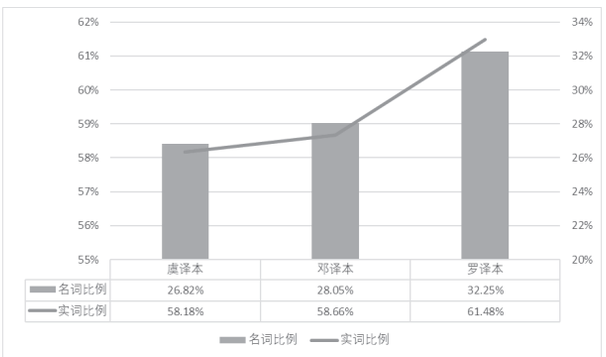


图 5 三译本中的名词比例和实词比例

如图 5 所示，罗译本在实词比例和名词比例上占比最高，分别为 61.48% 和 32.25%，名词属于实词的重要组成部分，实词的比例高实际上说明文章的词汇密度高、信息丰富度较大，同时这也说明罗译本能够适应原文较为文言化文体，可以译出不同的语言风格[11]。结合三个译本的总形符数，罗译本的正文总词数最少，我们也可以说罗译文用词简练灵活，和原文较为文言化的文体较为相近，同时“贴近原文的语言结构、行文方式，通过模仿汉语的形式，传递语言背后的民族文化传统、习俗风尚、审美心理、思维模式和精神气质[19]”。相反，虞译本的实词比例和名词比例最低，同时总形符数最高，这说明该译者多运用虚词来满足衔接、连贯等作用，翻译名词时也更为灵活，会综合运用意译、增译、转译或加注等翻译策略。究其原因，相较之罗、邓译本，虞苏美翻译过程中不仅重视原文当中富含的中国传统文化表达，同时在译文中增添虚词衔接、灵活转译名词，以达到扩展文化外延、行文流畅通顺、提高阅读体验的作用，力求英语读者能够准确识解原文的含义。因此有学者评价虞译本是“在精准传达《三国演义》原文内涵的基础上，也照顾了英语读者的阅读需求与期望”[24]。

另一个词汇特征是 Be 动词，统计结果显示，出现最多的是邓译本，共 20 865 次，每万字出现 348.73 次，虞译本出现频数略低，共 19 948 次，每万字出现 331.42 次，出现最少的是罗译本，只出现了 12 863 次，每万字出现 233.11 次。借助例（2）我们可以浅析 Be 动词的使用差异带来的文风区别。

例（2）：

a. 生得身长七尺五寸，两耳垂肩，双手过膝，目能自顾其耳。（原文）

b. He **was** tall of stature. His ears **were** long，the lobes touching his shoulders，and his hands hung down below his knees. His eyes **were** very big and prominent so that he could see backward past his ears.（邓译本）

c. He stood seven and a half spans tall，with arms that reached below his knees. His ear lobes **were** elongated，his eyes widely set and able to see his own ears.（罗译本）

d. In appearance，he **was** tall of stature. He had exceptionally long ears，the lobes touching his shoulders so that his eyes could see his own ears. His arms **were** long，too，with his hands hanging down below his knees.（虞译本）

例（2）是典型的描述文本，原文介绍的是刘备的外貌与众不同，较为奇特。在三个译文当中，邓译本使用了三次 Be 动词，分别描述出了刘备的身高、耳朵以及眼睛，并没有使用长难句或是从句来翻译原文，简单句式较多。罗译本读起来就更具有文学性，多为复杂句，也就导致了 Be 动词数量的减少，例子中只出现了一次。而虞译本则中规中矩，例子中使用了两次 Be 动词。同样的，Be 动词使用频率的高低可以说明两个方面：一是句式的简单或复杂，Be 动词出现频率高，则简单句较多，读起来流畅通顺，结构明了直白；二是词汇的丰富与否，Be 动词少的译本会借用其他的词汇来表达原文的意思，如例（2）中罗慕士使用 He stood seven and a half spans tall 来直观表现出刘备的身高，相比 he were tall 来说，在词汇上要更为丰富。

因此，邓译本 Be 动词较多，句式较为简单，主要由于该译本的受众主要是在华外籍人士，作为最早推出的英文全译本，起到了向外推介《三国演义》的作用，“用地道的英文传达原作者的话”，因此整体偏归化，所以有学者也认为“同 20 世纪 90 年代罗慕士的译本相比，邓罗译本较为简单，不但没有长篇的译者的话，没有附录，甚至注释都很少出现，而且没有严格忠实于原著，例如原著中的一些对话被改写成为文字叙述”[12]。而罗译本用词丰富、句式复杂，皆因罗慕士认为必须通过添加注释、编写评述，才能有效帮助西方读者顺畅理解原书的内容与情节[2]。这也使得罗译本成为诸多中国文化爱好者的案头书，他的译本也被称为是最好的《三国演义》英文译本[25]。

（二）关键词特征

经过特征选择产生的显著关键词特征可以有效区分不同译本之间的微观用词差异，通过观察关键词的在不同译本中的分布以及在具体语境的使用，能够在一定程度上体现出不同译者在行文时下意识的用词偏好。

表 6　关键词特征

	O		shall		wherefore	
	数量	每万字	数量	每万字	数量	每万字
虞译本	3	0.05	34	0.56	0	0.00
邓译本	266	4.45	624	10.43	0	0.00
罗译本	10	0.18	249	4.51	182	3.30

表 6 呈现的为三个关键词特征在三个译本中的统计数据，O、shall 和 wherefore 都是比较正式的词语，一般在非正式场合使用较少，因此，作为正式文体的代表，这三个词使用较多的译本，也同样体现出较为正式的文风。通过表 6，我们不难发现，不管是从数量上还是比例上，邓译本中三个关键词的使用上，都远远多于其他译本。在翻译可以吟诵的诗歌时，邓罗会选择使用正式诗歌中常用的古体词 O 来增加气势或达到押韵的效果；shall 作为情态动词的一种，其用法更为正式，在一些正式的文件中多用 shall 而不用 should，例（3）体现的便是邓译本中 shall 的大量使用；wherefore 则更是如此，罗、虞两个译本中均未出现，而邓罗使用了多达 183 次，甚至可以说，如果读者在阅读三国译本时，读到多个 wherefore，那么这本译本极大概率就是邓罗翻译的。

例（3）：

The Ruler **shall** live to the age of the eastern emperor，The dragon banner **shall** wave to the farthest limit. His glorious chariot **shall** be guided with perfect wisdom.

His thoughts **shall** reform all the world, Felicitous produce **shall** be abundant, And the people **shall** rest firm.

My desire is that these towers **shall** endure forever, And that joy **shall** never cease through all the ages.（邓译本）

邓译本中正式词汇，甚至是古体词汇较多的原因，或许跟他翻译的时代背景有关。首先，邓译本形成于20世纪初期，在翻译时或许会受到同时期英文小说创作的影响。如今的英语使用经历了科技发展与全球化，行文较为通俗、用词更加丰富，不再追求佶屈聱牙的大词、古词，而19世纪末、20世纪初的英语使用仍然以彼时的英式英语作为“标准体”，在行文写作中追求古典美，多运用拉丁语演变来的词，这个时代比较有代表性的小说有哈代的《德伯家的苔丝》，以及萨克雷的《名利场》等，因此邓罗在翻译同样富含历史和文化底蕴的《三国演义》时，会主观地向英语古典小说的词汇特征靠拢，同时客观上也会受到同时期英文文学作品的语言风格影响。

其次，根据邓罗传记记载[26]，邓罗于1857年出生在英国，接受过英国文学和诗歌的教育，后于1880年来到中国，在福州船政学堂教授课程，深深醉心于中国传统文化，于1888年开始尝试翻译《三国演义》。因此，掌握良好的英语文学功底，加之对中国文化的充分理解，这也在一定程度上解释了为什么邓罗会选择部分英文中的古体词来转译《三国演义》中的诗歌，此举既能准确传达中文诗歌的优美文体和浓厚底蕴，也能让英文读者阅读时产生熟悉感，不抗拒中国文化。

五、结论

本文借助机器学习中分类和聚类的方法，使用支持向量机、简单逻辑回归、决策树三个分类器和k-means聚类分析法对中国四大名著之一的《三国演义》的英文三译本进行翻译风格考察，从100个特征中挑选出20个显著性强的特征，并结合语篇对这些特征进行统计、分析和阐释。研究结果表明，显著特征足以有效区分三个英文译本，同时聚类分析也能够直观地将三译本分成界限鲜明的三个簇。研究发现，在符号比例上，罗译本中较多使用了缩写符，在翻译交流对话内容上更偏向非正式的行文风格；虞译本较少使用分号，贴合中文的表达形式，体现出一定的异化特征。在词汇层面，罗译本在实词比例和名词比例上占比最高，其译文的词汇密度高、信息丰富度较大，相反虞译本在译文当中增添了更多文化内容；Be动词的使用则间接体现出罗译本在词汇和句式上要更为丰富，而邓译本Be动词较多，其译文中富含简单句。最后的关键词特征验证了邓译本中存在古体词和正式词的大量使用，不同译本中较多或较少出现的关键词展现了译者的个人偏好，同样可以区分出不同译本的翻译风格。

本研究试图在翻译风格研究方法上进行创新，为中国典籍文学的翻译风格和译者风格研究提供了新思路和新方法。机器学习视阈无疑给翻译研究增添了科学性和客观性，这同“大数据时代‘系统性’‘整体性’和‘相关性’的理念”相辅相成[27]，从实践研究中探索语料库翻译学实现全面、深度、可持续的跨学科融合发展路径。

参考文献：

[1] 李鹏辉，高明乐 . 译者行为批评视域下19世纪英译群体行为研究——以《三国演义》为例 [J]. 外语学刊，2021（6）：55–60.

[2] 许多 . 试论罗慕士的文化立场与跨语际言说特质 [J]. 上海翻译，2020（6）：65-70.

[3] 郭瑶函，王东风 . “一带一路”背景下中国文化经典的传播与接受研究 [J]. 外语教学与研究，2021（2）：296-307.

[4] 胡开宝，谢丽欣 . 基于语料库的译者风格研究：内涵与路径 [J]. 中国翻译，2017（2）： 12-18.

[5] HERMANS T. The Translato's Voice in Translated Narrative[J]. Target，1996，01：23-48.

[6] BAKER M. Towards a Methodology for Investigating the Style of a Literary Translator[J]. International Journal of Translation Studies，2000，02：26.

[7] MIKHAILOV M，VILLIKKA M. Is there such a thing as a translator's style?[C]. Lancaster：Lancaster，UK.，2001：8.

[8] ILISEI I, INKPEN D. Translationese Traits in Romanian Newspapers： A Machine Learning Approach[J].

International Journal of Computational Linguistics and Applications，2011，02：319-332.

[9] LYNCH G , VOGEL C. The translator's visibility： Detecting translatorial fingerprints in contemporaneous parallel translations[J]. Computer Speech & Language，2018，06：79-104.

[10] 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（5）：80-85.

[11] 张浩然 .《三国演义》罗译本评析 [J]. 福建外语，2001（1）：49–54.

[12] 郭昱，罗选民 . 学术性翻译的典范——《三国演义》罗慕士译本的诞生与接受 [J]. 外语学刊，2015（1）：101-104.

[13] 贺显斌 . 从《三国演义》英译本看副文本对作品形象的建构 [J]. 上海翻译，2017（6）： 43-48.

[14] 彭文青 . 副文本视角下《三国演义》三个英文节译本研究 [J]. 明清小说研究，2021（2）： 240-250.

[15] 骆海辉，姜葵 .《三国演义》罗译本的骂词翻译研究——以目的论为观照 [J]. 漳州师范学院学报，2010（3）：116-121.

[16] 韩名利，陈德用 .《三国演义》译介模式与中国文学典籍之“走出去”[J]. 哈尔滨学院学报，2021（3）：115-118.

[17] 贺显斌 . 文化翻译策略归因新解——以《三国演义》Roberts 全译本为例 [J]. 天津外国语学院学报，2003（6）：1-6.

[18] 张晓光 . 基于语料库的《三国演义》罗慕士译本显化研究 [J]. 长春教育学院学报，2021（8）：17-25.

[19] 董琇 . 罗慕士英译《三国演义》风格之探析——以邓罗译本为对比参照 [J]. 中国翻译，2016（4）：93-99.

[20] 冉明志 .《三国演义》邓译本与罗译本战争军事术语英译策略对比研究 [J]. 译苑新谭，2021（1）：72-79.

[21] 张学工 . 关于统计学习理论与支持向量机 [J]. 自动化学报，2000（1）：36-46.

[22] 霍跃红 . 典籍英译译者文体分析与文本的译者识别 [D]. 大连：大连理工大学，2010.

[23] 刘颖 . 统计语言学 [M]. 北京：清华大学出版社，2014.

[24] 董雨晨 . 因果模型视角下的文学经典翻译——以《三国演义》虞苏美译本为例 [J]. 湖北第二师范学院学报，2018（3）：118-122.

[25] 文军，李培甲 . 国内《三国演义》英译研究：评述与建议 [J]. 北京第二外国语学院学报，2011（8）： 25-30.

[26] CANNON I. Public Success，Private Sorrow：The Life and Times of Charles Henry Brewitt-Taylor （1857—1938），China Customs Commissioner and Pioneer Translator[M]. Hong Kong：Hong Kong University Press，2009.

[27] 韩红建，蒋跃，袁小陆 . 大数据时代的语料库译者风格研究 [J]. 外语教学，2019（2）： 88–93.

[责任编辑：丁勇]

A ML-based Comparative Study on the Translation Styles of the Three English Versions of *Three Kingdoms*

KONG Delu

(School of Foreign Studies, Tongji University, Shanghai 200092, China)

Abstract: Classification and clustering methods in machine learning are used to investigate the translation style of the three English versions of *Three Kingdoms*. 20 salient features are chosen and analyzed. The research shows that the chosen features can effectively distinguish their stylistic differences with an average accuracy more than 95%. At the discourse level, each version shows its stylistic features on vocabulary and punctuation; in terms of keywords, the discrepancy in word frequency shows the translator's personal preference. This study aims to provide a practical path for the comparative study of translation style innovation.

Key words: Machine learning; translation style; *Three Kingdoms*