机器学习视域下政经语篇翻译风格对比研究①

——以《国富论》中文三译本为例

孔德璐②
（同济大学）

【摘　要】本研究采用机器学习算法，探索政经语篇翻译风格特征研究的新路径。自建亚当·斯密著作《国富论》中文三译本平行语料库，从特征集中筛选出10个显著特征，并结合语篇进行阐释和总结。研究结果表明，筛选后的显著特征能够有效区分三个中文译本的差异，分类器和k-means聚类分析能够实现的平均准确率均达95%左右。在篇章层面，各译本均在词汇、标点、语篇层面显示出不同的风格特征；在检索关键词特征方面，使用频次差异足以显示出译者的个人偏好。本研究以期为政经语篇翻译风格特征的创新研究提供一定实用方法。

【关键词】机器学习；政经语篇；翻译风格；《国富论》

①本文系国家社科基金年度项目“神经网络机器翻译的译后编辑量化系统模型研究”（项目编号：19BYY128）的阶段性研究成果。

②孔德璐，同济大学外国语学院博士研究生，研究方向：机器翻译译后编辑、计量文体学、语料库翻译学。

引言

亚当·斯密是西方经济学奠基人，被誉为“古典经济学之父”，其著作《国民财富的性质和原因的研究》（以下简称《国富论》）成为“西方经济学的‘圣经’”，将经济学研究聚焦国民财富增长问题，使经济学成为一门有独立体系的科学（任保平，2003）。《国富论》一经问世便引起很大反响，并译成多国文字在世界范围内传播。

鸦片战争后，随着“西学东渐”思潮的兴起，西方资产阶级经济学说逐步引入中国。1902 年，上海南洋公学译书院出版《原富》一书，由资产阶级启蒙思想家、著名翻译家严复先生翻译，这是《国富论》第一次被译成中文（张登德，2010）。不管是 20 世纪上半叶中国饱受内忧外患、政局风云变幻之时，还是新中国成立后的百废待兴，以及改革开放后的经济腾飞，翻译界对《国富论》的重译和阐释工作从未停止。据统计，从严复开始，《国富论》林林总总共有 124 个汉译本（肖海燕，2018：42）。纵使译本众多，但学界多针对严复译本进行研究，对再译本的讨论略显不足，同时多译本对比研究寥寥无几，译本间翻译风格的定量实证研究更为阙如。基于此，本文将针对三个不同译者的《国富论》中文复译本，对政经语篇翻译风格特征进行量化对比研究。

1　文献回顾

翻译风格研究是基于语料库的翻译研究中的热门话题之一，主要涉及翻译语言模式的发现和特征的阐释过程。一方面，研究翻译风格可以深入了解翻译的复杂性以及不同翻译版本之间的差异，进而发现在不同语言和文化背景下最适合的翻译策略，以改进翻译品质，提高准确性和自然度。另一方面，研究翻译风格可以帮助翻译人员更好地捕捉原著的艺术风格和语言特点，并将其转化为目标语言的等效表达，这对于保持

原作的风格和情感非常关键。

国外采用基于语料库方法的文学翻译风格研究出现较早，Hermans（1996）认为，译本中隐藏着“译者的声音（Translator’s Voice）”。Baker（2000）提出译文中会不可避免地展现“译者的指纹（Thumbprint）”，其提出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。Winters（2004；2007；2009）分别从外来词、报道动词和语气助词的角度剖析《美丽与毁灭》两部德语译本中体现出的不同译者风格。国内相关研究也逐步跟进，徐欣（2010）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；张继东、朱亚菲（2020）从语言特征和非语言特征两个层面考察《追风筝的人》两译本所呈现的翻译风格。可以说，基于语料库的翻译风格研究在国内已经较为成熟，研究成果丰富（刘泽权，闫继苗，2010；黄立波，朱志瑜，2012；黄立波，石欣玉，2018）。

然而，基于语料库的翻译风格研究范式仍存在完善的空间。Mikhailov 和 Villikka（2001）使用语料库的方法，对比多个俄语和芬兰语对应译本，发现以往研究中常用的参数，如词汇丰富度、高频词等，并不能有效判断译者风格。另外，大多数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，却忽略掉一些更能突显风格的隐性特征，削弱总体概括能力；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”（Ilisei & Inkpen，2011）。

机器学习方法能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，减少研究人员主观选择特征这一“诟病”。该方法缘起计量风格学（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。El-Fiqi 等人（2016）运用两两对比分类模型，有效区分阿语—英语和法语—英语译作中的译者风格。Lynch 和 Vogel（2018）利用支持向量机模型，证明英语中的 N 元语法可以识别译者风格。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（2017）使用支持向量机模型区分《傲慢与偏见》的两个中文译本；孔德璐（2021）运用分类聚类方法，从篇章特征和关键词特征出发，成功将《德伯家的苔丝》中文三译本进行区分。

目前研究成果多见于文学翻译领域，鲜有研究涉及非文学语篇的不同翻译风格。在对待非文学翻译风格和艺术创造上，人们更多认为“‘艺术手法和风格’说的肯定是文学翻译……非文学翻译有的时候甚至无须考虑风格问题”（韩子满，2019）。如果说文学译者为了追求“文学性”尚可对原作进行一定的主观再创作，那么政经语篇作为非文学体裁的一个大类，其表现出的严肃性、哲理性和逻辑性则为探究不同译本的翻译风格提出了挑战。前文述及，《国富论》作为经典的经济学著作，相关翻译领域研究多为对单一译作的探讨和批评（李涛，2014; 刘瑾玉，2015），或是对《国富论》百年汉译史的研究等（刘瑾玉，王克非，2020）。

基于此，本研究采用机器学习的方法，通过自建《国富论》政经语篇平行语料库，对三位译者的翻译文本进行分类和聚类，在特征解释部分，结合具体语料，对差异性语言特征进行定性探究。本研究拟回答以下问题：

1）机器学习方法是否可以区分政经语篇的译本风格？

2）哪些特征最能够解释不同译本间的风格差异？

3）如何解释这些差异特征并进行分类和归纳？

2　研究设计

2.1　语料选择

此次研究选择《国富论》原文以及三个不同译本作为研究语料。英文原本选择 2015 年《国富论》重印本（Smith，2015）；三译本分别选择 2009 年上海三联书店出版的郭大力、王亚南合译本（以下简称郭译

本），2010 年中央编译出版社出版的谢宗林、李华夏合译本（以下简称谢译本），以及 2011 年陕西人民出版社出版的第 3 版杨敬年译本（以下简称杨译本）。通过豆瓣读书网 ① 对所选译本进行检索，杨译本评分最高（9.3 分），五星评分比例最大（72.2%），郭译本（评分 8.7，五星评分比 54.9%）和谢译本相差不大（评分 8.9，五星评分比 63.7%）。可见这三个《国富论》译本接受度较广、好评度较高，具有可比性。

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，用 NLPIR-ICTCLAS 汉语分词系统 ② 对语料进行分词和标注。随后，按照 UTF-8 编码下 30kB 的文件大小，对三个汉语译文样本切分。固定文件大小的切分方式有助于样本容量基本统一，即每个中文样本中包含 5,000—6,000 词，因此不同译本间字数的差异也就带来样本数量上的不一致。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《国富论》英文源本不参与实验。最终建成的英汉双语平行语料库概况如表 1 所示，最终实验所用样本共计 205 个。

表 1　语料详情

作品	样本	形符数	类符数
《国富论》源本	—	378,392	9,496
郭译本	66	333,817	12,036
谢译本	75	387,568	13,612
杨译本	64	328,999	9,566

2.2　特征集建立

参照以往研究（詹菊红，蒋跃，2017; Lynch & Vogel，2018），本文建立的特征集主要分为两大类，一类为篇章级特征（Document-level），

① 豆瓣读书网是读者点评书籍的平台之一，网址为 https://book.douban.com/（访问时间：2024 年 3 月 21 日）。

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/（访问时间：2024 年 3 月 21 日）。

包括从词汇、句法到篇章的三个层次共计 30 个具体特征；第二类为关键词特征（Keywords），使用 Antconc 3.5.8，逐个将三译本与兰卡斯特现代汉语语料库（LCMC）做对比，提取关键词表，选取每个词表中关键值最高的 10 个主题词，汇总去重后得到 25 个关键词。最终选取 30 个篇章级特征、25 个关键词特征，共计 55 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；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

篇章级特征			关键词特征
（1）类形比	（11）文言虚词比	（21）分号比例	亦、价格、资本、所、之、其、税、此、及、于、的、这种、劳动、殖民地、利润、其他、镑、会、他们、或、那些、么、所有、能、货物
（2）标准类形比	（12）实词比例	（22）顿号比例	
（3）平均词长	（13）实词密度	（23）省略号比例	
（4）名词比例	（14）虚词比例	（24）括号比例	
（5）动词比例	（15）虚词密度	（25）引号比例	
（6）形容词比例	（16）平均句长	（26）把字句比例	
（7）副词比例	（17）陈述句比例	（27）被字句比例	
（8）介词比例	（18）问句比例	（28）成语比例	
（9）代词比例	（19）感叹句比例	（29）连词比例	
（10）数词比例	（20）逗号比例	（30）助词比例	

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。

借助 Python 和语料库工具 WordSmith 6.0，批量提取出所有样本的类形比、标准类形比、总词数、总句数等特征。最终经过统计，将形成的 205 个文本样本转换为基于 55 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 3.8.4[①]，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。

2.3 分类器和算法

本研究使用 SVM、朴素贝叶斯和 C4.5 决策树分类器。SVM 分类器在 1963 年由苏联的著名数学家弗拉基米尔·瓦普尼克等人设计（Žižka 等，2019：211），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（张学工，2000：37）。SVM 以实现风险最小化为目标，在诸多领域的研究中得到应用，并取得较好效果，如语音识别、图像识别、作者判定等。朴素贝叶斯和 C4.5 决策树分类器分别基于概率和信息熵概念，是机器学习和数据挖掘研究中发展比较成熟，且较为常见的两类分类器。

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，该工具旨在帮助用户将 SVM 便捷地应用于实践当中（Chang & Lin，2011）。在参数设置上，通过 GridSearch 算法对 LIBSVM 进行调参，朴素贝叶斯和 C4.5 决策树分类器均使用默认参数。

3 研究结果

3.1 分类实验结果

我们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。单次划分可能会导致评估结果对特定的训练集和测试集具有依赖性。为了减少这种依赖性，交叉验证方法被广泛应用。该方法主要

① Weka（Waikato Environment for Knowledge Analysis）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，详情参考 https://www.cs.waikato.ac.nz/ml/weka/（访问时间：2024 年 3 月 21 日）。

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，这种方法就叫作 K 折交叉验证。一般将 K 取值为 10，也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。十折交叉验证是一种常用且广泛验证的方法，能够在很大程度上平衡评估结果的稳定性和数据利用效率。

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，通过计算各类别中样本例的卡方统计量的值来评估特征贡献度，从结果中挑选出 10 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

	特征选择前（全部 55 特征）			特征选择后（显著 10 特征）		
分类器	准确率	召回率	AUC*	准确率	召回率	AUC
SVM	98.0488 %	0.980	0.985	94.1463 %	0.941	0.955
朴素贝叶斯	99.0244 %	0.990	0.991	97.0732 %	0.971	0.995
C4.5 决策树	94.1463 %	0.941	0.954	95.122 %	0.951	0.973

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 55 个特征构建的分类器模型已经达到了理想的结果，平均准确率达到 97%，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，运用特征选择后 10 个显著特征的分类实验效果较好，平均准确率达到 95.45%，平均召回率和 AUC 值均达到 0.9 之上。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，同时可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。

表 4 列出 10 个经过特征筛选后对应的特征，其中篇章级特征共 6 个，反映出宏观词汇、句法层面的特征；关键词特征共 4 个，体现不同

译本中较为微观的用词差异。值得一提的是，篇章级特征中的文言虚词比这一参数，使用了文言作品中常见的单字虚词共 12 个：之、或、亦、方、于、即、皆、因、仍、故、尚、乃。

表 4　经过特征选择的前 10 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
232.314 ±12.203	1	*逗号比例*	199.254 ± 9.927	6	*文言虚词比*
232.314 ±12.203	2	*顿号比例*	191.545 ± 7.126	7	*及*
213.615 ± 5.678	3	*亦*	173.486 ± 10.23	8	*之*
211.435 ± 6.879	4	*平均词长*	171.373 ± 6.877	9	*虚词密度*
204.334 ± 6.133	5	*的*	162.613 ± 10.847	10	*助词比例*

3.2　聚类实验结果

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相似度，而不同簇的对象间差异尽可能大（刘颖，2014：123）。作为一种无监督的机器学习方法，聚类分析能够直观展现可视化结果，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作用。

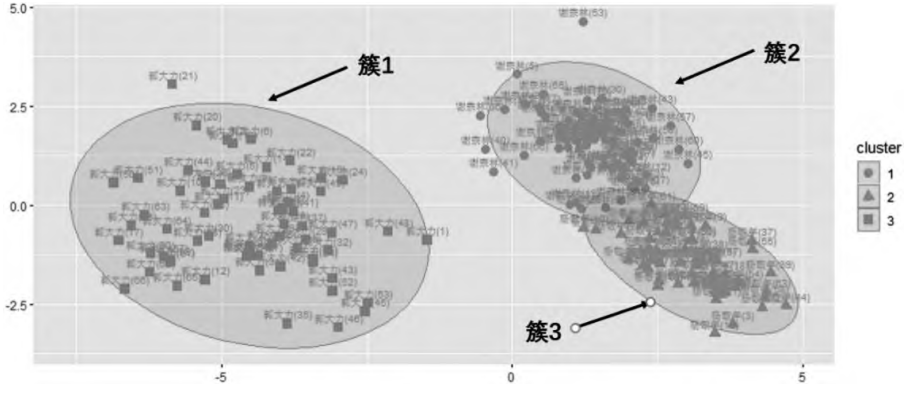

图 1　三译本基于 10 个显著特征的 K-means 聚类结果

本研究借助特征选取后的 10 个显著特征，使用 R 语言中的 K-means 函数，生成 205 个样本的聚类图（图 1），置信椭圆为 95%。如图 1 所示，所有样本明显区分出 3 个簇，簇 1 中的样本表示为方块，皆为郭译本样本；簇 2 样本表示为圆形，多为谢译本样本；簇 3 样本表示为三角形，多为杨译本样本。其中，簇 1 的椭圆面积较大，且可以明显与簇 2、簇 3 区分开，可见郭译本各部分之间风格活跃度较大，但在整体上仍保持有机统一。相反，簇 2、簇 3 的椭圆面积较小，样本位置较为集中，且两个簇之间距离很近，甚至有部分重合。这说明谢译本和杨译本的风格较为接近。同时，两译本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。

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 10 个显著特征可以有效区分《国富论》三译本的翻译风格，也能够说明机器学习算法可以完成较为困难的政经语篇翻译风格识别。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。

4　特征分析与讨论

4.1　篇章级特征

4.1.1　逗号比例与顿号比例

从表 4 的排名可以看出，逗号比例和顿号比例可以显著区分出三个译本。通常，研究者较少关注标点符号所起到的作用。在此次研究中，相较于谢译本和杨译本来说，郭译本在逗号比例和顿号比例有着显著的差异。结合表 5 我们可以发现，在逗号的使用上，郭译本整体占比约为 71%，要高于谢译本的 63% 和杨译本的 60%，均呈现显著差异（郭谢：Loglikelyhood= 269.536, p<0.000）。而在顿号的使用上，郭译本全文只出现了 36 次，占比不及千分之一，相反谢译本和杨译本中均出现超过一千多次，远多于郭译本（郭谢：ll= 1820.47, p<0.000）。

表 5　逗号与顿号比例

译本	标点总数	逗号总数	逗号比例	顿号总数	顿号比例
郭译本	59,598	42,602	71.48%	36	0.06%
谢译本	53,167	33,727	63.44%	1,404	2.64%
杨译本	39,685	23,848	60.09%	1,540	3.88%

注：逗号比例为逗号总数和标点总数的比值，顿号比例为顿号总数和标点总数的比值。

例 1 为从平行语料库中抽取的实际语料，这段话包含众多从业者的具体名称。结合实例能够发现在表达多个并列关系的名词时，谢译本和杨译本会灵活将原文中的逗号转换成连接并列成分的顿号。英文中逗号的作用包含连接并列成分，而郭译本则直接借鉴了英文逗号的用法，同样使用逗号来连接不同的并列成分，在一定程度上这也解释了郭译本中相对逗号比例较高、顿号比例较低的现象。

例 1：

（1）矿工、熔矿炉制造者、伐木者、烧制木炭供熔炉使用的工人、制砖者、叠砖者、照料熔炉的工人、安装或修理熔炉的工人、锻冶工与打铁匠等等，所有这些各行各业的人必须联合起来，才能做出这种简单的剪刀。（谢译本）

（2）采矿工、熔矿炉制造工、伐木工、熔矿炉所用焦炭的烧炭工、造砖人、泥水匠、护炉工、磨坊设计与建筑人、锻工、铁匠，全都必须把他们的不同手艺结合起来，才能生产出剪刀。（杨译本）

（3）为了制造这种剪刀也就须有许多种工人把他们各色各样的手艺结合起来，例如矿工，熔炉的建造者，木材的采伐者，烧炭工人，制砖工人，泥水匠，熔炉的工人，锻工，铁匠，以及其他等等。（郭译本）

究其原因，一方面，谢译本、杨译本灵活使用顿号来转换英文中逗号所发挥的连接并列成分的功能，体现出译者不拘泥于原文，会根据原文的具体含义和词汇之间的关系调整标点符号的使用。另一方面，郭译

本多选择沿袭英文中的逗号来表达并列关系是具有一定时代印记，该译本创作于 20 世纪 20 至 30 年代，在那个动荡年代为西方先进经济学说在中国的译介与传播做出了巨大的贡献（蔡强，徐偲，2020）。新文化运动同样为那个时期的语言使用（包括标点符号）带来了巨大的变革。五四运动之后，随着《新式标点符号议案》的颁布，顿号逐步从逗号中分离出来，并在教学实践过程中逐步展现出自身的意义和使用方法，成为汉语自有的一种标点符号（韩秋红，2011）。由此可见，顿号的分离是自新文化运动开始，并经历较长的实践阶段才被民众所接受，所以在郭译本产生的时代，逗号仍发挥着并列的词或词组之间的停顿作用。

4.1.2 平均词长

在三个译本当中，郭译本平均词长为 1.51（66 个样本中最低为 1.46，最高 1.55，标准差为 0.022），谢译本为 1.62（75 个样本中 Min.=1.56，Max.=1.7，S.D.=0.03），杨译本为 1.59（64 个样本中 Min.=1.54，Max.=1.67，S.D.=0.028）。可见郭译本在平均词长上要明显低于另外两个译本，且译本中各部分平均词长变化较小，整体保持了较为固定的字词长度。为进一步观察译本间的词长差异，我们逐一提取三译本的词表，并设置停用词，去除较为常见且影响研究结果的词语。随后从词表中选取前 100 个高频词，按照一字词、二字词、三字词和四字词的词长进行统计，结果如表 6 所示。

表 6 前 100 个高频词词长统计

词长	郭译本		谢译本		杨译本	
	类符	形符	类符	形符	类符	形符
一字词	43	33,432	27	27,493	27	26,533
二字词	53	31,301	70	45,271	68	41,834
三字词	4	1,743	3	1,641	4	1,962
四字词	-	-	-	-	1	373
总计		66,476		74,405		70,702

从统计结果来看，郭译本平均词长较短的原因是一字词比例较大，在同一分词标准下，郭译本中一字词的类符形符在高频词中几乎占到了一半的比例，而谢译本和杨译本中的一字词的形符比例只有 36% 左右。相反，在二字词上，郭译本的形符占比约为 47%，而谢译本和杨译本中的二字词形符比例约为 60%，类符比例甚至达 70% 左右。在三字词上，三译本的差别不大，而四字词只有杨译本词表中存在，为“大不列颠”，全文出现 373 次。总的来说，平均词长的差别在一定程度上能够反映出郭译本行文简洁明了，善于应用一字词来准确表达原文思想，这和以往研究结果较为吻合（蔡强，徐偲，2020）。

4.1.3　文言虚词比

为了体现译文中的独特语言风格特征，我们统计了 12 个文言虚词在各个译本中的比例，特征选择的结果也证明了文言虚词比例可以有效区分三个译本。通过图 2 我们可以看出，郭译本每万词中约出现 260 个文言虚词，而谢译本和杨译本中每万词的文言虚词只有约 50 次，前者约是后两者的 5 倍。

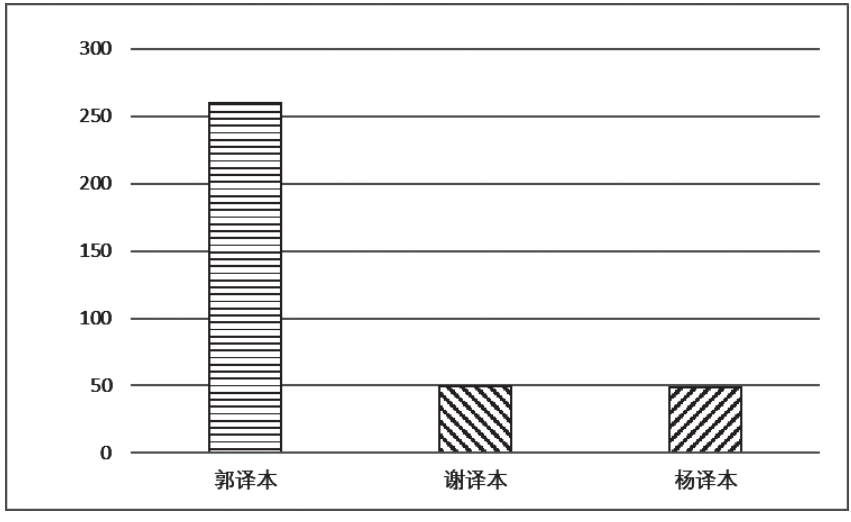

图 2　文言虚词在各译本中出现的频数（次 / 万词）

一方面，文言虚词具有“虚且活”的特点，所谓虚，即其词性和意

义均不及实词那样具体、实在；而所谓活，即用法灵活多变（周绪全，1995）。郭译本中丰富的文言虚词，能够帮助语句停顿，起到强调作用、变化语气语意、加重感情表达或调节音韵等，在句子中也能起到很好的表达效果。另一方面，谢译本和杨译本作为建国后的重译本，在文言虚词的使用上不及郭译本。然而，对于现代读者而言，郭译本这种“时代特点”反而会带来局限，例如在亚马逊网站[①]购买郭译本的顾客评价中，就有留言称“（郭译本是）比较早的译本，用词和时代有些脱节了”，或有“语言风格完全是民国时代的，不符合现在的阅读习惯”等等；当然也有读者评论称“翻译的语言简洁有力，现在的翻译版本篇幅太长”。整体看来，郭译本富含大量的文言虚词，对于现代读者来说不算友好，但因其形成于“觉醒年代”，虽囿于时代但不困于时代，仍不失为是一部经典翻译著作。

4.2 关键词特征

经过特征选择产生的 4 个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，通过观察这四个关键词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 7 显著关键词特征分布统计

	亦		之		的		及	
	形符数	比例	形符数	比例	形符数	比例	形符数	比例
郭译本	1,641	0.49%	2,544	0.76%	22,472	6.73%	2,544	0.76%
谢译本	223	0.06%	285	0.07%	32,893	8.49%	285	0.07%
杨译本	39	0.01%	330	0.10%	31,765	9.66%	330	0.10%

表 7 中呈现的是 4 个显著关键词特征在三个译本中的统计数据，可将这 4 个词分成两组，左边的两个词“亦、之”都属于文言虚词，而右

① 具体网址为：https://www.amazon.cn/dp/B00AE3Z9R2（访问时间：2024 年 3 月 21 日）

边的“的、及”中，前者属于助词，后者属于连接词。整体来看，郭译本中除了“的”字出现频数和比例明显低于谢、杨译本之外，其他三个词的比例都显著高于另两个译本。通过检索语料库，我们可以借助具体语境观察译者的用词偏好。

例 2 中的语料论述资本市场里利润和利率的关系。郭译本中出现 3 次“亦”，谢译本和杨译本多使用“也”。“亦”在古代多做副词，表示同样、也是；而“也”在古代多用做文言语气助词，现代多使用“也”来表示同样、并行的意思。可见郭大力在翻译《国富论》时会偏好“亦”这个具有古风的词来表达“同样的”意思。这也是郭译本中文言虚词丰富的特点之一，反映出译文中具有时代特点的偏文言译风。

例 2：

（1）我们由此确信：一国普通市场利息率变动了，资本的普通利润，亦不得不相应而一同变动。利息率下落，利润亦随而下落；利息率上腾，利润亦随而上腾。（郭译本）

（2）所以，无论在哪一国，当寻常的市场利率发生变动时，就可以推定资本的平常利润也有同样的变动。利率降时，它也降，利率升时，它也升。（谢译本）

（3）因此，根据任何一国通常的市场利息率的变动，我们可以肯定，资本的普通利润一定会随之变动，利息下降利润也下降，利息上升利润也上升。（杨译本）

其他三个词虽然不如“亦”一词的区分力度高，但相较于其他的语言特征，这三个词仍可以将三个不同的《国富论》译本区分开。囿于篇幅，不再结合语料展开分析。其中，“之”字在郭译本中出现频率较高，一方面，该词作为虚词，在一些结构中不提供实际意义，只帮助形成结构；另一方面，该词也能够作为助词，表领有、连属关系，如郭译本中第五篇标题“论君主或国家之收入”；而谢、杨译本均翻译为“论君主或国家的收入”。这表明在郭译本创作时代，“之”字普遍用来表示连属关系，在现代则更多使用“的”字，这也在一定程度上解释了郭译本中

“的”词数量不及其他两个译本的原因。

5 结语

本文借助机器学习中分类和聚类的方法，对政经语篇《国富论》中文三译本进行翻译风格考察，从 55 个特征中挑选出 10 个显著性强的特征，并结合语篇进行统计、分析和阐释。政经语篇本身具有严肃性、哲理性和逻辑性，为探索其不同译本的翻译风格提出了挑战。研究结果表明，机器学习分类算法可以有效区分三个中文译本，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。显著特征（表 4）在分类过程中发挥了重要作用，各译本均在词汇、标点、语篇层面显示出不同的风格特征。

在篇章层面，郭译本中逗号的使用较为普遍，在顿号比例上要明显少于其他两个译本。这体现出郭译本创造过程的时代特点，即逗号和顿号使用场景不明晰，两者功能在当时的语言环境中有所交叉。另外，郭译本较其他两个译本的平均词长较小，反映出郭译本用语简洁明了。郭译本中使用的文言虚词也远超其他两个译本，使其译风颇具文言气息，但用户评价对该现象褒贬不一。最后的关键词特征也验证了郭译本中文言虚词的大量使用，这和篇章层面特征所呈现的结果相辅相成，并且不同译本中较多或较少出现的关键词同样也展现了译者的个人偏好，区分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，为非文学文本的翻译风格特征研究提供了新思路和新方法。机器学习方法为翻译研究增添了科学性和客观性，通过实践研究可进一步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

蔡强，徐偲．基于语料库的郭大力翻译风格研究［J］．江西理工大学学报，

2020（2）：79-84.

韩秋红.“五四时期”国语（文）类教科书实践新式标点符号的讨论［J］.辽宁大学学报（哲学社会科学版），2011（2）：151-153.

韩子满．翻译批评的惟文学思维［J］．上海翻译，2019（5）：1-6.

黄立波，石欣玉.《到灯塔去》两个汉译本基于语料库的翻译风格比较［J］.解放军外国语学院学报，2018，（2）：11-19.

黄立波，朱志瑜．语料库翻译学：研究对象与研究方法［J］．中国外语，2012，9（28-36）.

孔德璐．基于机器学习的文学译者风格考察［D］．上海：上海外国语大学硕士学位论文，2021.

李涛．翻译与“国家富强”：析严复翻译之用意［J］．上海翻译，2014（2）：27-30.

刘瑾玉，王克非．岂一个“富”字了得？——《国富论》百年汉译史述论［J］．上海翻译，2020（2）：62-67.

刘瑾玉．严复手批《国富论》英文底本研究［J］．中国翻译，2015（5）：33-39.

刘颖．统计语言学［M］．北京：清华大学出版社，2014.

刘泽权，闫继苗．基于语料库的译者风格与翻译策略研究——以《红楼梦》中报道动词及英译为例［J］．解放军外国语学院学报，2010（4）：87-92.

任保平．论亚当·斯密《国富论》的方法论基础与特征［J］．经济评论，2003（2）：81-84.

肖海燕.《国富论》在中国的重译研究［D］．北京：对外经济贸易大学，2018.

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.

张登德．亚当·斯密及其《国富论》在近代中国的传播和影响［J］．理论学刊，2010（9）：95-99.

张继东，朱亚菲．基于语料库的《追风筝的人》两译本风格对比研究［J］.外语电化教学，2020（5）：50-57.

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）：34-46.

周绪全．文言虚词源流初探［J］．重庆师院学报（哲学社会科学版），1995（2）：112-115．

Baker, M. Towards a methodology for investigating the style of a literary translator[J]. *Target*, 2000, 12(2): 241-266.

Chang, C-C & Lin, C-J. LIBSVM: A library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): 1-27.

Hermans, T. The translator's voice in translated narrative[J]. *International Journal of Translation Studies*, 1996, 8(1): 23-48.

Ilisei, I & Inkpen, D. Translationese traits in Romanian newspapers: A machine learning approach[J]. *International Journal of Computational Linguistics and Applications*, 2011(2): 319-332.

Lynch, G & Vogel, C. The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations[J]. *Computer Speech & Language*, 2018, 52: 79-104.

Mikhailov, M & Villikka, M. Is there such a thing as a translator's style[C]. *Proceedings of the Corpus Linguistics.* Lancaster, UK, 2001: 378-385.

Smith, A. *The Wealth of Nations*[M]. London: Xist Classics, 2015.

Winters, M. F. Scott Fitzgerald's Die Schönen und Verdammten: A corpus-based study of loan words and code switches as features of translators' style[J]. *Language Matters,* 2004, 35(1): 248-258.

Winters, M. F. Scott Fitzgerald's Die Schönen und Verdammten: A corpus-based study of speech-act report verbs as a feature of translators' style[J]. *Meta*, 2007, 52(3): 412-425.

Winters, M. Modal particles explained: How modal particles creep into translations and reveal translators' styles[J]. *Target*, 2009, 21(1): 74-97.

Žižka, J, Dařena, F & Svoboda, A. *Text Mining with Machine Learning: Principles and Techniques*[M]. Boca Raton: CRC Press, 2019.

A ML-based Investigation into the Translation Style of Political Economic Discourse: A Case Study on Three Chinese Translations of *The Wealth of Nations*

Kong Delu

(Tongji University)

Abstract: This paper applies machine learning methods (ML) to explore new approaches to the study of translation styles of political economic discourses. Three different translations of *The Wealth of Nations* by Adam Smith are taken, first of all, to build up a parallel corpus. Then we select 10 distinctive features through ML algorithms, elaborating them with actual concordances. The result shows that the selected features can effectively distinguish three different Chinese translations, and the average accuracy of classifiers and K-means clustering result can reach about 95%. At the document level, three translations show distinctive style features at different aspects; in terms of keyword features, the difference in frequency of occurrence is sufficient to show different translator's personal preferences. This research provides a practical and innovative approach on the translation style of political economic discourses.

Keywords: Machine Learning; Political Economical Discourse; Translation Style; *The Wealth of Nations*