基于机器学习方法的《德伯家的苔丝》中文译本翻译风格考察

孔德璐

摘　要　研究使用机器学习中的分类和聚类方法，基于自建平行语料库，考察哈代名作《德伯家的苔丝》中文三译本的翻译风格。从68个全部特征中筛选出15个显著特征，并结合实例进行量性融合的阐释和总结。结果表明，显著特征能够有效区分三译本风格差异，分类、聚类实验的平均准确率均达到97%左右，提示出各译本在词汇、句法、语篇上的不同风格特征和译者的个人偏好。研究在为既往质性研究提供数据支持和细粒度分析的同时，也提出了一些纠正性结论，如张谷若译本词汇密度更大、被字句比例极少、成语比例差别不大等，并为翻译风格和译者风格研究方法提供了一定改进和补充。

关 键 词　机器学习；　翻译风格；　平行语料库；　《德伯家的苔丝》

分 类 号　H319.3

作者简介　孔德璐，同济大学外国语学院博士研究生，Email：kongdelu2009@hotmail.com。

0　引言

始于20世纪50年代的描写译学研究，打破了以往规约性翻译研究对“正统翻译”和“固定翻译范式”的限制，开始关注翻译的实际现状，并推崇基于实际语料进行实证研究，分析翻译文本中反复出现的模式及其成因。[1]伴随着描写译学和语料库语言学在90年代的融合，基于语料库的方法也逐渐在翻译研究中得到了具体运用。[2]贝克（Baker）[3]率先展开基于语料库的翻译研究可行性和现状讨论，并声称该方法可以揭示仅凭直觉和内省无法归纳的翻译规律。作为该领域热点之一的翻译风格研究，主要涉及译文语言模式挖掘和译文语言特征阐释，这一过程随着语料库语言学的发展成熟而变得更加高效。在此背景下，数字人文理念和方法得以融入翻译研究，使用先进的文本分析和可视化工具可进一步加深对翻译文本规律的洞察，本研究即为此类尝试之一——通过机器学习方法对《德伯家的苔丝》中文三译本进行翻译风格考察。研究利用分类和聚类实验从语言特征层面区分不同译本的风格，探讨翻译风格所体现出的译者个人偏好，为翻译风格和译者风格研究领域贡献新的视角和方法。

1 研究设计

1.1 文献回顾

赫尔曼斯(Hermans)认为译本中隐藏着“译者的声音”;[4]贝克认为译文中会不可避免地展现“译者的指纹”[5],其所提出的文学翻译风格考察的框架已被广泛用于基于语料库的翻译研究,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。在中文世界,徐欣基于多译本语料库分析《傲慢与偏见》的三个译本,从词汇、句法方面讨论不同译本的特有风格;[6]孔德璐和张继东采用定性和定量相结合的方法,从四字格的使用上对比考察了《德伯家的苔丝》三译本所呈现的翻译风格[7]。

批判性研究也同时在展开,如米哈伊洛夫(Mikhailov)和维利卡(Villikka)通过语料库方法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,如词汇丰富度、高频词等,并不能有效判断译者风格。[8]另外,在此类研究中,研究者往往会预先设定并过于强调某些特征,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,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”,因此伊利塞伊(Ilisei)等人批评这种方法“总是基于学者的自觉……显然是‘人工设计’的”[9]。鉴于此,研究者开始借助机器学习方法,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。该方法首先起于计量风格学(Stylometry),即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。如埃尔-菲奇(El-Fiqi)等人运用两两比对分类模型,成功地区分了阿拉伯语—英语和法语—英语译作中的译者风格,其表现要远优于传统的 C4.5 决策树模型;[10]林奇(Lynch)和沃格尔(Vogel)使用支持向量机模型,验证英语中的 N 元语法可有效识别译者风格[11]。在国内,詹菊红、蒋跃使用支持向量机模型成功地区分《傲慢与偏见》的两个中文译本,并对部分显著特征进行阐释。[12]

计量风格学研究固然弥补了基于语料库方法的一些缺陷,但也存在一些需要解决的问题。首先,在特征选择上,不同特征可能导致不同的研究结果,因此需要尽可能扩大研究所使用的特征集,以确定最适合特定文学译作分类任务的特征。其次,在模型的验证上,个别研究仅使用单一分类或聚类方法,缺乏对模型的验证,可能会导致结果偏差,最终影响研究结论的可靠性。本研究即以解决以上问题为初衷展开。

研究选取哈代经典之作《德伯家的苔丝》中译本翻译风格作为考察对象。《德伯家的苔丝》中译本众多,但针对翻译风格的定量研究较少,结合机器学习方法并结合具体特征来探讨译文风格的研究更是阙如。在既有研究中,对张谷若译本、王忠祥和聂珍钊合译本、吴笛译本翻译风格的讨论较为充分[13-16,37-39],其中不乏基于语料库的研究[14-16],且确实存在选取特征有交叠重复之处(形符比、平均句长、四字格、方言等)、部分隐性特征被忽略等问题,因而能为研究结论提供足够的比较空间,故本研究选取此三译本作为基础语料,并以解决以下两个问题为核心进行实验设计和结果分析:(1)机器学习方法如何更好地区分和验证不同译本间的风格差异;(2)如何改进特征阐释,以获得更具意义的结论。

1.2 研究路线

本研究技术路线如图 1 所示。就问题(1),针对前述人工预设特征、忽略隐形特征、特征容量有限的问题,

在参考既有研究方法的基础上,通过对实验材料的预处理和标注,尽可能多地扩展特征数量,构建较全面的多维特征集。随后采用有监督学习的文本分类实验来区别不同译本间的风格,同时为避免单一算法的偶然性[12],采用 SVM、简单逻辑回归和 C4.5 决策树作为分类器,互为参考。为了最大化研究效率和准确性,进一步采用卡方算法进行特征筛选,识别分类任务中贡献最大的"显著特征",并再次检验这些特征的分类效果。这种方法不仅精简了特征集,也确保显著特征属于区分三译本的"最优解"。偏重技术化的研究往往止步于呈现模型正确率,缺少不同特征影响实验结果的讨论[11],以上步骤同时解决了这个问题。最后为避免单一分类实验的偶然性,采用无监督学习的聚类分析方法来进一步验证分类实验结果,以增强研究结论的稳健性。

就问题(2),本研究将基于分类、聚类实验得到的结果,统计"显著特征",也即高贡献度语言特征在不同译本中的分布情况,并结合真实语料进行阐释说明,再参考和对比既有量化或质性研究的结论,对三译本的翻译风格进行考察。这一过程也同时检验了上一步的实验结果,体现了机器学习与深度学习"黑箱子"的不同,即机器学习方法所需要的特征集可用以检验分类效果。

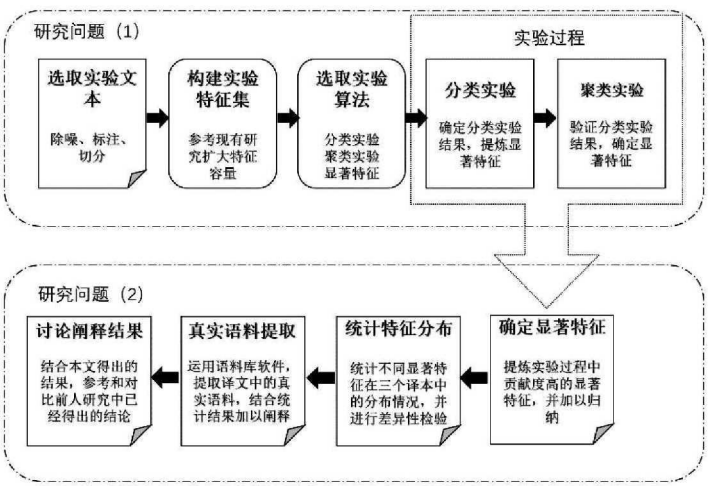


图 1 研究技术路线图

2 研究材料

2.1 语料处理

本研究选择《苔丝》原文以及三个不同译本作为研究语料,借助 Paraconc 软件实现句对齐,以便在质性分析部分检索并抽取语料进行分析。英文原文使用的是 2003 年版重印本[17];译本选取的是 1984 年人民文学出版社出版的张谷若译本(以下简称张译本),2011 年长江文艺出版社出版的王忠祥和聂珍钊的合译本(以下简称王聂译本),及 2016 年上海文艺出版社出版的吴笛译本(以下简称吴译本)。三译本中,张谷若译本创作于 1930 年前后,是在中国传统文化背景下对《苔丝》进行的早期翻译尝试,而王聂合译本及吴笛译本则代表了更现代的翻译视角和技巧。选用三译本的理由如下:(1)三个译本能够实现最佳的实验可操作性,仅选择两个译本则可对比的空间有限,三个以上则数据量、计算量、分析量大增;(2)这三个译本影响力较大,好评度较高,读

者接受度较好,具比较价值①;(3)如前所述,学界针对这三个译本的讨论较多,已有不少研究成果,可为本研究的结论提供讨论空间。

在语料处理阶段,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,并分别用 TreeTagger 和 NLPIR-ICTCLAS 汉语分词系统②对英汉语语料进行分词和标注。我们按照固定规模标准,即 UTF-8 编码下 30KB 的文件大小,切分三个汉语译文样本。固定文件大小的切分方式有助于样本容量的基本统一,为 5000—6000 词。最终建成的英汉双语平行语料库概况如表 1 所示,实验样本共计 112 个。

表 1 **语料详情**

作品	样本	形符数	类符数
《苔丝》源本	—	153233	12133
张译本	39	194385	12525
王聂译本	36	184926	11125
吴译本	37	174041	12449

2.2 建立特征集

参照以往研究[11-12,18],本研究所采集的特征主要分为两个大类。第一类为篇章级特征(Document-level),包括从词汇、句法到篇章的三个层次共计 33 个具体特征,其中新增《中国成语大辞典》[19]作为自定义用户词典,旨在考察文本中成语比例和成语密度差异,同时在用户词典中收纳常见方言词,构成自建方言词库,以揭示不同译本中的方言表达是否具有显著性差异。第二类为关键词特征,使用 Antconc 3.5.8,将三个译本逐个与“兰卡斯特现代汉语语料库”(LCMC)中的文学子库(LCMC Fiction)作对比,以提取各译本的关键词表,然后选取每个词表中关键值最高的 20 个主题词,并人工剔除与小说情节相关的专有名词,包括人名(“苔丝”“克莱尔”等)、地名(“沙氏屯”“马勒村”等),总词表去重后得到 35 个关键词。关键词特征以比例形式呈现,以确保其不受样本长度所影响,即某个关键词特征=该关键词频次/总词数,以体现该关键词在译本中的分布情况。

两个分类相辅相成,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维度来考察各译本的特点;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,体现译者的个人用词偏好。最终选取的 68 个实验特征如表 2 所示。借助自编 Python 代码和语料库检索工具 WordSmith 6.0,我们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、总成语数等所需要的特征数据。

① 通过豆瓣读书网对所选译本进行检索,张译本评分最高(8.8 分),五星评分比例最大(46.1%),且影响受众最广(所有版本共有超过 15000 人评价);王聂译本(评分 8.4 分,五星评分比 39.7%)和吴译本相差不大(评分 8.2,五星评分比 36.1%),统计时间为 2023 年 12 月。

② NLPIR-ICTCLAS 汉语分词系统是张华平博士负责开发的中文分词和词汇标注平台,详情参考:http://ictclas.nlpir.org/。

表 2 实验所使用的特征集

篇章级特征			关键词
(1)形符比	(12)语气词比例	(23)顿号比例	她、什么、俺、时候、把、那儿、怎么、不过、那么、但是、因为、所以、那、儿、从前、哪、这种、她们、都、他们、的、奶牛、在、没有、就、已经、他、这儿、为什么、牧师、一样、马车、这么、牛奶、觉得
(2)标准形符比	(13)实词比例	(24)省略号比例	
(3)平均词长	(14)实词密度	(25)括号比例	
(4)名词比例	(15)虚词比例	(26)引号比例	
(5)动词比例	(16)虚词密度	(27)把字句比例	
(6)形容词比例	(17)平均句长	(28)被字句比例	
(7)副词比例	(18)陈述句比例	(29)成语比例	
(8)介词比例	(19)问句比例	(30)成语密度	
(9)代词比例	(20)感叹句比例	(31)方言词比例	
(10)数词比例	(21)逗号比例	(32)连词比例	
(11)叹词比例	(22)分号比例	(33)助词比例	

注:①词类比例=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=句型频次/总句数,以句号、问号、叹号、分号、冒号作为判断句子结尾的标准。③符号比例=符号频次/总符号数。④本研究将名词、动词、形容词和副词定义为实词,这些词类均表达稳定词义;将介词、连词、助词、语气词和叹词这些不具备稳定词义或意义模糊而主要起语法功能作用的词语定义为虚词。则实词比例=总实词数/总词数,实词密度=总实词数/总虚词数;虚词比例=总虚词数/总词数,虚词密度=总虚词数/总实词数。⑤成语比例=总成语频数/总词数,成语密度=总成语频数/总句数。

3 实验及验证

3.1 分类器与算法

支持向量机(Support Vector Machine, SVM)在 1963 年由著名的苏联数学家弗拉基米尔·瓦普尼克(Vladimir Vapnik)等人设计,并在随后的广泛应用中展示出良好的性能[20]。在译者和译作风格研究中,SVM 模型执行分类任务效率较高、效果较好,受到国内外学者的关注和推广[11-12]。但单独使用 SVM 算法进行分类实验,缺少其他算法对照,结果不免存在一定的偶然性,因此本研究将 112 个文本样本转换为对应的基于 68 个特征的向量模型,使用 SVM、简单逻辑回归(Simple Logistic Regression)、C4.5 决策树作为分类器。

本研究使用了台湾大学林智仁教授等开发设计的便捷软件包 LIBSVM[22]。SVM 包含三个关键参数:(1)核类型(Kernel Type):线性核(Linear)意味着构建一个简单、高效且易于解释的 SVM 模型;(2)惩罚系数(cost):控制错误分类的惩罚力度,设置为 1000 能够达到最佳分类效果,同时避免过拟合;(3)伽马值(gamma):通过 GridSearch 算法调参,对(cost, gamma)参数组合进行系统遍历和评估,选择在交叉验证数据集上表现最好的参数组合,即对应伽马值为 0。最终,LIBSVM 的参数设置如图 2 所示,其他分类器均使用默认设置。

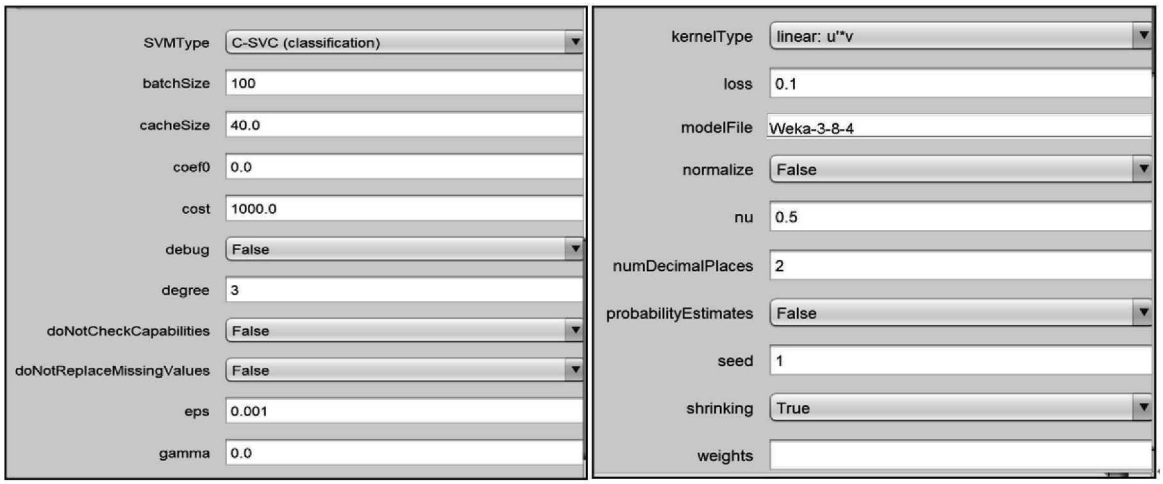


图 2　LIBSVM 参数设置

为选取分类过程中贡献最大的特征，本研究利用机器学习平台 Weka 3.8.4① 中内置的特征选择分类器（Attribute Selection Classifier），使用卡方评估法，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，从前一步结果中挑选出 15 个卡方值最高的参数，即 15 个显著特征。

3.2　分类实验结果

我们首先将预处理后的数据（68 个特征）导入三个分类器，使用十折交叉验证法进行分类模型精度评估②。接着利用卡方算法对特征进行筛选，获取 15 个显著特征（表 3），重复以上实验步骤。两次实验结果对比如表 4 所示。

实验结果包含三个主要指标：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；AUC，用来反映模型预测能力。召回率和 AUC 越接近 1，分类结果和模型预测能力越好。表 3 显示，用全部 68 个特征构建的分类器模型已经达到较为理想的结果，三个分类器准确率均高于 95%，其中以简单逻辑回归分类器表现最佳。取 15 个显著特征后，SVM 和简单逻辑回归分类器的三项指标均进一步提升。可见，基于显著特征进行分类是可行的，模型效果甚至更好，这种做法不仅可以提高数据处理效率，而且可在一定程度上避免部分噪音特征的影响，在特征阐释阶段也便于找到切入口和处理依据。

图 3 展示了基于 15 个显著特征的 SVM 分类器结果，在 112 个总样本中，共有 109 个样本被正确分类。底部的混淆矩阵反映出 3 个错分样本的具体情况，有两个张译文样本被错分入了王聂译本组，一个王聂译本样本被错分入张译本组，吴译本所有样本都被正确分类，且没有其他译者样本错误分入该类。这证明了 15 个显著特征是可以满足译本风格区分需求的。

① “Weka”是新西兰怀卡托大学开发的一款免费开源的机器学习和数据挖掘软件，全称为“怀卡托智能分析环境”（Waikato Environment for Knowledge Analysis），详情参考：https://www.cs.waikato.ac.nz/ml/weka/。

② K 折交叉验证即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。一般将 K 取值为 10，就是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。

表 3 **15 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
113.378±11.644	1	儿	72.659±2.974	9	都
100.80±0.4	2	分号比例	71.753±4.597	10	实词密度
99.600±3.872	3	哪	71.753±4.597	11	虚词密度
97.007±1.387	4	被字句比例	72.076±7.605	12	在
91.519±7.91	5	这么	69.334±3.25	13	助词比例
85.637±1.974	6	虚词比例	69.782±2.323	14	从前
86.296±10.901	7	没有	68.163±1.489	15	方言词比例
79.599±3.681	8	副词比例			

注:表中斜体字属于篇章级特征,其他为关键词特征。

表 4 **显著特征选择前后分类结果**

分类器	全部 68 个特征			15 个显著特征		
	准确率	召回率	AUC	准确率	召回率	AUC
SVM	95.5357%	0.955	0.967	97.3214%	0.973	0.980
简单逻辑回归	97.3214%	0.973	0.998	98.2143%	0.982	0.999
C4.5 决策树	96.4286%	0.964	0.973	96.4286%	0.964	0.973

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

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

Correctly Classified Instances         109               97.3214 %
Incorrectly Classified Instances         3                2.6786 %
Kappa statistic                          0.9598
Mean absolute error                      0.0179
Root mean squared error                  0.1336
Relative absolute error                  4.0175 %
Root relative squared error             28.3365 %
Total Number of Instances              112

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.949     0.014     0.974      0.949     0.961      0.941     0.968     0.942     TessZhang
                 0.973     0.027     0.947      0.973     0.960      0.940     0.973     0.931     TessWang
                 1.000     0.000     1.000      1.000     1.000      1.000     1.000     1.000     TessWu
Weighted Avg.    0.973     0.014     0.973      0.973     0.973      0.960     0.980     0.957

=== Confusion Matrix ===

  a  b  c   <-- classified as
 37  2  0 |  a = TessZhang
  1 36  0 |  b = TessWang
  0  0 36 |  c = TessWu
```

图 3 基于显著特征的 SVM 分类器结果

3.3 聚类分析验证

前文述及，仅通过分类实验来区分译本翻译风格而不加验证，其结果不足令人信服，因此本研究运用聚类分析进一步检验分类结果的信度和效度。聚类分析（Clustering 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[24]。本研究借助 15 个显著特征，利用 R 语言中的 k-means 函数，得到 112 个样本的聚类结果图（图 4）。

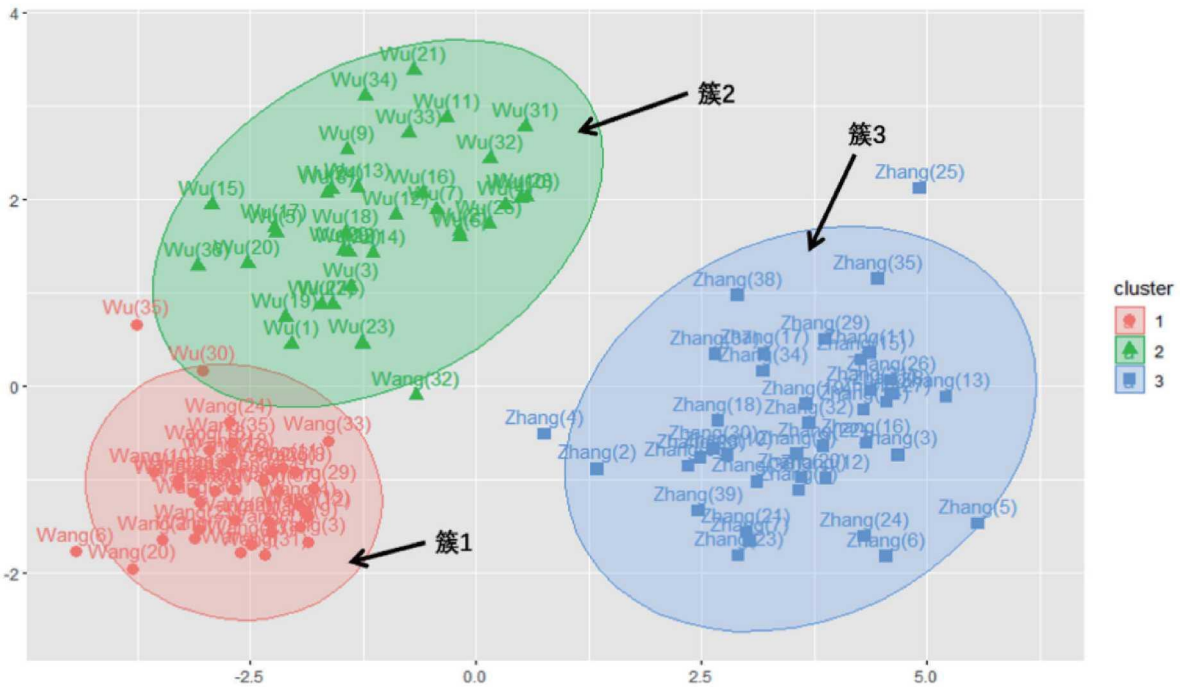


图 4　三译本基于显著特征的 k-means 聚类结果

注：圆形（红色）代表簇 **1**，三角形（绿色）代表簇 **2**，正方形（蓝色）代表簇 **3**；聚类各簇外围的椭圆表示 **95%** 的置信区间。

图 4 呈现显著可辨的 3 个簇，簇 1 的样本表示为红色圆点，主体为王聂译本样本；簇 2 的样本表示为绿色三角形，主体为吴译本样本；簇 3 的样本表示为蓝色正方形，均为张译本样本。簇 1、簇 2 外围椭圆（即 95% 的置信区间）的相交处及附近出现了被错分的样本——Wu（35）、Wu（30）和 Wang（32），张谷若的译文样本全部分簇正确，也即 112 个样本中只有三个样本被错误分组，基于 15 个显著特征的 k-means 聚类正确率为 97.32%。

至此，无论是机器学习中的文本分类还是聚类分析，均证明在广泛建立特征集基础上提取显著特征的方法可以有效区分译本的翻译风格。

4　基于显著特征的翻译风格分析

实验完成后，我们基于 15 个显著特征，结合平行语料库中统计得到的数据以及具体语料，从篇章级特征

和关键词特征两个层次对三译本进行翻译风格分析。

4.1 篇章级特征

4.1.1 词汇特征

在8个篇章级显著特征中,词汇层面特征有虚词比例、实词密度、虚词密度和副词,这首先意味着实词和虚词的使用特征对于区分三个译本有重要贡献。张译本虚词密度为0.32、虚词比例为19.2%、实词密度为3.08,为方便分析,我们统计了实词比例,为59.09%;王聂译本对应四个数值为0.38、21.65%、2.65、57.47%;吴译本为0.37、21.34%、2.72、58.01%。参考王克非[25]63 对英译汉文学文本中各类词的分布统计(本文对于实词和虚词的划分与该研究一致),可以发现三译本虚词比例均高于原创汉语的平均值(13.04%);实词比例均高于文学翻译汉语的平均值(57.3%),但未能达到原创汉语的程度(61.4%)。王克非和胡显耀认为虚词和实词的占比分别反映了翻译语篇中"虚词显化"和"词汇密度"。[26]三个译本的虚词比例均高于原创汉语,说明均受到英文原文的影响,体现出翻译汉语在语法标记上的外显(形合)特征。学界普遍认为词汇密度是衡量信息量大小和文本难度的重要指标之一[25]82-85:实词属于开放区间,不断有新词产生,而虚词属于闭合区间,虚词数量有限。因此,实词密度高和虚词密度低代表词汇信息度相对较大,且文本难易度较高[26]。三译本中,张译本实词密度最高、虚词密度最低,可以说其词汇密度最大,文本信息更为丰富,且更接近原创汉语文本;与之相反,王聂译本实词比最低、虚词比最高,说明其受到原语的影响相对较大,更突显出翻译语言"显化"特点。

针对副词,《苔丝》源本的副词比例为7.22%,张译本为10.44%,王聂译本为9.19%,吴译本为9.7%。表5单因素方差分析的数据显示,源本中使用副词比例要远低于中文译文,p值均小于0.05;三译本中又以张译本副词比例最高,且和其他两译本均呈现显著差异($p<0.000$);王聂译本和吴译本在副词使用上大同小异,无法检验出差异性($p=0.224$)。由于汉语是意合的语言,而且缺乏在英语中使用较广泛的限定词和不同的动词时态、体态等,所以在翻译时需要使用大量副词来填补空缺。同时,汉语中的副词是语篇衔接的重要手段之一[32],且汉语副词主观性强,能表达说话者丰富的主观态度或感情[34]。这是《苔丝》源本中使用副词比例要远低于中文译文的原因。王克非[25]63 提出汉语原创文学中副词的占比为10.42%,三译本中只有张译本的副词比例达到原创汉语的水平,这意味着张译本在人物动作和景物描写上更加生动形象,文本可读性更强。因此,副词比例可被视为将张译本与其他两译本区分开的重要词汇特征。并且综合虚实词特征来看,张译本都更接近原创汉语。

表5 **副词差异性检验**

差异检验	源本/张译本	源本/王聂译本	源本/吴译本	张译本/王聂译本	张译本/吴译本	王聂译本/吴译本
p值	.000**	.044*	.000**	.000**	.000**	.224

注:单星号上标表示具有差异性,双星号上标表示差异十分显著,未标星号表示在统计意义上不具有差异。

4.1.2 句法特征

分号比例和被字句比例在表2中分列第2和第4位,说明是两个能够有力区分三译本翻译风格的特征。分号比例被筛选为有力的显著特征比较出乎意料,鲜有研究会关注翻译中标点符号的运用。英文源本共使用了1386个分号,张译本中有1405个,比例为3.71%,数量甚至超过了源本;王聂译本中出现了1198个分号,比例为3.65%,虽然不及英文原文和张译本,但相较于吴译本(136个,比例为0.39%)仍显现出英文源本的强烈影响。观察例(A),英文原文中共出现5次分号,起到并列作用,对此吴译本全部使用逗号来翻译,体现出译者不拘泥于原文,在标点的使用上发挥更多的"译者主体"作用。在这个例子中,张译本使用了2个分号、王聂译本使用了4个分号(囿于篇幅未将原文列出)。可见标点符号非常能够凸显译者的个人选择,后续研究应当更多关注这一较为隐形的特征。

例(A)

(1)A gaunt four-post bedstead which stood in the room afforded sitting-space for several persons gathered round three of its sides;a couple more men had elevated themselves on a chest of drawers ;another rested on the oak-carved" cwoffer ";two on the wash-stand;another on the stool;and thus all were,somehow,seated at their ease.

(2)房间里,一张破旧的四柱床充当了座位,好几个人聚集在床铺的三面,还有两个人高高地坐在五斗橱上,另有一人坐在橡木雕花的箱子上,另有两人坐在盆架上,另有一个坐在板凳上,于是,所有的人好歹都舒舒服服地有了坐处。(吴译本)

"被"字句是另一个可显著区分三译本的句法特征。张译本中"被"字句平均每百万词出现41次,王聂译本每百万词出现1801次,吴译本每百万词出现1539次,显然,相较于张译本,王聂译本和吴译本受到英语中高频使用被动语态的"源语干扰"[27]较强影响。例(B)集中体现了多个被动语态连用下张译本中的翻译策略。英文原文包含5个被动语态,多个被动从句层层嵌套,而张译本只使用了一个半语法化的显性被动标记——"叫",同时诉诸句法上的处理,如调整句子主、被动结构,或通过重复专有名词("犏牛"和"牛犊"),完成了将被动语态转变为主动语态。

例(B)

(1)Long thatched sheds stretched round the enclosure,their slopes encrusted with vivid green moss,and their eaves supported by wooden posts rubbed to a glossy smoothness by the flanks of infinite cows and calves of bygone years,now passed to an oblivion almost inconceivable in its profundity.

(2)院子四围草棚长列,坡着的棚顶上,长着一层鲜明的绿苔,前檐都有多年以来叫无数的犏牛和牛犊用肚子摩擦得光滑发亮的木头柱子支着;那些犏牛和牛犊如今好像是坠入了深得不可思议的遗忘之渊里去了。(张译本)

胡显耀和曾佳[28]指出,"兰开斯特现代汉语语料库"(LCMC)文学子库中平均每百万词出现1553次被动式(x),当代汉语翻译小说语料库中平均为1249次(y),二者比值(y/x)为80.4%,即在文学文本中,翻译汉语所使用的被动式比原创汉语约少20%。他们将这一现象归纳为"传统化"(conventionalization),意指翻译文本趋从或夸大译入语传统,以提高译本的可接受性,反映在"被"字句上就是翻译文本夸大汉语传统,使用"被"

字句比原文文学更少。若以此视角来看,王、吴译本“被”字句比例均高过翻译文本,比较接近于原创文本,“传统化”并不明显。但张译本中每百万词被动式仅41次(z),和汉语原创文学该比例的比值(z/x)为2.64%,即相比汉语原创文学“被”字句竟少97%,显然超出了“传统化”的内涵,反映的是张谷若的个人翻译风格:善于利用句法转换、名词复现等其他方式来翻译原文被动式。

4.1.3 篇章特征

方言词比例是区分三译本翻译风格的显著篇章特征。在文学作品中,方言的功能多种多样,如刻画人物形象、体现乡土风情,等等[39]。我们将方言词划为篇章特征的主要依据是该类词所带来的篇章效果,里奇(Leech)和绍特(Short)将方言的使用归纳为“对作家使用语言的远离,因此同样远离小说的主要判断标准”[29]。由此,方言表达有助于带来不同的篇章效果,建立特殊的篇章含义。在三个译本中,张译本使用了最多的方言词,每万词平均出现40.52个方言词,吴译本中则每万词出现16.23个,而王聂译本自始至终没有使用方言。

称呼语是方言词的一大突出表现形式,我们以第一人称单数称呼语“我”和其方言变体“俺”为例,验证了三个译本之间方言词的使用差异。如表6所示,英文中第一人称数目总数比三个中文译本都少,这反映出英文小说中常常省略代词的现象,而汉语译文通常需要增加代词来确保行文流畅和通顺;使用方言词“俺”占比最高的是张译本,吴译本其次。以上统计也证明了以方言词区分三个译本的有效性。

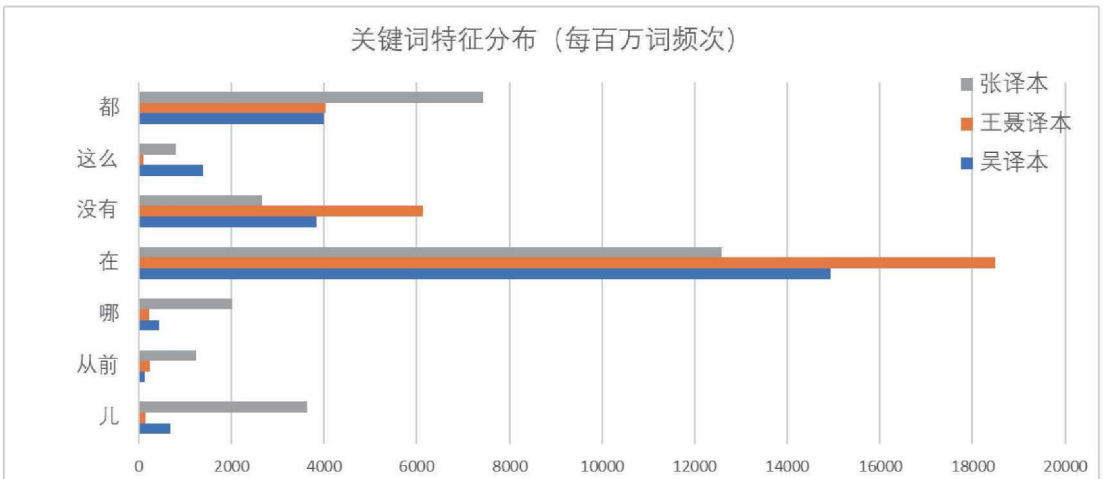
表6 **方言称呼语差异对比**

称呼语 文本	方言变体“俺” 数量(个)及占比	标准语“我/I/me” 数量(个)及占比	总计(个)
《苔丝》源本	-	2049(100%)	2049
张译本	496(15.8%)	2644(80.24%)	3140
王聂译本	-	3042(100%)	3042
吴译本	280(9.8%)	2577(90.20%)	2857

4.2 关键词特征

15个显著特征中有7个关键词特征,分别为代词“这么”“哪”、介词“在”、时间指示词“从前”、副词“没有”“都”和词缀“儿”。这些词类实际意义较少,在语篇中多起到补充衔接的作用。

由图5可知,整体而言,张译本中的“都”“哪”“从前”“儿”使用频率明显高于其他两个译本,王聂译本中“在”和“没有”使用频率突出,吴译本中“这么”一词使用频率十分之高。这些用词差异反映了译者下意识的个人偏好,并和宏观篇章级特征具有一定的联系。如在“儿”一词的使用上,张译本中大量使用了儿化词,如“小房儿”“今儿”“趁早儿”等等,儿化词是北方方言的特征之一[31],佐证了张译本偏好使用方言词。这也说明在三译本当中,张译本更富有乡土气息。



	儿	从前	哪	在	没有	这么	都
张译本	3640.77	1240	2010.51	12581.28	2653.33	790.77	7432.05
王聂译本	136.22	242.97	228.65	18498.11	6136.76	91.62	4036.22
吴译本	684.17	130.56	450.56	14945.28	3842.22	1371.67	3999.72

图 5　关键词特征分布情况

5　讨论

本研究从词汇、句法、篇章三个层面构建语言特征集，运用机器学习算法对《苔丝》张谷若译本、王忠祥和聂珍钊合译本、吴笛译本中文三译本翻译风格进行考察，以下为考察结论和方法总结。

5.1　与既往研究结论的对比探讨

首先，本文部分结论佐证了既有研究。统计数据证明张译文和中文原创文学语料相似，说明其在翻译时更加偏向中文原文行文习惯，避免受到过多的“源语影响”，迎合汉语读者阅读习惯，符合其翻译生涯中一以贯之的准则——努力打造“地道的译文”[33]。在副词的使用上，张译本中的副词数量较多，也从数据上支持了其译本在人物动作和景物描写上更加生动形象的说法[7,15]。同时在方言词上，张译本使用了大量的方言俗语，形成了其独特“方言对译”的翻译风格，也间接佐证前人的研究[39]；王聂译本自始至终都选择标准语来翻译，没有使用任何的方言词。

其次，本文部分结论补充了既有研究。现有文献往往忽视了某些标点符号和特殊句式对于区分不同译本翻译风格的作用。在标点符号方面，本研究实验结果揭示分号比例是三个译本之间的重要区别。通过进一步对源本和译本的对照考察，发现吴笛译本多使用逗号来转译分号，使用分号较少成为其译本风格特点之一，体现出该译者的创造性以及主体性。这一点是既有研究没有注意到的。同时在“被”字句比例上，通过更加深入的研究，我们发现张谷若译文中被动句的超低比例并没有遵循“被字句出现频率相较原创汉语文本更高”[27]的结论，也远远超越文学翻译汉语中的“传统化”[28]范畴，这反映出译者鲜明的个人译语偏好，形成其独一无

二的译者风格。

最后,本文部分结论更新了既有研究。现有成果基于个别章节,称张谷若译本的词汇密度最低[14],但本文的研究语料为全译本,是对《苔丝》多译本的全面考察,发现张谷若译本的实词密度在三个译本中是最高的,且虚词密度最低,所以张译本词汇密度最大,文本信息更为丰富,且更接近原创汉语文本。在这一点上,本文结论和既往研究有所出入。另外,既往研究认为张谷若的译文"最突出的特点是偏爱使用四字结构和成语"[16]。就固定成语①的使用上来看,本研究采用基于《中国成语大辞典》的自定义用户词典对三译本进行考察,结果发现和成语相关的两个参数(成语比例和成语密度)均未能进入 15 个显著特征当中,分列第 27 和第 40 位。单独采用这两个参数对三译本进行分类,SVM 准确率只有 50.89%,即仅一半的样本被正确分类,说明固定成语难以区分三个译本的翻译风格。这证明,就固定四字成语的使用上来看,张译本和其他译本差别不大。

5.2 对既有研究方法的补充

前文述及,基于语料库的翻译风格研究能够在海量数据的观察和分析中发现规律性特征,从而为传统的质性研究提供大量数据,可以证实、补充或推翻已有结论[35]。然而,此类研究仍存在人工预设特征、忽略隐形特征、特征容量有限等问题。本研究结合语料库研究以及计量文体学研究的方法,在实验设计上尽可能弥补上述缺陷。

在特征挖掘方面,本研究对文本进行特征提取和语言参数区分度计算,以发现前人研究中可能被忽略或遗漏的文本特征。我们建立了涵盖词汇、句法、篇章三个层面共 68 个语言特征的特征集,其中含既有研究未使用过的成语比例、方言词比例、关键词比例等特征。随后运用卡方算法提取出 15 个在分类中具有高贡献度的显著特征,并证明了其在分类和阐述上的有效性和实用性,这是本研究在方法上又一独到之处。而且,大部分基于语料库的翻译风格研究所使用的贝克[5]框架中的风格参数 TTR、STTR,并未入选该 15 个显著特征,说明这两个参数对于本研究的文本对象区分能力并不优越。这再次凸显了本研究的提示意义:不加筛选便使用前人研究框架中预先设定好的特征是不妥当的。

在验证阶段,我们采用机器学习中的聚类实验来验证分类实验结果,避免了使用单一分类实验造成的偶然性,并增加研究结论的科学性和可靠性。本研究实验过程可重复、模型可验证,具有较好的科学性[12],进一步证实了数据驱动的机器学习方法在翻译风格研究领域的适用和有效。

5.3 局限性与未来展望

首先,本研究只选取了三个译本,而如今市面上《苔丝》重译本屡屡问世,将学界讨论较多的其他译本,如孙法理译本等纳入,是未来研究方向之一。

其次,本研究通过量化分析对文本进行"远读",结论得到数据支持,但因为篇幅限制,部分语言特征未经深入挖掘,同时在质性探讨上还有所欠缺,较难对个别论点展开详尽论述。例如张译本、吴译本、王聂译本的

① 根据方梦之先生定义[42],四字格(或四字结构)是"由四个汉字组成的词组,包括大量成语和自由组合的词组"。由于松散的四字结构较难自动提取,此处仅讨论属于"固定成语"的四字结构。

某些区别特征形成的背后原因有哪些？译本出现年份的不同是否会对译文的语言特征产生影响？这些问题待后续研究讨论。

最后，本研究所采用的实验特征局限于词性标注后的字词层面，囿于技术限制，更为复杂的语义、语篇层面未能充分涉及。因此，后续研究将从特征选择入手，借助文本可读性等指标，深挖语义、语用、语篇等方面的文本特征，在数字人文领域进一步丰富翻译风格的研究方法和深度。

6 结论

本研究借助机器学习中分类和聚类的方法对哈代名篇《德伯家的苔丝》三个中文译本——张谷若译本、王忠祥和聂珍钊合译本、吴笛译本进行翻译风格考察。具体而言，即在建立涵盖词汇、句法、篇章三个层面共68个语言特征的特征集后挑选出15个显著特征，使用SVM、简单逻辑回归、决策树三个分类器和k-means聚类分析法对三个译本进行基于特征的分类和验证，最后结合文本对这些特征进行分析和阐释。实验结果显示，此15个显著特征足以有效区分三个中文译本，分类器平均准确率达到97%。研究发现，在篇章层面，张译本包含更多的副词和方言词，极少使用“被”字句，体现出译者遣词造句的灵活性，其译文中实虚词比例也更接近汉语原创文本；王聂译本和吴译本在词汇层面差距较小，实虚词和副词使用比例较为接近；吴译本在分号的使用上富有创造性，试图摆脱源语干扰；而王聂译本中不使用“方言对译法”是其主要的特征之一，即选择标准语来对应翻译英文中富有乡土气息的部分。在关键词层面，不同译本都有使用频率特别突出的关键词，展现出细微的译者个人偏好。

总之，本研究在为既往质性研究提供数据支持和细粒度分析的同时，也提出一些纠正性的结论，并在方法论上进行了一定创新。机器学习方法的应用无疑给翻译研究增添了科学性和客观性，研究过程体现了“大数据时代‘系统性’‘整体性’和‘相关性’的理念”[36]，提示了探索翻译学在数字人文背景下实现全面、深度、可持续的跨学科融合的发展路径。

参考文献

[1]胡开宝. 语料库翻译学：内涵与意义[J]. 外国语(上海外国语大学学报)，2012(5)：59-70.

[2]杨惠中. 语料库语言学导论[M]. 上海：上海外语教育出版社，2002.

[3]BAKER M. Corpora in translation studies：an overview and some suggestions for future research[J]. Target：international journal of translation studies，1995(7)：223-243.

[4]HERMANS T. The translator's voice in translated narrative[J]. Target：international journal of translation studies，1996(1)：23-48.

[5]BAKER M. Towards a methodology for investigating the style of a literary translator[J]. Target：international journal of translation studies，2000(12)：241-266.

[6]徐欣. 基于多译本语料库的译文对比研究——对《傲慢与偏见》三译本的对比分析[J]. 外国语(上海外国语大学学报)，2010(2)：53-59.

[7]孔德璐，张继东. 基于平行语料库的 *Tess of the D'Urbervilles* 三译本四字格的对比研究[J]. 翻译研究与教学，2022

(2):105-113.

[8]MIKHAILOV M, VILLIKKA M. Is there such a thing as a translator's style[C]//Proceedings of the corpus linguistics. Lancaster, 2001:378-385.

[9]ILISEI I, INKPEN D. Translationese traits in Romanian newspapers: a machine learning approach[J]. International journal of computational linguistics and applications, 2011(2):1-12.

[10]EL-FIQI H, PETRAKI E, ABBASS H A. Pairwise comparative classification for translator stylometric analysis[J]. ACM Transactions on Asian and low-resource language information processing, 2016(16):1-26.

[11]LYNCH G, VOGEL C. The translator's visibility: detecting translatorial fingerprints in contemporaneous parallel translations[J]. Computer speech & language, 2018(52):79-104.

[12]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.

[13]韩子满. 试论方言对译的局限性——以张谷若先生译《德伯家的苔丝》为例[J]. 解放军外国语学院学报,2002(4):86-90.

[14]王蓉. 基于语料的英汉翻译语言风格对比研究——以《苔丝》三译本为例[J]. 外文研究,2018(2):84-89.

[15]吴逾倩,赵文通. 以张谷若译《苔丝》为例的译者风格研究[J]. 北京建筑工程学院学报,2011(1):76-80.

[16]张乐金,徐剑.《苔丝》两译本的译者风格对比研究[J]. 江苏师范大学学报(哲学社会科学版),2013(5):69-73.

[17]HARDY T. Tess of the D'Urbervilles[M]. Chicago: World Book Inc, 2003.

[18]黄伟,刘海涛. 汉语语体的计量特征在文本聚类中的应用[J]. 计算机工程与应用,2009(29):25-27.

[19]王剑引. 中国成语大辞典[M]. 上海:上海辞书出版社,1987.

[20]ŽIŽKA J, DAŘENA F, SVOBODA A. Text mining with machine learning: principles and techniques[M]. Boca Raton: CRC Press, 2019.

[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]CHANG C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM transactions on intelligent systems and technology, 2011(2):1-27.

[23]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,2010:129.

[24]刘颖. 统计语言学[M]. 北京:清华大学出版社,2014.

[25]王克非. 语料库翻译学探索[M]. 上海:上海交通大学出版社,2012.

[26]王克非,胡显耀. 基于语料库的翻译汉语词汇特征研究[J]. 中国翻译,2008(6):16-21.

[27]XIAO R. How different is translated Chinese from native Chinese? A corpus-based study of translation universals[J]. International journal of corpus linguistics, 2010(15):5-35.

[28]胡显耀,曾佳. 翻译小说"被"字句的频率,结构及语义韵研究[J]. 外国语(上海外国语大学学报),2010(3):73-79.

[29]LEECH G, SHORT M. Style in fiction: a linguistic introduction to English fictional prose[M]. London/New York: Longman, 1981.

[30]辜正坤. 翻译主体论与归化异化考辩——序孙迎春教授编著《张谷若翻译艺术研究》[J]. 外语与外语教学,2004(11):59-63.

[31]姜静. 英汉文学作品中方言的翻译比较研究[D]. 北京:北京外国语大学,2017:113.

[32]张谊生. 副词的篇章连接功能[J]. 语言研究,1996(1):130-140.

[33]张谷若. 地道的原文,地道的译文[J]. 中国翻译,1980(1):19-23.

[34]翁义明. 基于语料库的副词"倒"的主观性英译研究[J]. 山东外语教学,2018(4):89-98.

[35]胡开宝,黑黟.数字人文视域下翻译研究:特征、领域与意义[J].中国翻译,2020(2):5-15.

[36]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.

[37]朱益平.阐释学三大原则对文学翻译的启示——以《德伯家的苔丝》多译本为例[J].江西社会科学,2010(1):217-20.

[38]智雨婷,李晓红.方言的不可译问题——以张谷若《苔丝》译本为例[J].华北理工大学学报(社会科学版),2017(5):145-148.

[39]YU J. Various voices in dialect and the frequency issue in the Chinese translations of *Tess of the D' Urbervilles*[J]. Neohelicon, 2021(1):415-429.

[40]NASSERI M,THOMPSON P. Lexical density and diversity in dissertation abstracts:revisiting English L1 vs. L2 text differences [J]. Assessing writing,2021(47).

[41]MALVERN D,RICHARDS B,CHIPERE N,et al. Lexical diversity and language development[M]. Berlin:Springer,2004.

[42]方梦之.译学词典[M].上海:上海外语教育出版社,2003:149.

An Investigation of Literary Translation Style Through ML Method:A Case Study of *Tess of D'Urberville*

Kong Delu

Abstract This paper applies classification and clustering methods in machine learning studies, builds a parallel corpus, and examines the translation styles of the three versions of Hardy's masterpiece *Tess of the D' Urberville*. From a total of 68 features, 15 significant ones are selected and quantitatively synthesized with examples for detailed explanation. The results show that these salient features can effectively distinguish the stylistic differences among the three translations, with both classifying and clustering experiments achieving an average accuracy rate of about 97%. The study found that at the document-level, each translation shows different style features at the vocabulary, syntax, and discourse aspects; in terms of the keyword level, the frequency differences of certain keywords also present the translator's personal preferences. The article provides data support and fine-grained analysis for previous qualitative research, while also proposing some corrective conclusions, such as the higher lexical density, extremely lower proportion of passive *bei* sentence, and similar number of idioms in Zhang's translation compared to the others. Eventually we attempt to provide some improvements and supplements to the research methodology in translation style and translator's style studies.

Key words machine learning; translation style; parallel corpus; *Tess of the D' Urbervilles*