Modern Linguistics 现代语言学, 2023, 11(11), 5194-5203
Published Online November 2023 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2023.1111698

Hans汉斯

《大学》英译本特征比较

——基于Python数据分析技术

董艳华

昆明理工大学外国语言文化学院，云南 昆明

收稿日期：2023年10月7日；录用日期：2023年11月9日；发布日期：2023年11月20日

摘　要

典籍外译研究是推动中国优秀传统文化走出国门，面向世界的重要部分。为探究不同典籍译本的文本特征，本文自建四位译者《大学》英译本的小型语料库，基于Python语言统计各译本的词汇密度、平均句长和特殊句式数量等数据，从词汇、句法、篇章层面进行分析比较。分析数据发现，四译本用词都较为正式，句式都较为复杂，其中理雅各和陈荣捷的译本较为简练，连贯程度略低；林语堂译本的衔接更加连贯，更多地使用“释译”的翻译方法，更易于读者理解；庞德的译本最长，译本中较多的“释译”来说明原文含义，相较于其他三译本，更加通俗易懂。

关键词

《大学》英译本，文本特征，译文比较，Python数据分析

Comparison of Textual Features of English Translations of the *Great Learning*

—Based on Python Data Analysis Technology

Yanhua Dong

Faculty of Foreign Languages and Cultures, Kunming University of Science and Technology, Kunming Yunnan

Received: Oct. 7th, 2023; accepted: Nov. 9th, 2023; published: Nov. 20th, 2023

Abstract

The translation of Chinese classics plays an important role in promoting the dissemination of Chinese excellent traditional culture. In a bid to explore the textual features of different translations of Chinese classics, this paper builds a small corpus of four English translations of the *Great Learning*. Based on Python language, the data of the textual features, such as lexical density, aver-

文章引用: 董艳华. 《大学》英译本特征比较[J]. 现代语言学, 2023, 11(11): 5194-5203.
DOI: 10.12677/ml.2023.1111698

age word length and average sentence length of each translation are analyzed and compared from lexical, syntactic and textual levels. The results show that all the four translations utilize formal words and complex sentence structures. And the translations of James Legge and Chan Wing-tsit are more concise with weak coherence relatively. Lin Yuyang's translation is more coherent with more interpretive translation, which is more accessible to foreign readers. Pound's translation is the longest one and there are more paraphrases to explain the meaning of the source text. Compared with the other three translations, it can be understood more easily.

Keywords

English Translation of the *Great Learning*, Textual Features, Translation Comparison, Python Data Analysis

1. 引言

更好地推动中华文化走出去，典籍外译是至关重要的一部分，构建典籍译本语料库并应用语料库技术进行典籍外译研究，有助于增强我国文化软实力，促进中国传统文化更好地在海外传播。《大学》为“四书”之首，原是《礼记》中的一篇。程颐认为其是“孔氏之遗书，而初学入德之门也”[1]。《大学》阐明了古代成人教育的基本原则和途径，提出了“在明明德，在亲民，在止于至善”的三纲领和“格物、致知、诚意、正心、修身、齐家、治国、平天下”的八条目[1]。近年来，对《大学》英译本的研究主要为单译本研究或者多译本对比研究，少有学者基于 Python 语言对《大学》英译本的文本特征进行研究。本文建立了詹姆斯·理雅各(James Legge)、林语堂、陈荣捷(Chan Wing-tsit)和埃兹拉·庞德(Ezra Pound)四位译者《大学》英译本的小型语料库，基于 Python 相关的自然语言处理(Natural Language Processing, NLP)技术，从词汇、句法、篇章层面分析了四译本的文本特征。

2. 文献综述

语料库与翻译研究相结合最早可追溯至莫娜·贝克(Mona Baker)于 1993 年发表的论文，她在文章中指出“语料库可用于描写和分析大量客观存在的翻译语料，揭示翻译的本质”。[2]基于语料库对《大学》英译本展开研究的学者较少，其中徐兴梅借助 ParaConc 等工具从词汇与句子两个层面对《大学》英译本进行了研究[3]；徐欣基于文本数据挖掘技术研究了《大学》英译本的翻译风格的异同[4]。一些学者基于某一理论对《大学》英译本展开研究，如蔡昕辉研究了《大学》两译本在翻译对等理论视角下句子结构和选词方面的差异[5]；谭坤媛从认知识解理论视角出发，探讨了理雅各《大学》英译本中的概念隐喻[6]。还有一些学者研究了《大学》英译本的译介效果，如李耀研究了翻译传播过程中各因素对《大学》英译本译介的影响及其发生机制[7]；王炬炬考证了《大学》英译本的译者、书名、出版年代和出版社等因素，探究了代表性译者的译介目的和效果[8]。但少有学者基于 Python 语言对《大学》各英译本的词汇、句法和篇章进行整体分析与研究。

3. 语料库建立与处理

《大学》英译者主要包括：1) 来华传教士，如罗伯特·马礼逊(Robert Morrison)、约翰·马士曼(John

Marshman)、高大卫(David Collie)、詹姆斯·理雅各等；2) 本土华人，如辜鸿铭、林语堂等；3) 外裔华人，如陈荣捷等；4) 汉学家，如修中诚(Ernest Richard Hughes)、埃兹拉·庞德等。本文选择了影响力较大的四位译者的译文进行分析，即詹姆斯·理雅各[9]、林语堂[10]、陈荣捷[11]和埃兹拉·庞德[12] (译本分别简称为理译、林译、陈译和庞译)。所收集的语料来源于电子书籍，书籍详情如下表所示(表 1)：

Table 1. Details of the corpus texts
表 1. 语料文本详情

译者	书籍名称	出版社	出版年份
詹姆斯·理雅各	The Chinese Classics	Hurd and Houghton	1870
林语堂	The Wisdom of Confucius	Carlton House	1938
陈荣捷	A Source Book in Chinese Philosophy	Princeton University Press	1963
埃兹拉·庞德	Confucius: The Great Digest, The Unwobbling Pivot, The Analects	New Directions Pub. Corp.	1969

本文使用软件 ABBYY Finereader PDF 15 中的 OCR 编辑器和 Screenshot Reader 功能进行文本提取；将提取的文本存储为 TXT 格式文件，借助正则表达式进行文本清洗，删除多余的标点符号并统一为英文标点；而后人工进行逐字校对，确保文本的准确性。

4. 语料库检索与结果分析

4.1. 词汇层面

4.1.1. 类符/形符比

类符/形符比(Type-Token Ratio, TTR)是衡量文本中词汇变化程度的指标，用于衡量文本中词汇的多样性和丰富程度。其中，“类符”是指文本中词汇、符号的种类；“形符”是指文本中所有词汇、符号的数量。计算公式为：

$$\text{类符/形符比} = \frac{\text{类符数}}{\text{形符数}} \times 100\%$$

对同一规模的语料库而言，这一比值越大，说明文本所用词汇的变化程度越高，用词越丰富；反之，则变化程度越低，用词越匮乏。本文将所有单词进行小写处理，利用 NLTK 库中的相关函数计算形符和类符，所得结果如下(表 2)：

Table 2. TTR of the corpus texts
表 2. 语料文本的类符/形符比

译本	类符	形符	类符/形符比
理译	753	3658	20.59%
林译	755	3814	19.80%
陈译	706	3576	19.74%
庞译	1068	4159	25.68%

陈译和林译的类符/形符比最低，为 19.74%和 19.80%；理译略高，为 20.59%，庞译这一比值最高且远高于其他三位译者，为 25.68%，说明陈译和林译词汇丰富度最低，理译略高一些，庞译词汇丰富度最高。

4.1.2. 词汇密度

词汇密度也是用于衡量词汇变化程度与文本信息量的指标，计算公式为：

$$\text{词汇密度} = \frac{\text{实词数}}{\text{词汇总数}} \times 100\%$$

其中英文实词包括名词、动词、形容词和副词，这一比值越高，说明文本的词汇量越丰富，蕴含信息越多。本文去除译文中的停用词，利用 NLTK 库中的相关函数，进行分词和词性标注，统计实词数量，得到以下结果(表 3)：

Table 3. Lexical density of corpus texts
表 3. 语料文本的词汇密度

译本	实词数量	单词总量	词汇密度
理译	1118	3085	36.24%
林译	1225	3315	36.95%
陈译	1134	3124	36.30%
庞译	1343	3602	37.28%

理译和陈译的词汇密度最低，分别是 36.24%和 36.30%，林译略高一些，为 36.95%，庞译这一数值最高，为 37.57%，说明理译、陈译和林译呈现的信息更为简洁，而庞译所表达的信息相对更多。

4.1.3. 词长统计与平均词长

词长统计是计算文本中各种长度词汇的数量。平均词长在一定程度上可以体现文本的正式程度，越正式的文本其平均词长越长，反之则越短。本文在去除停用词后，统计了四译本各种长度词汇的数量与平均词长，得到结果如下(图 1)：

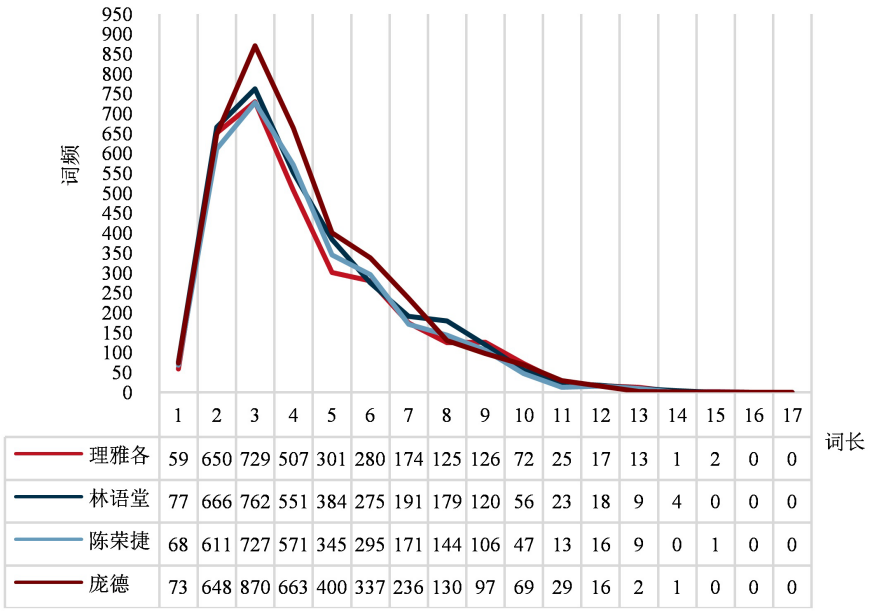

Figure 1. Number of different word length
图 1. 不同词长的词频

如图，四译文词长的整体变化趋势较为相似，在 1~3 的词长区间中，词频随着词长的增加而增加，在词长为 3 处达到极大值，而后随着词长的增加而减少，直至词长为 15 处，四译本都未使用长于 15 个字母的词汇。从词频数量来看，高频词长词汇主要集中在 2~6 这个区间，符合一般文本的词频分布[13]，其中庞译所用词汇最多，林译、陈译和理译依次之，词长数量在 300~900 之间；在 7~9 区间中，词汇数量偏低，四译本的词长分布较为相似，其中庞译词长为 7 的频数较高，林译词长为 8 的频数略高，词长数量约在 100~200 之间；词长为 1 或词长为 10~15 的词汇数量最低，在 0~100 之间(表 4)。

Table 4. Average word length of corpus texts
表 4. 语料文本的平均词长

译本	文本长度	单词总量	平均词长
理译	16,614	3085	5.39
林译	17,831	3315	5.38
陈译	16,607	3124	5.32
庞译	19,095	3602	5.30

通常情况下，一个单词的长度在 2~5 之间，平均词长为 4，平均词长长于 4 的文本语言较为复杂[14]。从平均词长的数据来看，四译本的平均词长在 5.30~5.40 之间，译文用词都较为正式，理译和林译的平均词长较高，复杂程度略高于陈译和庞译。

4.1.4. 高频实词与词云图

英文文本的高频词汇通常为虚词，如 the、and、a 等虚词，而文本的实际含义主要由实词体现。本文统计了四位译者译本的前 10 位的高频名词、动词、形容词和副词，结果如下(图 2)：

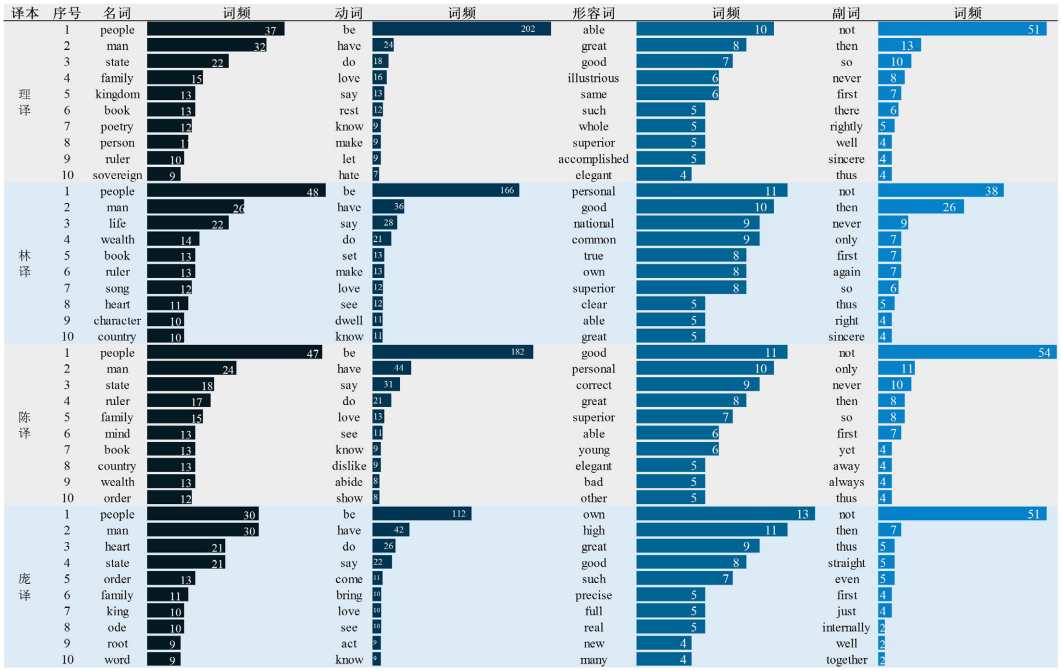

Figure 2. High-frequency notional words
图 2. 高频实词

从统计结果来看，形容词的高频词汇分布最为均匀，数量较少；名词与副词次之，数量居中；动词的高频词汇分布差异最大，数量最多。

名词层面，四译本中，频数最高的两个词都是 people 和 man，其他共同的高频名词有 state、family、ruler 等，数量在 10~50 之间，围绕着“明德、亲民和止于至善”三大纲领展开。《大学》中多引用《诗经》，四译本对《诗经》的译法不尽相同，理译为“Book of Poetry”，林译为“Book of Songs”，陈译为“Book of Odes”，庞译为“Odes”，因此 book、song、ode 也在不同译本中各为高频词汇。

动词层面，be 动词数量最多，在 110~200 之间，远高于其他词汇。其他高频动词有 have、do、say、know、love、hate、dislike 等，say 反映了原文多旁征博引的特点，know、love 和 hate 等体现“知本”、“好恶”等论证内容。

形容词层面，高频词汇数量较少，在 5~15 之间，分布较为密集，高频词汇有 good、great、own、superior、personal、national 等。原文中多次出现“君子”一词，既指品德高尚之人，也指地位高之人，如君王等。君子意为品德高尚之人时，四位译者的译法也各有不同，理译为“superior man”，林译为“gentleman”和“superior man”，陈译为“superior man”，庞译为“real man”和“true man”。

副词层面，频数最高的词都是 not，数量在 40~55 之间，远高于其他高频副词。理译、林译和庞译本居于第二位的高频副词为 then，其中林译中出现次数最多。

利用 Python 的 WordCloud 函数将前 60 的实词绘制成了词云图，结果如下(图 3~6)：

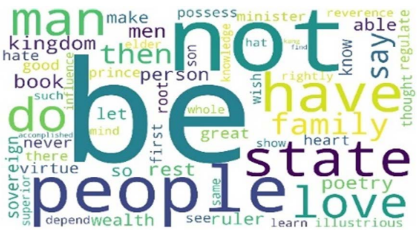

Figure 3. Word cloud of the translation of Legge
图 3. 理译词云图

Figure 4. Word cloud of the translation of Lin
图 4. 林译词云图

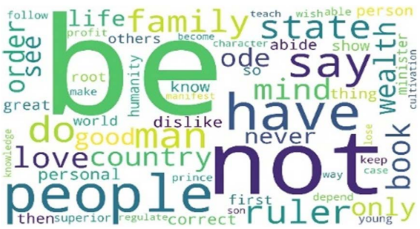

Figure 5. Word cloud of the translation of Chen
图 5. 陈译词云图

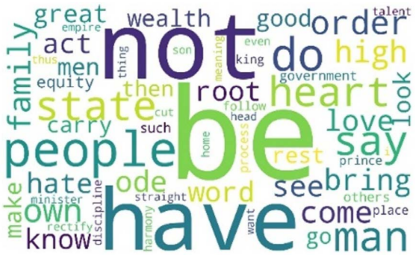

Figure 6. Word cloud of the translation of Pound
图 6. 庞译词云图

4.2. 句子层面

4.2.1. 平均句长与句段长

在句子层面，平均句长与句段长也是用来反映文本复杂度与正式度的指标，其数值越高，说明文本越复杂、越正式。平均句长是用形符的数量除以划分句子标点符号(句号、问号和感叹号)得出的，计算公式为：

$$\text{平均句长} = \frac{\text{形符数}}{\text{句点符号}}$$

平均句段长是用形符的数量除以句段标点符号(冒号、引号、逗号、句号、问号和感叹号)得出的，计算公式为：

$$\text{平均句长} = \frac{\text{形符数}}{\text{句段符号}}$$

笔者利用 NLTK 词性标注和正则表达式的功能，统计了文本中句子数量与句段数量，结果如下(表 5)：

Table 5. Average sentence and clause length of corpus texts
表 5. 语料文本的平均句长和句段长

译本	单词数量	句子数量	平均句长	句段符号数量	平均句段长
理译	3665	164	22.35	462	7.93
林译	3815	124	30.77	348	10.96
陈译	3576	177	20.20	324	11.04
庞译	4169	132	31.58	483	8.63

可以看出，庞译与林译的平均句长最长，为 31 个单词，远高于理译和陈译，约 21 个单词；而陈译与林译的平均句段长最长，约 11 个单词，高于理译与庞译，约 8 个单词。理译的平均句长与句段都较短，林译都较长；陈译的平均句长较短而句段较长，庞译的平均句长较长而句段较短。

4.2.2. 不同句式占比

基于语料库进行翻译语言特征研究时，句法特征分析包括句法总体特征和典型句式特征[15]。针对英汉译本的分析研究，相关学者着眼于“把”字句、“被”字句、“使”字句等。本文研究的为汉英译本，利用词性标注和正则表达式统计了被动句和定语从句的数量，来反映译本的复杂程度。这两类句式数量越多，文本越正式，具体统计结果如下(图 7)：

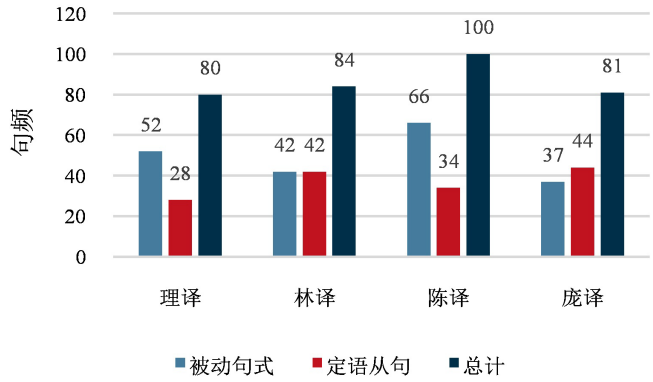

Figure 7. Ratio of different sentence patterns
图 7. 不同句式占比

被动句层面，陈译使用数量最多，为 66 句，理译略少，为 52 句，林译和庞译较少，为 42 句和 37 句；定语从句层面，庞译和林译使用数量最多，为 44 句和 42 句，陈译和理译略少，为 34 和 28 句。从两种句式的总和来看，陈译数量最多，共计 100 句，其他三位译者数量相近，在 80~85 句之间。

例 1:

原文: 大学之道，在明明德，在亲民，在止于至善。

理译: What the Great Learning teaches, is—to illustrate illustrious virtue; to love the people; and to rest in the highest excellence.

林译: The principles of the higher education consist in preserving man's clear character, in giving new life to the people, and in dwelling (or resting) in perfection, or the ultimate good.

陈译: The Way of learning to be great (or adult education) consists in manifesting the clear character, loving the people, and abiding (chih) in the highest good.

庞译: The great learning [adult study, grinding the corn in the head's mortar to fit it for use] takes root in clarifying the way wherein the intelligence increases through the process of looking straight into one's own heart and acting on the results; it is rooted in watching with affection the way people grow; it is rooted in coming to rest, being at ease in perfect equity.

分析:

以《大学》开篇之句为例，理译、林译和陈译句子与句段较短，整体较为简练，而庞译较为繁冗，句子句段较长。原文基本句式为“大学之道，在……，在……，在……”，理、林和陈的译文分别重复使用 to do、in doing 和 in doing 的结构来展现原文结构，而庞重复使用 take/be root in doing 并结合分号来复现，用词形式多变。“明明德”为“V + N”结构，理的译文“illustrate illustrious virtue”最为简洁，与原文形式一致；林与陈增译了限定词，分别译为了“preserve man's clear character”和“manifest the clear character”，而庞采用“增译”的方法，将其译为“clarifying the way wherein the intelligence increases through the process of looking straight into one's own heart and acting on the results”，蕴含的信息更多，译文也更为繁长。

4.3. 语篇层面

在语篇层面，为探究四译本的衔接手段与连贯程度，本文统计了四位译者译文中并列连词 and、or、nor 和 but 与介词/从属连词如 while、although、therefore、before 等的数量，结果如下(图 8)：

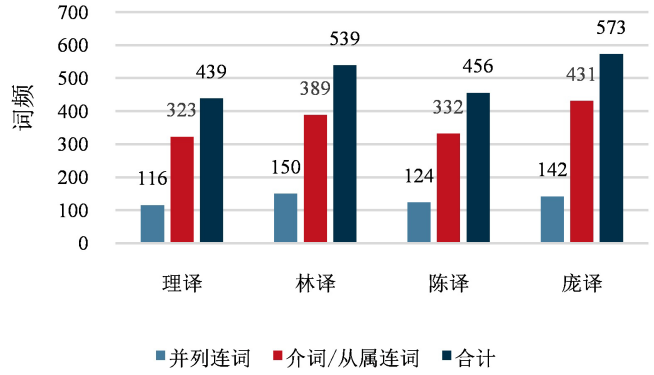

Figure 8. Number of conjunctive words
图 8. 关联词数量

并列连词层面，林译使用数量最多，为 150 词，庞译略少，为 142 词，理译最少，为 116 词。介词/从属连词层面，庞译使用数量最多，为 431 词，林译次之，为 389 词，陈译和理译较少，为 332 和 323 词。从总计的数量看，庞译最多，为 573 词，林译略少，为 539 词，陈译和理译较低，分别为 456 词和 439 词。

例 2:

原文：德者本也，财者末也。

理译： Virtue is the root; wealth is the result.

林译： Thus character is the foundation, while wealth is the result.

陈译： Virtue is the root, while wealth is the branch.

庞译： The virtu, *i.e.*, this self-knowledge [looking straight into the heart and acting thence] is the root; the wealth is a by-product.

分析：

中文为意合语言，句子的语法意义和逻辑关系通过词语或分句表达，无需借助语言形式手段。而英文为形合语言，句子的词语或者分句之间通常采用某种语言形式手段(如关联词)连接起来，表达语法语义和逻辑关系[16]。在本例中，原文为对仗句式，强调“德”为“本”而“财”为“末”，并未使用衔接词，理译与庞译的句式与原文一致，两分句之间用分号连接而未用连接词，庞译对“德”的概念进行了解释；林译与陈译在两分句之间添加了转折词“while”来显化原文的对比关系，林译在句前增译了“thus”来承接上句。

5. 结语

经过以上分析，可以发现四译本在词汇、句法、篇章层面的数据结果略有差异，但整体较为相似。四译本用词都较为正式，其中，理译和陈译都较为简练，词汇特征相似，文本长度相近，连贯程度相仿，平均句长接近，但陈译句段更长，林译对抽象概念如“大学”、“止”等词会补充释译，理译则不会，林译使用的复杂句式也更多，整体来看，两译本都呈现简而精的特征。林译也较为精炼，文本长度略长，平均句长和句段长都较长，连贯程度略高，对抽象概念会进行补充释译，更易于读者理解。庞译的文本最长，词汇密度最高，连贯程度较高，平均句长也最长，但句段长略低，复杂句式较低，译本中有大量

较长的释译对原文进行解释，相比于其他三译本，庞译更加通俗易懂。本文利用 Python 语言从词汇、句法、篇章层面对四译本进行了文本特征比较，存在一定的局限于不足，后续还可以容纳更多译本，研究文本相似度，结合译者的成长环境、教育背景和意识形态等因素对译文进行分析研究。

参考文献

[1] 朱熹. 四书章句集注[M]. 北京: 中华书局, 4-14.

[2] 胡开宝, 朱一凡, 李晓倩. 语料库翻译学[M]. 上海: 上海交通大学出版社, 17-29.

[3] 徐兴梅. 基于语料库的《大学》英译本译者风格研究[J]. 品位·经典, 2021(10): 55-57.

[4] 徐欣. 基于文本数据挖掘的《大学》英译本翻译风格研究[J]. 东方翻译, 2020(5): 36-40.

[5] 蔡昕辉. 翻译对等理论视角下的《大学》英译版本对比——以辜鸿铭与理雅各英译版本为例[J]. 牡丹江教育学院学报, 2015(7): 37+46.

[6] 谭坤媛. 认知识解视角下《大学》中概念隐喻的英译研究——以理雅各的英译本为例[J]. 东莞理工学院学报, 2020, 27(6): 92-97.

[7] 李耀. 传播学视域下《大学》英译史及译介过程研究[J]. 常州工学院学报(社科版), 2022, 40(4): 92-98.

[8] 王炬炬. 《大学》英译译介钩沉[J]. 齐鲁师范学院学报, 2022, 37(3): 151-156.

[9] Legge, J. (1870) The Chinese Classics. Hurd and Houghton, New York, 112-123.

[10] Lin, Y.T. (1938) The Wisdom of Confucius. Carlton House, New York, 139-152.

[11] Chan, W.-T. (1963) A Source Book in Chinese Philosophy. Princeton University Press, Princeton, 86-94.

[12] Pound, E. (1969) Confucius: The Great Digest, the Unwobbling Pivot, the Analects. New Directions Publishing, New York, 17-91.

[13] 刘泽权, 刘超朋, 朱虹.《红楼梦》四个英译本的译者风格初探——基于语料库的统计与分析[J]. 中国翻译, 2011, 32(1): 60-64.

[14] 祝朝伟, 李润丰. 基于语料库的庞德中国典籍英译译者风格探析[J]. 外语教学, 2023, 44(4): 75-82.

[15] 张跃伟, 潘宁, 贾鹰. 基于语料库的《红楼梦》霍译本中双及物小句句式翻译的认知语义动因研究[J]. 中国外语, 2023, 20(4): 86-94.

[16] 连淑能. 英汉对比研究[M]. 北京: 高等教育出版社, 73-88.