语言计量特征在译者身份判定中的应用*
——以《傲慢与偏见》的两个译本为例

詹菊红 蒋 跃**
(西安交通大学,西安 710049)

提 要: 本研究提出将语言计量特征应用于语言风格对比及译者身份判断的方法。通过对各 10 万字的两个训练译本语料库中 14 个语言结构特征分布的统计对比,发现两个译本中 5 个具有显著性差异的语言计量特征。以这 5 个特征作为译本表征,对各 10 万字的两个未知译者文本作相关分析并进行译者身份识别实验,根据这 5 个计量特征准确地判定未知译本的译者。研究证明,将基于语料库的计量研究方法与统计学方法相结合,用语言结构的计量特征标识译本的方法有助于加强译者身份、译本的判定以及译者风格辨别的可解释性和客观科学性,有助于弥补传统的译者风格定性研究的不足,使翻译研究趋于客观、科学并具有可解释性。

关键词: 译者身份判定; 语言计量特征; 相关性分析
中图分类号: H059 **文献标识码:** A **文章编号:** 1000 - 0100(2016)03 - 0095 - 7
DOI **编码:** 10.16263/j.cnki.23 - 1071/h.2016.03.019

Application of Quantitative Linguistic Properties to Translatorship Recognition

Zhan Ju-hong Jiang Yue
(Xi'an Jiaotong University, Xi'an 710049, China)

This paper proposes a method of using quantitative linguistic properties to contrast different translation styles and to recognize the translators of unknown translation texts. The Chinese translations of *Pride and Prejudice* by two different translators were cleaned, segmented, tagged and used as two corpora for analysis. The two corpora were divided into halves: the first halves to be used as training translation texts(TTT1 and TTT2) and the second as experimental translation texts(ETT1 and ETT2) whose translators were assumed unknown. Fourteen linguistic properties were explored and compared for significant differences between two training translation texts (TTT1 and TTT2). It eventually discovered five contrastive and differentiative linguistic properties typical of two different translators. Data of the five linguistic properties of two experimental translation texts (ETT1 and ETT2) whose translators were assumed unknown were also acquired and cross-compared with those of ETTs for statistical significance of differences by using Significance Test of Difference. It was found out that the five linguistic properties show no statistical significant difference between TTT1 and ETT1, and between TTT2 and ETT2; while the five properties show highly significant difference between TTT1 and ETT2, TTT2 and ETT1, and ETT2 and TTT2. Thus, it can be concluded that ETT1 and TTT1 belong to the same translator, WKY; and ETT2 and TTT2 belong to the other author, ZJH. The paper has successfully determined the contrastive linguistic properties between two translators, and based on these quantitative linguistic properties, ideally recognized the translators of experimental texts. The paper provides a new method for the research of different translation styles, translators' styles and translatorship recognition or even identification, which is intended to improve the accuracy, objectivity and explainability of the traditional studies of translation and translator styles.

Key words: translatorship recognition; quantitative linguistic properties; significance test of difference

---

* 本文系教育部人文社科研究项目"中医汉英平行语料库的构建与应用研究"(15YJC740127)和"在线机译与人工翻译的语言计量特征对比"(15YJA740016)的阶段性成果。

** 蒋跃为本文通讯作者。

1　引言

近 20 年来，作为一门新兴的语言学分支，计量语言学成为研究语言结构和语言演化的一个利器（刘海涛 2015: 40），也成为文学文本、翻译语言研究的重要工具。计量语言学认为，作者不同的语言风格是由于语言单位使用频率差异而形成的（黄伟 刘海涛 2009: 25）。近 30 年来，人们开始利用计量语言学的统计方法进行语言风格特征比较、文本年代的判定、文章作者的判定和作者风格的识别等（Hanlein 1999，Oakes 1998: 139）。黄伟、刘海涛（2009: 26）和陈芯莹（2012: 137）等曾通过对不同作家作品中语言结构特征的统计得出语言风格的一致性或区别性特征，利用语言结构的分布数据测量作家语言风格的计量特征，并且用以判断陌生文本的作者。同理，语言单位在译作中的分布数据成为体现译者语言风格的语言计量特征。因此，本研究假设，利用计量的方法对译作的语言进行统计分析，提取一些能够标识译者语言风格的语言计量特征集（set of quantitative properties），然后基于这些语言特征进行未知译本的译者识别或判定，它应该像陌生文本作者的判定一样是可行的。霍跃红（2010）用语料库方法分析典籍英译译者的文体并进行文本译者识别的实验。王青等（2012: 81 – 93）也用类似的方法进行作者指纹的建立和识别。但目前，采用基于语料库的计量研究方法判定未知译本译者，尤其是对翻译汉语文学作品的译者判定和识别的研究尚未见报道。

本研究基于语料库，运用计量语言学的方法，获取王科一和张经浩《傲慢与偏见》译本中有区别性的语言计量特征集，然后将这些计量特征与未知文本语言结构方面的数据进行文本相关性比对，进行未知文本译者的判定实验。研究的问题包括：（1）王科一和张经浩在《傲慢与偏见》译本中表现出的区别性的语言计量特征有哪些；（2）如何获取这些特征；（3）如何运用这些特征判定未知译本的译者？通过回答上述问题，以期为译者译本风格研究及译本译者的判定提供比较客观的和科学的研究方法和途径。

2　语料与工具

本研究语料的收集处理具体包括：语料的选取、语料清洗和分词赋码、分词语料的人工干预等。使用的语料库语言学软件主要包括 Ant Conc，CLAWS7，ABBYY，Ultra Edit 和 ICT CLAS 2015 等。

2.1 语料选取

秉承权威性和代表性的原则，本研究选取简·奥斯汀的小说 *Pride and Prejudice* 为源语文本。该小说的中译本有 30 多部，在中国读者中影响广泛。选取的译文语料样本分别为王科一和张经浩的译本。前者是 20 世纪后半叶的著名翻译家，其外国文学功底深厚且对英语语言驾轻就熟，曾翻译过多部世界文学名著，如《远大前程》、《十日谈》等，而且是第一位《傲慢与偏见》的中文译者，译著在 1955 年 2 月首次出版，并由上海译文出版社于 1980 年再版，已为广大读者所熟知，并且广受好评。王科一的翻译语言准确通顺、形神兼备、极其负责并忠实地传达原作的精神实质（李雪娇 2014: 48 – 49），是研究翻译语言语体和语用特征的理想样本。20 世纪末，随着对外开放很多杰出的中国翻译家致力于把经典文学名著介绍给中国读者，对很多名著开展重译，《傲慢与偏见》的重译本如雨后春笋般出现。其中最具代表性的当属 1996 年张经浩的译本。张经浩是这一时期最杰出的英译汉翻译家之一，翻译经验丰富，翻译风格与王科一的迥异。他主张在翻译过程中“去翻译腔”（张经浩 1999: 38），语言通达流畅，力求贴近目标语言。上述两译本首版时间相差 40 年，译者所处的不同时代背景和不同的翻译主张造成明显的译本语言结构差异和计量特征方面的不同，具有较强的可比性，是较理想的语料来源。

2.2 语料清洗

语料清洗是建立语料库的首要工作。通过扫描获得两译本的 PDF 格式文件，使用 ABBYY 软件进行格式转换，生成 Word 文档。由于 Word 文本均存在错别字和乱码的现象，因此，为了保证研究的准确性和客观性，须进行语料清洗和人工纠错，然后生成两个完整的译本分别保存为 wkypap. doc（218，425 字）和 zjhpap. doc（196，513 字）。清洗后的两个样本字数差别小于 10%，对语言特征的计量结果不产生实质性影响，可忽略。

2.3 语料加工

本研究采用中国科学院计算技术研究所开发的 ICTCLAS 2015 汉语词性标注软件对清洗后的两个译本进行分词标注。由于标注软件自身的局限性，词性标注加工完成后，不可避免地会出现一些错误。例如，由于电脑对句子语义的识别能力有限，会导致分词和标注错误。如“丽迪雅”是人名，软件将其标识为“丽/*ag* 迪雅/*n_new-word*”，而不能正确地处理成“丽迪雅/*nrf*”. 因此，本研究对标注加工后的译文再检查和校对，校对过程中，参照《CLAWS 7 词性赋码集》和《中科院计算所汉语词性标记集（3.0）》（刘群等 2008）以及《汉语现代词典》（第五版），对软件标注错误的地方进行人工纠错。分词标注后的文件分别保存为 wkypap. txt 和 zjhpap. txt.

3　方法

研究方法涉及语言结构计量特征的确定与甄选、数据统计以及数据处理与分析。使用的软件主要是 AntConc 3.4.3、WordSmith6 和 SPSSv21。

3.1 语言结构计量特征的确定

本研究考察语言结构的指标均为词汇层面和句子层面的语言结构特征。按照语料库翻译学和计量语言学文本聚类分析的惯例，主要选择部分代表语言结构长度、词汇丰富程度、词类和句式使用等方面的语言结构特征作为考察对象。参照黄伟和刘海涛(2009: 25 – 26)提出的用于文本聚类的汉语计量特征选择词长、句长、型例比、副词比例等 12 种结构类型作为考察对象，用以有效区分文本语体特征。须特别指出的是，在翻译文本中，形容词比例和四字惯用语比例也是体现译者风格(translator style)差异的一个重要计量特征(蒋跃 2014: 99 – 100)。所以，本研究把这两项也纳入考察指标中。基于文本聚类的汉语计量特征，本研究最终确定 14 种语言结构特征作为分析对象: 词长、句长、型例比、形容词比例、副词比例、名词比例、代词比例、助词比例、惯用语比例、标点符号比例、陈述句比例、疑问句比例、感叹句比例和单现词。

3.2 实验设计

为了测试和探寻到底哪些语言计量特征可以有效地判定未知译本的译者，本研究拟以文本为单位计算特定语言结构在文本中的频率和百分比，基于样本的均值比较这些语言结构在两组样本中的分布是否具有差异。即用训练样本获取的具有显著性差异的语言结构数据作为依据，对测试样本的相同语言结构特征进行检索和显著性差异分析，通过该实验实现译者身份的判定。

3.21 训练文本和测试文本的确定

首先，将王译本(wkypap. doc)和张译本(zjhpap. doc)对半平分成前后两部分，平分时保留截分处句子的完整性。前一部分作为训练文本，后一部分作为测试文本。也就是说，两个训练文本取自两译本的前一半字数; 待判定译者的两个测试文本为两个译本的后一半文本，与训练文本数据无交叉。截分的结果是，王译本均分为训练文本 TTT1. doc(109，211 字)和待判定译者文本 ETT1. doc(109，214 字); 张译本均分为训练文本 TTT2. doc(98，262 字)和待判定译者文本 ETT2. doc(98，251 字)。之后，在相同截分处将分词赋码后的王科一全译本 wkypap. txt 和张经浩全译本 zjhpap. txt 进行相应的均分，生成两个训练文本 TTT1. txt 和 ETT1. txt 以及两个测试文本 TTT2. txt 和 ETT2. txt.

3.22 研究样本的获取

为求得更加客观和准确的统计数据，并获取足够的样本量，本研究进一步对两训练文本和两测试文本的 4 个. doc 文件以每 5，000 字为单位进行均分，均分结果既保证每个样本字数基本相同，又保证样本之间的可比性; 同时，使得每组产生 20 个左右的样本，样本量比较合理。为保证截分处句子的完整性，每个文档的字数并非 5，000 整，但基本确保每个文档字数在 4，990 – 5，010 之间。而后，在相同截分点处，对 4 个. txt 文档进行截分，最终生成王科一训练样本 22 个，张经浩训练样本 20 个，未知译者 A 测试样本 22 个以及未知译者 B 测试样本 20 个。所有样本均为分词赋码过的文档，扩展名均为. txt，用 utf(8)格式保存。

3.3 数据统计与分析

首先对 wky 和 zjh 的两组训练样本中的 14 个语言计量特征进行显著性差异分析，获取其中具有显著性差异的语言计量特征，组成一个指标集(set of index)。然后对未知译者文本 A 和未知译者文本 B 中的两组样本中的相同指标用 AntConc 3.4.3 进行检索计算，得出这两组样本中相应的语言指标的数据。之后，以这些指标为变量，4 组样本两两交叉，进行独立样本 T 检验，以验证 4 组样本的计量语言特征之间是否具有显著性差异。如果两组数据间具有显著差异，说明它们并非来源于同一总体，即为不同译者所译。如果两组数据间的差异不大，且无统计学意义，则可判定它们来源于同一个总体，即为同一译者所译。

4　结果

4.1 训练样本中语言计量特征显著性差异分析

对王科一的 22 个训练样本和张经浩的 20 个训练样本中的词数、句数(以句号、问号和感叹号为标记)、副词数、形容词数、名词数、代词数、助词数、惯用语数、陈述句数、疑问句数、感叹句数、标点符号数、单现词以及型例比(TTR/每 5，000 字)以及惯用语用 AntConc 3.4.3 进行统计检索，然后将数据导入 Excel 中进行计算，得出前述 14 个语言特征的计量数据。

用 SPSS 对两组训练样本中的 14 个语言计量特征的数据进行独立样本 T 检验，最终从两组样本中发现 14 个语言特征中有 5 个特征的分布数据存在显著性差异。表$_1$ 和表$_2$ 所示是这 5 个特征的描述性统计结果和 T 检验结果。

表$_1$　两组训练样本中 5 个语言特征的描述性统计结果

指标	译本	样本量	均值	标准差	标准误
句长	TTT1	22	32.0975	3.6724	0.7830
	TTT2	20	25.0553	3.1613	0.7069
标点符号比例	TTT1	22	0.1038	0.0093	0.0020
	TTT2	20	0.1105	0.0096	0.0022
名词比例	TTT1	22	0.1643	0.0149	0.0032
	TTT2	20	0.1933	0.0159	0.0036
代词比例	TTT1	22	0.1484	0.0130	0.0028
	TTT2	20	0.1215	0.0144	0.0032
惯用语比例	TTT1	22	0.0116	0.0023	0.0005
	TTT2	20	0.0162	0.0040	0.0009

注: 表中数据均四舍五入到小数点后 4 位。(1)句

长 = 字数(不含标点符号) /句数; (2) 名词比例 = 名词数/词数; (3) 代词比例 = 代词词数/词数; (4) 惯用语比例 = 各类惯用语总数/词数; (5) 标点符号比例 = 标点符号数量/字数。

从表$_1$ 可看出,通过比较均值可知 TTT1(wky 译本) 的句长和代词比例高于 TTT2(zjh 译本) ,而 TTT2(zjh 译本) 在标点符号比例、名词比例和惯用语比例上高于 TTT1(wky 译本) 。

表$_2$　两组训练样本中 5 个语言特征的独立样本 T 检验结果

指标	F 值	T 值	自由度	显著值(双侧)	标准误
句长	1.0627	6.6279	40	.0000***	1.0625
标点比例	0.1539	-2.3217	40	.0254*	.0029
名词比例	0.0898	-6.1016	40	.0000***	.0048
代词比例	1.0274	6.3592	40	.0000***	.0042
惯用语比例	5.7628	-4.6274	40	.0000***	.0010

注: 独立样本检验前要假设具备两个条件,一个是两个总体呈现正态分布,另一个是两个总体具有相同方差。因此,先用 SPSS 对两组数据做正态分布检验,如果不符合正态分布,则对数据进行转化,使之符合正态分布; 另外,独立样本 T 检验自带方差齐性检验(Levene' s test) , T 检验的报表显示方差齐性检验的结果,只要 sig > .05, 就代表两组数据方差相同,取 T 检验结果中的第一行 sig 值; 检验结果发现,所有样本符合方差齐性。

由表$_2$ 可知,经过独立样本 T 检验得出的 5 个语言特征的 sig 值均小于临界值. 05。两译本的训练样本在句长、标点符号比例、名词比例、代词比例和惯用语比例这 5 个语言计量特征上均存在显著性差异。尤其值得注意的是,句长、名词、代词和惯用语比例的 sig 值均为. 000。在统计学上,P 值小于. 001 说明两组数据之间呈现极其显著的差异。这一数据表明,王科一和张经浩译本中这 5 个语言结构特征的分布差异极其显著,两译者在 4 个语言结构的使用方面差别非常大。因此,上述 5 个语言结构特征形成本研究要重点考察的语言计量特征集,提示两位译者在《傲慢与偏见》训练文本中这 5 个语言结构的使用方面具有区别性差异。那么,这些语言特征在测试样本中的表现如何,在训练样本与测试样本间是否具有显著性差异? 本研究将进一步分析检测,并据此验证未知译本的译者。

4.2 测试样本中语言计量特征显著性差异分析

为了验证两组训练样本间的 5 个语言计量特征是否存在显著性差异,研究对未知译者文本 ETT1 的 22 个测试样本和未知译者文本 ETT2 的 20 个测试样本中的 5 个语言结构特征频次,用正则表达式在 AntConc 3.4.3 中进行统计检索,之后将数据导入 Excel 进行计算,得出各个语言结构的计量数据。并用 SPSS 对两组数据进行独立样本 T 检验,得出这 5 个特征描述性统计结果(见表$_3$) 和显著性差异检验数据(见表$_4$) 。

表$_3$　两组测试样本中 5 个语言特征的描述性统计结果

指标	译本	样本量	均值	标准差	均值的标准误
句长	ETT1	22	31.1441	5.9203	1.2622
	ETT2	20	24.5776	3.2335	.7230
标点符号比例	ETT1	22	.1013	.0109	.0023
	ETT2	20	.1119	.0110	.0024
名词比例	ETT1	22	.1567	.0143	.0031
	ETT2	20	.1851	.0151	.0035
代词比例	ETT1	22	.1522	.0190	.0041
	ETT2	20	.1229	.0190	.0043
惯用语比例	ETT1	22	.0128	.0051	.0011
	ETT2	20	.0169	.0053	.0012

从表$_3$ 可以看出,通过比较均值可知未知译者文本 ETT1 的句长和代词比例均高于未知译者文本 ETT2,而未知译本 ETT2 在标点符号的比例、名词比例、惯用语比例的使用频率上高于未知译本 ETT1。

表$_4$　两组测试样本中 5 个语言特征的独立样本 T 检验结果

指标	F 值	T 值	自由度	显著值(双侧)	标准误
句长	7.3640	4.3967	40	.0001***	6.5665
标点比例	.0009	-3.1250	40	.0033**	-.0106
名词比例	.0970	-6.1632	40	.0000***	-.0283
代词比例	.0001	5.0010	40	.0000***	.0293
惯用语比例	.4256	-2.5689	40	.0140*	-.0041

表$_4$ 显示,经过独立样本 T 检验得出 5 个语言特征的 sig 值均小于临界值. 05。由此可知,两个未知译者测试样本在句长、标点符号比例、名词比例、代词比例和惯用语比例这 5 个语言结构特征方面均存在显著性差异。其中,句长、名词以及代词比例 sig 值为. 000,小于. 001,表明两未知译者文本在这 5 个方面存在极其显著的差异,尤其是在标点符号和惯用语的分布上存在更为显著的差异。据此可以判定,未知译者文本 A 和未知译者文本 B 来自于不同的译本,为不同译者所译。

4.3 4 组样本间语言计量特征的显著性差异分析

将两组训练样本和两组测试样本交叉进行语言特征的比对。为了判定未知译本 ETT1 是否来自 wky 译本,是否为王科一所译,需要将未知译者 A 样本中的 5 个语言特征数据与 wky 训练样本 TTT1 中相应的语言特征数据

进行显著性差异分析，如果存在显著性差异，则可断定未知文本 ETT1 并非王科一所译；如果 5 个指标均不存在显著性差异，则可判定未知文本 ETT1 来自于 wky 译本，为王科一所译。同理，为判定未知译本 ETT2 是否为王科一所译，需要将未知译本 ETT2 中的 5 个语言特征的数据与 wky 样本 TTT1 中的语言特征数据进行显著性差异分析，如果存在显著性差异，则可判定未知译本 B 并非王科一所译，反之亦然。同样，将未知译本 ETT1 与未知译本 ETT2 和张经浩文本 TTT2 中的 5 个语言计量特征分别进行显著性差异检验，判定未知译本 B 是否为张经浩所译。表$_5$ 显示 4 组样本两两交叉的独立样本 T 检验结果中的显著值，即 P 值的情况。

表$_5$　4 组样本间 5 个语言计量特征独立样本 T 检验的显著值（P 值）

比较样本	句长	标点比例	名词比例	代词比例	惯用语比例
ETT1vs. TTT1	.5244	.4264	.0944	.4345	.3011
ETT2vs. TTT1	.0000***	.0132*	.0001***	.0000***	.0001***
ETT1vs. TTT2	.0002***	.0061**	.0000***	.0000***	.0218*
ETT2vs. TTT2	.6394	.6859	.1039	.7953	.6383

由表$_5$ 可知，未知译者文本 ETT1 与 wky 训练文本 TTT1 的 5 个语言特征经过独立样本 T 检验得出的 sig 值均远远大于临界值.05，所显示的差异不显著，无统计学意义。未知译者文本 ETT1 与 TTT1 在上述 5 个语言结构特征上均不存在显著性差异。换言之，ETT1 与 TTT1 来自于同一译本。结合本研究实际情况可以判定，ETT1 来自于 wky 译本，为王科一所译。ETT2 与 wky 训练文本的 5 个语言特征独立样本 T 检验得出的 sig 值均小于临界值.05。其中，句长、标点符号比例、代词比例和惯用语比例 sig 值均为.000，P 值小于.01 说明两组数据之间的差异具有显著的统计学意义。可知，ETT2 与 TTT1 在 5 个语言结构特征方面均差异显著。据此可以判定，ETT2 不是来自于 wky 译本，非王科一所译。

未知译者文本 ETT1 与 zjh 训练文本 TTT2 的 5 个语言特征经过独立样本 T 检验得出的 sig 值均小于临界值.05，而且所有 sig 值均小于.01。当 P 值小于.01 时，说明两组数据间存在极其显著的差异。因此可知，未知译者文本 ETT1 与 zjh 训练文本 TTT2 在上述 5 个语言结构特征的比例方面均差异显著。据此可以断定，ETT1 并非来自于 zjh 译本，非张经浩所译。未知译者文本 ETT2 与 zjh 训练文本 TTT2 的 5 个语言特征经过独立样本 T 检验得出的 sig 值均远远大于临界值.05，因此可知，ETT2 与 TTT2 在上述 5 个语言结构特征比例方面的差异很小，或者均不存在显著性差异，没有统计学意义。据此可以判定，ETT2 来自于 zjh 译本，为张经浩所译。

5　讨论

本研究对王科一训练文本 TTT1、张经浩训练文本 TTT2、未知译者文本 ETT1 和未知译者文本 ETT2，分别以句长、标点符号比值、名词比值、代词比值和惯用语比值为考察对象，进行语言计量特征显著性差异分析，并进行译者身份识别实验。实验结果验证出我们的研究假设，证明运用译文中获取的语言计量特征可以准确地判定未知文本的译者，这一点我们通过回顾未知译者文本内容进行验证。

尽管黄立波等（2011）通过对照《红楼梦》几个英译本以及英国“翻译英语语料库”（TEC）翻译小说子库文学类翻译英语在类符/形符比、平均句长、叙述结构（这里指选择性 that 的使用）方面的比较，发现译者风格差异不明显，这些语言结构和特征“并不足以区分不同译者的翻译风格”（黄立波 2011：917）。另外，黄立波等人考察葛浩文英译中国小说的翻译风格后发现，利用语料库统计数据（如标准类/形比和平均句长等）并不能够有效地将一个译者与另一个译者的翻译风格区分开，这些统计结果更像是翻译文本表现出的一种共性。但是，本研究采用显著性差异分析方法，能够准确区分王科一和张经浩两译者的 5 个语言结构计量特征，证明尽管译者风格在宏观层面可能差异不大，尽管翻译文体表现出明显的共性，但是不能否认在某些语言形式的使用偏好上，不同译者、不同译本确实存在显著性差异，并且这些特征可以用来判定不同译者及其译作。

5.1 名词与代词

通过统计整个译本发现，王科一译本的代词比例为 15.03%，而张经浩的则为 12.22%，王科一代词的使用频率比张经浩的高出约 3%。此外，通用英汉对应语料库（CEPC）中汉语原创文学的代词比例为 8.83%，汉语翻译文学作品的代词使用频率为 11.81%（王克非 2012：63）。两译本的代词比例均远远高出汉语原创文学，这是因为汉语代词的类型较少，没有主宾格之分，使用频率较低，从而证实文学翻译中指代关系的“显化”现象（王克非 2012：87）。王科一的代词使用频率远高于汉语翻译文学的整体水平，而张经浩则略高，比较接近汉语原创作品的代词使用情况。

王科一在翻译中本着“极其负责任和忠实原文的思想态度”（李雪娇 2014：48－51），在翻译方法上趋向于靠近源语。王克非认为，英语作为一种形式化比较高的语言，代词使用的频率要明显高于汉语原创文学，“译者的忠实”会使代词从英语向汉语译本迁移，导致译本代词冗余现象。这个现象也与译者所处的年代有关，20 世纪 50 年代新中国刚刚成立，社会主义建设正在起步阶段，汉语

与英语之间的社会地位悬殊造成翻译策略和翻译产品的“逆差”(王克非 2012:114),代词的使用便是这种逆差的见证。虽然王科一译本形神兼备,迄今依然被推崇为《傲慢与偏见》最为经典和备受欢迎的译本之一,但译者身上深深的时代烙印藉此可见一斑。我们在 1:2 平行语料库中对两译本代词使用情况进行检索观察,也能看出两译本的差异。比如:

① She_PPHS1 felt_VVDanew_RRthe_AT justice_NN1 of_IOMr._NNB Darcy_NP1 's_GE objections_NN2; and_CC-never_RRhad_VHD she_PPHS1 before_RTbeen_VBNso_RG-much_RRdisposed_VVNto_TOpardon_VVIhis_APPGE interference_NN1 in_IIthe_AT views_NN2 of_IOhis_APPGE friend_NN1. //王译:她/rr 重新/d 又/d 想到/v 达西/nrf 先生/n 的确/d 没有/d 冤枉/v 她们/rr,/wd 他/rr 指出/v 她们/rr 的/ude1 那些/rz 缺陷/n 确/d 是/vshi 事实/n,/wd 她/rr 深深/d 感觉/v 到/v,/wd 实在/d 难怪/d 他/rr 要/v 干涉/v 他/rr 朋友/n 和吉英/nr 的/ude1 好事/n. /wj

张译:她/rr 又/d 在/p 想/v,/wd 达西/nrf 先生/n 看不惯/v 自/p 有/vyou 其/rz 看不惯/v 的/ude1 道理/n. /wj 他/rr 替/p 朋友/n 着想/vi,/wd 插/v 了/ule 一/m 手/n,/wd 现在/t 看来/v 的确/d 情有可原/vl. /wj

例①对同一英文原句的翻译,王译本中代词出现 8 次,而张译本中则只有 3 次。两位译者代词使用频率的悬殊显而易见。

王译本的名词比例为 16.05%,而张译本的名词比例为 18.92%,可以看出王科一的名词使用频率低于张经浩将近 3 个百分点。两译本中名词与代词比例总和分别是 31.08%与 31.14%,对比汉语翻译文学的研究数据,翻译汉语名词与代词的比例总和为 31.33%(王克非 2012:63)。两译本及汉语翻译文学作品的名词与代词所占的词频比例总和基本都在 31%左右。数据显示,名词词频比例高的文本,则代词比例低;而代词比例偏高时,则名词比例偏低。其原因可能是因为“汉语常规的指代方式主要以“名词复现”和“零指代为主,显性人称代词的使用频率一般较低;而代词是替代名词的一种词类,汉语中多数代词具有与名词相同的指别(deixis)功能”(王克非 2012:87),因此,代词与名词在指称功能上呈现一种“词频互补的关系”。张经浩的译本更加靠近汉语,翻译中刻意弱化源语的影响,尽量消除“翻译腔”。因此张经浩译本的代词使用频率相比于源文和王科一译本都大大降低。但是,为了避免因指代不明而引起的歧义,译者势必会使用相对多的名词明示原文的指代关系,这或许是张经浩译本名词比例增高的原因。

② Mr._NNB Bennet_NP1 saw_VVDthat_CSTher_APPGEwhole_JJ heart_NN1 was_VBDZin_IIthe_ATsubject_NN1;_; and_CCaffectionately_RRtaking_VVGher_APPGE hand_NN1,_, said_VVDin_II reply_NN1. //王译:班纳特/nrf 先生/n 看到/v 她/rr 钻进/v 了/ule 牛角尖/n,/wd 便/d 慈祥/a 地/ude2 握住/v 她/rr 的/ude1 手/n 说/v:/wp

张译:贝内特/nrf 先生/n 看/v 得/ude3 出来/vf,/wd 女儿/n 说/v 的/ude1 这/rzv 番/qv 话/n 完全/ad 是/vshi 内心/n 话/n,/wd 亲切/a 地/ude2 拉/v 起/vf 她/rr 的/ude1 手/n,/wd 答道/v:/wp

③ Very_RGfrequently_RRwere_VBDR they_PPHS2 reproached_VVNfor_IF this_DD1 insensibility_NN1 by_II Kitty_NP1 and_CC Lydia_NP1,whose_DDQGEown_DA misery_NN1 was_VBDZextreme_JJ,and_CCwho_PNQScould_VMnot_XXcomprehend_VVIsuch_DA hard-heartedness_NN1 in_IIany_DDof_IOthe_AT family_NN1. //王译:可是/c 吉蒂/nrf 和/cc 丽迪雅/nrf 已经/d 伤心/a 到/v 极点/n,/wd 便/d 不由得/d 常常/d 责备/v 两/m 位/q 姐姐/n 冷淡/a 无情/a. /wj 她们/rr 真/d 不/d 明白/v,/wd 家里/s 怎么/ryv 竟/d 会/v 有/vyou 这样/rzv 没有/v 心肝/n 的/ude1 人/n! /wt

张译:基蒂/nrf 与/p 莉迪亚/nrf 则/d 难过/a 至极/vi,/wd 多次/mq 埋怨/v 她们/rr 太/d 冷漠/a 无情/a. /wj 虽然/c 是/vshi 自家/rr 的/ude1 姐姐/n,/wd 这样/rzv 没/v 心肝/n 基蒂/nrf 与/p 莉迪亚/nrf 看/v 不/d 过去/vf. /wj

在例②中,王译本中出现两例人称代词“她”,而张译本中“她”只出现一次,但却用“女儿”来明示代词“她”的所指。另外,从名词的整体比例来看,王译本中名词的使用频率为 4 次,而张译本中达到 7 次。可以看出,两译者名词使用频率的差别,单凭名词使用频率的差异也能区分两个译本;同时也直观地呈现出两译本中名词频率与代词频率的“互补关系”。而例③中,张译本“基蒂”、“莉迪亚”分别出现两次,符合汉语以“名词复现”的方式明确指代关系的语言特征。

5.2 句长与标点

从句长与标点两个指标来看,王科一译本的句长均值为 31.63%,张经浩的则是 24.82%。相比原创汉语平均句长 25.46%和汉语翻译文学作品的平均句长 25.81%(王克非 2012:58-59),王科一译本的平均句长显然远远高于原创汉语文学作品、翻译文学作品和张经浩译本的句长,而张经浩的平均句长则接近翻译汉语的整体水平。王克非认为,“汉语翻译时的句子扩张与英语源语有关”(王克非 2012:60),译者越靠近源语,译本扩张越明显。张经浩译本的标点符号使用比例略高于王科一,这与他的译文尽量靠近汉语目标语,语言凝练、句子短小精悍有很大的关系。例如:

④ I_PPIS1 am_VBMonly_RRashamed_JJof_IOhis_APPGEasking_VVGso_RG little_DA1. //王译:他/rr 要/v 得/

ude3 这么/rz 少/a，/wd 我/rr 倒/d 觉得/v 不好意思/a 呢/y. /wj//张译: 我/rr 还/d 嫌/v 他/rr 开价/v 太/d 低/a. /wj

从例④可以看出两译文句长的差异，王译文句子较长，张译文的句子比较简洁。同时王译本在形式上比张译本更加贴近源语，译文从词汇与句子层面都与原文形成鲜明的对应，却并非死译硬译，译语温婉徐缓，这应该也是王译本备受推崇的原因之一; 而张译本更加贴近目标语，一个“嫌”字尽显汉语的凝练之妙，这种翻译取向使张译本语言紧凑简练，比王译本总字数少 1.5 万字之多。

5.3 惯用语

本研究所界定的惯用语包括习惯用语和四字成语，由软件自动分词标注人工校对的方法进行确定。张经浩的惯用语比例为 1.66%，而王科一的为 1.22%。惯用语形式短小、含义丰富，可使语言生动形象、含蓄幽默，意在言外。“常用词、习惯用语比例增高说明译本语言趋向于汉语目标语”(胡显耀 2010: 457); “惯用语使用的密集程度或者分布特征，往往能表明译者对译入语惯用语掌握的熟练程度，也能表明其语言和词汇的丰富度和变化度”(黄立波 2009: 84)。张译本比王译本更频繁地使用习惯用语和四字成语，使译文语言精辟凝练，富有文采，更贴近目标语文化。例如:

⑤ They_PPHS2 must_VMhave_VHIseen_VVN them_PPHO2 together_RL for_RR21 ever_RR22. //王译: 照理/v 应该/v 常常/d 看到/v 他们/rr 两/m 人/n 在/p 一起/s 呀/y. /wj//张译: 他们/rr 一定/d 看见/v 了/ule 这/rzv 两/m 人/n 形影不离/vl. /wj

⑥ I_PPIS1 am_VBMso_RGgrieved_JJfor_IF him_PPHO1. //王译: 我/rr 真/d 为/v 他/rr 难受/a. /wj//张译: 我/rr 拿/v 他/rr 也/d 无可奈何/vl. /wj

在以上两例中，张译文使用“形影不离”和“无可奈何”两个成语，译文读起来丝毫没有“翻译腔”，带有原创汉语文学作品的韵味。“负责任的译者在穿梭于两种语言之间进行协调时，总会尽量减少信息传输过程中的损耗和丢失，便于读者的理解和吸收。”(柯飞 2005: 307) 比照源语不难发现，王译本的语言表达更加精准严谨，这也印证其“忠实地传达原文精神实质”的翻译主张。

6　结束语

本研究以《傲慢与偏见》译本为例，获取能够有效区分王科一和张经浩语言特色的 5 个语言结构指标; 并经过实验和统计学分析，证明使用上述 5 个语言结构的分布数据作为文本的表示特征，可以在未知文本的译者身份判别方面取得可信任的结果。本研究证明，即使在译者风格不存在整体差异的情况下，也可以运用计量语言特征进行未知译者的身份判定，并为之提供可以借鉴的方法和基本思路。

参考文献

陈芯莹　李雯雯　王 燕. 计量特征在语言风格比较及作家判定中的应用——以韩寒《三重门》与郭敬明《梦里花落知多少》为例 [J]. 计算机工程与应用，2012(3).

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(6).

黄立波　王克非. 语料库翻译学: 课题与进展 [J]. 外语教学与研究，2011(6).

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).

霍跃红. 译者研究: 典籍英译译者文体分析与文本的译者识别 [M]. 上海: 中西书局，2014.

蒋　跃. 人工译本与机器在线译本的语言计量特征对比——以 5 届韩素音翻译竞赛英译汉人工译本和在线译本为例 [J]. 外语教学，2014(9).

柯　飞. 翻译中的隐和显 [J]. 外语教学与研究，2005(4).

李雪娇. 从译者序分析译本特色——以《傲慢与偏见》的两个中译本为例 [J]. 河北北方学院学报，2014(4).

刘海涛　潘夏星. 汉语新诗的计量特征 [J]. 山西大学学报(哲学社会科学版) 2015(2).

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).

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

张经浩. 重译《艾玛》有感 [J]. 中国翻译，1999(3).

Adolphs，S. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies* [M]. London and New York: Routledge，2006.

Hanlein，H. *Studies in Authorship Recognition — A Corpus-based Approach* [M]. New York: Peter Lang Pub Inc，1999.

Oakes，M. P. *Statistics for Corpus Linguistics* [M]. Edinburgh: Edinburgh University Press，1998.

Wang，Q.，Li，D.-F. Looking for Translator's Fingerprints: A Corpus-based Study on Chinese Translations of Ulysses [J]. *Literary & Linguistic Computing*，2012(1).

定稿日期: 2016-03-01　　【责任编辑　孙　颖】