艺术与文化研究
ISSN: 2789-5165 (print)
雙清學術出版社
Contents lists available at www.qingpress.com
Journal homepage: qingpress.com/zh-cn/journals/6

基于 R 的《细雪》两译本文体研究

王子睿[1*] 刘善钰[2]

1. 王子睿，广西大学外国语学院，研究方向：日本语言与文化，文本挖掘，a18523343480@126.com
2. 刘善钰，广西大学外国语学院
*通讯作者：王子睿

摘要：本文尝试构建一套基于开源软件 R 的语料库翻译文体学路径，以《细雪》的两个译本为研究对象，从词汇多样性、文本特征词、成语、句长等层面统计了两个译本存在的参数差异，考察了两译本在文体上的差异。本研究不仅证明 R 语言在语料分析方面卓越的能力，还尝试以数据的可视化的方式展示了《细雪》两个译本的文体差异。
关键词：翻译文体, 语料库翻译学, R 语言, 文体研究

1. 引言

《细雪》是日本唯美主义作家谷崎润一郎于 1943 年-1948 年之间分三部写成一部长篇小说，小说以旧贵族蒔冈家族的四个姐妹鹤子、幸子、雪子、妙子为主人公，围绕着三妹雪子的相亲故事展开。该小说不仅叙事节奏舒缓悠然，而且一改谷崎以往以“虐恋”为主要特征的的恶魔主义文学风格，又因融入了关西地区独特的风土人情、关西方言、上流社会的独特面貌等因素，被称为谷崎文学创作的巅峰作品。《细雪》又被称为谷崎版的《源氏物语》，是其用现代日语翻译了《源氏物语》之后创作的小说，该作透露着谷崎对以《源氏物语》为代表的古典日本美学的吸收、借鉴。《细雪》以日本传统写实主义为创作手法，具有典型的物哀色彩[1]，不管在研究日本美学还是日本文化方面都有极大的研究价值。《细雪》被名家多次翻译，那么小说独具的文体价值在经过译者创造性的翻译过程后，又体现出怎么样的译者的文体风格呢？针对此问题，本文将基于语料库的翻译文体学，对《细雪》的周逸之译本[2]、储元熹译本[3]进行多维形式参数进行量化分析，探究两个译本之间存在的文体差异，并尝试了通过基于 R 语言的语料库的翻译研究路径。

2. 基于语料库的翻译文体学文献回顾

文学作品带有鲜明的作者风格，这种风格普遍被认为是文体(styles)，而译者在翻译作品中体现

Article history: Received August 21, 2022 Accepted August 26, 2022 Available online August 27, 2022
To cite this document: 王子睿,刘善钰(2022). 基于 R 的《细雪》两译本文体研究.艺术与文化研究，卷 2，第 2 期，9-18 页.
Doi: https://doi.org/10.55375/jacr.2022.2.6

出的风格则被认为是译者的翻译文体。著名翻译家林少华(2009)曾言及翻译是一种创造艺术，在这种文学的转换过程中，必然会伴随着译者个性、译者风格(即译者文体)的介入[4]。这种由译者的无意识的习惯性语言特征与有意识表达出的语言特征的即为译者的风格，也可以称其为译者的文体。随着语料库技术、语料库相关学科的辐射，翻译作品的文体研究引起了文体学者和翻译学者的关注。Mona baker(2000)年指出基于语料库的翻译文体研究不仅继承了文学研究对作家创造性文体的关注，还继承了语言学对语言使用群体的关注[5]。申丹(2002)在讨论文学的文体学在翻译研究的应用时，指出文学文体学是连接语言学与文学批评的桥梁，其分析方法适合引入翻译学科[6]。在国内方面，黄立波(2009，2014)率先提出语料库翻译文体学翻译研究框架，并提出根据语言对比模式可以将语料库的翻译文体学分为语际对比与语内类比，又可以根据研究的参数不同分类为统计文体、叙事文体、语言文体三种参数[7]。。随着数字人文研究的发展，基于语料库的人文研究空前繁盛，对此，语料库翻译文体研究也应该借鉴语料库文体学、计算语言学、计量语言学等相邻领域的研究方法(黄立波，2018)。胡开宝(2018)指出当代数字人文研究愈来愈重视文本深度挖掘和智能分析等方法的应用，强调数据的可视化，翻译研究也该吸收先进技术，深度分析翻译本质和翻译规律[8]。并且由于可视化技术的应用，基于数据可视化的翻译研究以及相关关系的呈现变得生动、直观(胡开宝，黑黟，2020)[9]。

3. 语料处理与开源工具 R 包介绍

根据 IEEE Spectrum 平台对 2021 年最欢迎的编程语言进行排序，R 语言排在 Python、Java、C 等语言后，位列第六，IEEE Spectrum 对其定义为：一门专门应用于数据分析与数据挖掘的语言[①]。也正是其具有灵活的数据分析、以及卓越的绘图与可视化能力所以备受统计学研究者的热爱。Python 和 R 功能有所重叠，前者因其拥有较长的发展历史，还有核心团体更重视语言的底层架构，优于性能，而后者因其建设人员的背景较多的来源于统计学、生物学、心理学及人文社科领域，所以较性能而言，更重视任务导向，其操作更为简单[10]。1997 年 R 语言问世后，因其拥有强大的 R 语言社区，各种数据处理、统计、画图包(package)如雨后春笋般涌现，现在包的数量已经达到一万余个。特别是在文本分析方面，从低级的文本处理包到高级文本建模技术据统计数量至少超过了 50 个，而且 R 社区还在不断推动包与包之间的互融互通、研究者之间的交流(WELBERS 等，2017)[11]。在分词方面，中文分词包 jiebaR、Rsegword 等包出现解决了 R 平台中文分词进行后续 NLP 处理的难题，Stringr 包可以对字符串进行基于正则表达式的处理。在语料库分析包方面，Ingo Feinerer 教授的 TM 语料分析包、伦敦政治经济学院的 Ken Benoit 教授与因斯布鲁克大学的渡辺耕平教授等研究者受欧洲研究院资助携手开发的 Quanteda 包，其不仅具备超过了 WordSmith、Antconc 等集成软件的语料分析能力，也大幅度降低了用户使用 NLP 技术进行相关研究的学习成本，可以说 R 语言的 Quanteda 包在语料库分析方面存在着巨大的潜力[12]。

① 该排行榜是由 CareerBuilder、GitHub、Google、Hacker News、IEEE、Reddit、Stack Overflow 和 Twitter 八个信息元中的 11 个指标进行加权得出的结果，详情可以搜索 IEEE Spectrum 官网

笔者整理了目前在人文社科领域常用的包如下图所示：

表格一　文本挖掘类 R 包

分词包		**语料库分析工具包**		**可视化与绘图**		**机器学习**	
名称	用途	名称	用途	名称	用途	名称	用途
jiebaR	中文分词	Quanteda	多功能语料分析	ggplot	可视化	MLR	机器学习
Quanteda	多语种分词	tm	语料库	formattable	词频图	topicmodels	主题建模
Rwordseg	中文分词	tidytext	整洁数据	wordcloud	词云图	sentiment	情感分析

R 语言随着国内数字人文学科的兴起而受到重视，文本挖掘技术、计算机语言应用在语言学、翻译学、文体学等领域渗透的还不够，故本研究尝试提出一种基于 R 进行翻译文体研究的研究思路。

4. 语料库翻译研究方法与拓展

语料库翻译文体研究模式的统计参数通常以类形符比、高频词表、词汇密度、平均句长等基础数据切入，在此之上也有与从叙事学角度对文体加以阐释的研究(黄立波，2014），从利用多维分析模型对文体进行分析(刘泽权等，2017），也有从对拆译、减译的统计数据对文体进行分析的研究(张继东等，2020）。以上的研究都在基础数据的统计、分析的基础上加入了新的研究方法，为基于语料库的翻译研究注入了新的活力。但随着技术的进步以及研究的需要，一些年代久远的语料库分析软件疏于更新、维护，其对数据的分析的可信性难以保证，研究者在对加入新的研究方法、理论时，对需要对基础数据的统计方法、路径进行重新审视。例如：词汇丰富度是否应该引入 STTR 以外的统计方法同时作为观察数据？高频词表中存在大量的“停用词”，只观察高频词表是否会导致研究者只能观察到大量无意义词？汉语语料的句长分析是否应该引入句段长概念？等等诸如此类的问题都需要研究者重新反思。

基于语料库的翻译研究也需要吸收相邻学科的新技术、新手段，不仅要对基础数据与集成软件持不断反思、推陈出新的态度，也应该结合语言学、叙事学、文体学、计量语言学等临近学科的的成果与理论，吸收欧美、日韩等国家与港澳台地区人文研究的经验与方法，促进学科的在技术层面上的进步。基于以上的观点，本研究尝试性地基于 R 的 Quanteda 包，jiebaR 包，ggplot2 包等包对两个译本进行多角度的定量分析与可视化。

4.1.词汇多样性

词汇丰富度(Lexical richness)是一个宽泛的概念，该指标包括词汇多样性、词汇复杂性、 词汇密度和错误数量(张艳等，2012)[13]。其中词汇的多样性一直是语料库翻译学研究中词语层面不必可少的参考指标，其揭示了作者或译者在一篇文章中文章词汇的使用的广泛程度，该指标越高说

明用词更为广泛，该指标越低说明用词更趋于统一。在对两个文本进行分词后，使用 tokens_remove 函数去除所有符号、非汉字成分，使用 Quanteda 包中 textstat_lexdiv 函数可以快速计算出已分词文本的词汇丰富度，如下图所示(精确到小数点后三位)：

表格二　两译本词汇多样性对比

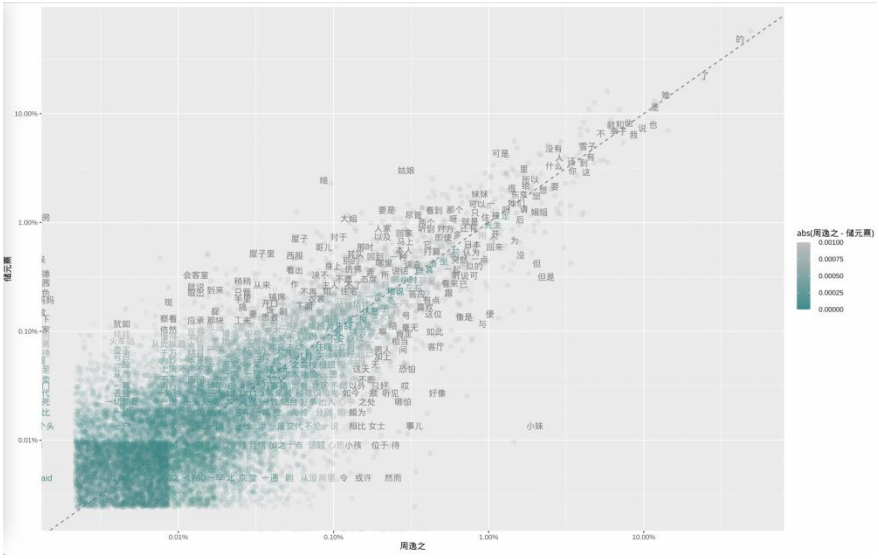
译本	C	TTR	R	token
周逸之	0.821	0.112	50.613	204086
储元熹	0.812	0.101	45.526	204269

从上图的 C、TTR、R 三个指标可以明显看出周逸之译本的词汇使用范围更广，从用词类型角度来看丰富度较高，储元熹译本词汇变化性稍低，在同等篇幅的译本中，较周译本词汇更趋于统一。因为两个译本在篇幅上并无明显的差异，故本研究没有采取 STTR 和 MATTR(移动形符比)，加入了 TTR 指标。

4.2.高频词可视化分析

以往的词表分析往往通过 Wordsmith Tools 等软件制成词表进行对比分析，这样粗略的对比往往会导致“的”“了”等不具备显著性差异的词频率较高，故本研究尝试通过高频词共现图来寻找两个译本中存在突出差异的词。通常来说词频大于 0.1%的通常被成为高频词。本研究借鉴并改写了 SILGE（2017)的词频可视化代码[14]，将《细雪》两译本中所有词频大于 0.1%的单词通过 ggplot2 包构图可视化，如下图所示。图中的虚线表示两个译本的词频相等，越往 x 轴偏离的部分说明该词在周译本中使用更为频繁，反之则意味着在储译本中使用更为频繁。

图 1 基于 ggplot2 的词频共现图



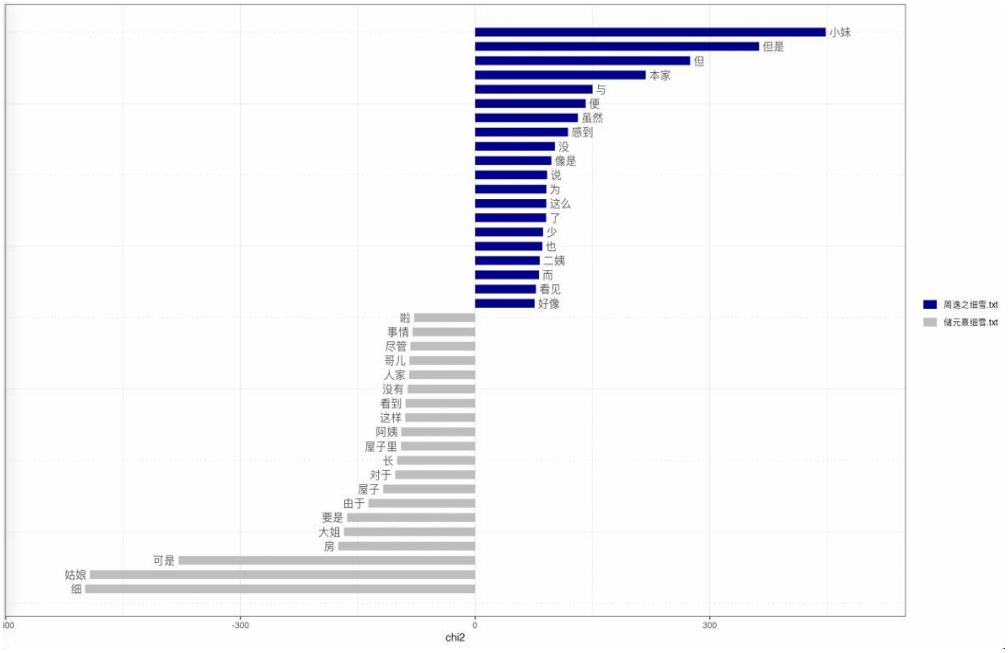
该图可以非常清晰地看出大于 0.1%的词频情况。例如储译本中典型的人物称谓方言化特色词

细姑娘，在该译本中多次出现，所以更偏离 x 轴，靠近 y 轴。在周译本中则出现了相对高频的“但”“但是”“便”等转折连词，所以更向 x 轴偏离，周译本还出现了更多的“感到”一词。反之储译本的“可是”一词出现频率明显大于周译本。显然高频词的可视化图只能起到辅助作用，我们需要对两个译本凸显的差异进行进一步的分析。

4.3.文本关键词

在高频词可视化环节虽然可以明显看出部分出现频次突出的词，但仅仅靠肉眼的识别不能客观地证明某个译本中哪些词可以区别译者文体特征，调用 Quanteda 包中的 textstat_keyness 函数可以比较目的语料库和参照语料库的特征词与特征词的重要性，该包相较于其他算法更为优越，集成了费希尔精确检验、卡方检验、likelihood ratio 似然比检验，在计算文本间特征词的精确度较高。textplot_keyness 可以用绘图技术可视化两个相对语料库之间的差异，笔者用其内置的 likelihood ratio 似然比检验算法得出了两个译本的特征词并根据其重要程度来构图如下所示。

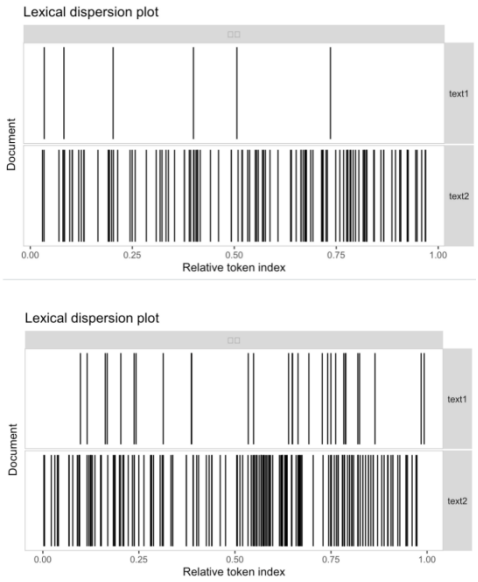
图 2 基于 textplot_keyness 可视化两译本前 20 个特征词



如图所示周逸之版本的《细雪》与储元熹版本具有最大区别的特征词就是对于四妹妙子的称谓翻译，前者将其译为小妹，后者则选用了我国湖南湖北四川等地的方言人称“细妹”。在翻译大姐鹤子家时，周译本选择了异化并加注的方式将其译为“本家”，后者进行了归化处理，译为了“长房”，此外前者还善用但、与、便、而等连接词，反映出与储译本在连词上具有明显的差别。很明显周逸之大量使用了偏文言形式的单音节连词，这一点有待进一步的考察。其次前者还倾向于使用“像是”“好像”等比喻处理手法，以“像是”“好像”两词为例，使用 quanteda. textplots 模块对“像是”“好像”两词的分布进行文本分布构图，如下所示我们很容易就可以看出周逸之使用了更多的“像是”“好像”两词，且词语的使用上在文本分布上有明显的集

聚倾向。

图 3 基于 quanteda. textplots 的词频分布图



4.4.成语使用率

成语通常是指经过语言使用群体长期使用、约定俗成的代表性词语，最为常见的是四字成语，也有三字、五字成语，其字数并不统一，其具有精简、通俗易懂的特点，是中华文化的结晶，具有丰富的文化历史底蕴。而且四字格中往往还有大量的虚词，如“天渊之别”“假以时日”等词含有“之”“以”两个虚词。成语的出现符合语言的经济原则②，起到了补足音节、增加音律感的作用，成语语体风格庄重而典雅(史维国, 2020)[15]。也有学者认为译者在译本中大量使用成语往往会使译本归化倾向更为明显，对译本中的成语使用持有保守态度(蒋志辉等,2014)[16]。笔者认为，译者在译文中使用成语主要源于三个方面的文体原因：其一是译者为提高译文的美学价值，试图用目的语文化负载词弥补翻译过程中的文体价值流失，比如说《细雪》中存在大量的日本地区独有的方言，其独特的语言价值、文学的美学价值在翻译过程中不免遭到破坏，译者为了补足这部分文体的流失则使用了大量凝练、简洁的成语，即：译出语文体与文化特色在翻译的过程中遭到了损失，而在译者的巧妙的加工下，同等价值的译入语文化被载入到译本之中。大量成语的运用不仅需要考验母语译者对成语的熟练运用以及较高的语言功底。其二是源于母语文化背景影响导致的译者在译本中留下母语语言的文体痕迹，例如在两个译本中“若无其事”一词在两个译本中都出现 22 次，“不知不觉”一词在储译本中出现 17 次在周译本中出现了 32 次，“若无其事”出自《苦菜花》，而后者出自《两教辨》，显然这些成语因其简洁、形象的原因已经被母语者所接受，母语者在形容类似语言现象时首先想到的就是此类成语，这也成为了母语语言的

② 法国语言学家马丁内提出，指的人们会自觉不自觉地使用更少的词语表达更多的意思，

文体特征。其三是避免用语重复，在翻译两个甚至多个同样意义的译出语时，译者会将其分别翻译为不同的表达，成语因其具有高度抽象、简练的特点往往会被考虑作为补充材料。

表格三 译者成语使用的原因

产生原因	译者意向	表现
提高文学文体的审美价值	有意识	名译本普遍大量使用成语
被母语文化影响的文体特征	无意识	“不知不觉”“若无其事”等常见成语
避免用语重复	有意识	增译、加译

综上所述，翻译作品中成语文体的主要是由这三方面的原因。另一方面成语的使用程度还意味着翻译文体的归化程度，较多的成语使用意味着译文的归化程度越高。笔者为了探究两个译本在成语层面的语言特征，利用了搜狗成语词库、jiebaR、Quanteda 包的 tokens_select 函数对两个译本进行了统计，结果如表格四所示。

表格四 两译本成语占比

成语	储元熹译本	周逸之译本	显著性检验	
字数	频数	频数	Loglikelihood	p 值
4	1749	2062	26.02	0.000 ***-
3	136	146	0.36	0.546 -
4 及以上	16	17	0.03	0.860 -
合计	1901	2225	25.76	0.000 ***-

根据统计的结果可以发现储译本较周逸之译本成语使用率更小，周逸之的全部成语使用率显著大于前者(LL=29.67 P=0.000 ***-)，说明周译本更倾向于使用成语来提高译本的文体价值、并且避免词语的重复性，这一点与前面得到的词汇丰富度指标是一致的。两个译本在三字成语、四字以上的成语方面不具备显著性差异，而且数量较四字成语更少，也说明了四字成语的使用在翻译文本中的体现更为明显。

4.5.文言文连词抽取

在进行高频词可视化后我们发现周译本出现了明显使用文言连词的情况，为了进一步探究译者的翻译文本文言连词的使用情况，我们通过对以下连词按照作用分类，共分为 8 类文言连词，对两部译作进行统计、可视化。

表格五 文言文抽取与显著性检验

抽取词类		储元熹译本	周逸之译本	显著性检验	
连词词类	具体检索词	频次	频次	Loglikelihood	p 值
承接	乃、遂、而、便、则	357	713	121.05	0.000-

转折	却、而、致	497	728	44.03	0.000-
因果	因、是故、以至、以	96	195	34.45	0.000-
选择	或、亦、即、孰与	7	42	27.77	0.000-
假设	若、譬如	2	27	25.67	0.000-
比较	如、同、不及、不如、则	198	153	5.74	0.017+
递进	况、并、且	78	186	45.60	0.000-
目的	以、以便	56	131	31.01	0.000-
共计		1291	2175	228.77	0.000-

通过组间两两对比我们可以明显看出：储译本的文言形式的连词使用率普遍低于周译本，只有比较类的词语显著高于周译本，前者倾向于使用对比类连词。而周逸之作为中华诗词学会的成员，笔名为楚之氓，在国内的报刊上发表了大量诗作，古诗中存在大量文言虚词，这一点也影响了译者文体在翻译作品中的体现。

4.6.句长

句长作为文体研究指标已经历史悠久。根据秦洪武(2010)的观点，句段长度与结构容量比句长分析更具有解释力，句段长度直接影响行文的流畅与语言运用的质量[17]。平均句长的计算方法是字数除以句号、问号、 感叹号的总数，句片段长的计算方法是字数除以问号、句号、逗号、感叹号、 分号的总数。平均句长从侧面上反映出译者发挥主体性对原文进行了有效重组[18]。

在句长方面笔者使用自行制作制作的函数包，其中含有三个函数可以分别对平均句长、平均句段长、平均句内字长，前二者是对句子和句段内所含词数进行统计，后者则是对字数进行统计，结果表格六所示：

表格六　两个译本句长比较

句长数据	储元熹译本	周逸之译本
汉字数	389114	386156
句子个数	11866	11609
句片段个数	32273	34875
平均句内句片段数	2.720	3.004
平均句长/字	32.792	33.264
平均句段长/字	12.057	11.073

通过上表可以明显看出，周逸之译本的句长、句段长都低于储元熹译本，而在平均句内句片段数方面，周逸之译本每句内的句片段数量更多，易读性变强，储元熹每句内的句片段数量更少，更偏向使用长句，易读性降低。

5. 结果与讨论

研究证明,周逸之作为中华诗词学会的成员在从事文学翻译的同时也创作了大量诗作，其较好的母语写作功底影响到了译本文体的表现，呈现出四字成语使用率显著高于储译本，文言虚词中的连词使用率显著高于储译本的特点，具有较强的典雅、音韵美，更符合目的语读者的文体审美。此外在句子层面上，平均句内句片段数(3.004)大于储译本(2.720)，呈现出行文结构灵活，行文归化程度更高的特点。在译者加注方面，译者还通过大量的加注试图弥补归化带来的原文的文体损失。储元熹译本的文体风格较前者而言异化程度更高，对成语与文言形式的连词较前者而言更为慎重储元熹译本整体偏向异化处理，在文言虚词方面更倾向运用比较类的文言连词，且显著大于前者，为了弥补原文中对人称描述的文体损失，创造性地使用了中国南方地区的方言，平均句内句片段数低于前者，体现出句片段普遍较长。

6. 结语

本研究基于 R 语言，通过多维度的客观的量化分析对比了储元熹译本与周逸之译本的文体特征，尝试了通过 R 语言的各种语料分析工具进行语料库翻译研究，从不同角度讨论了两个译本在文体上的差异。国内数字人文研究方兴未艾，对语料库翻译学各个领域的渗透还不够，语料库研究需要对既往基于集成软件的浅层数据分析进行反思和推陈出新。基于 R 的语料库翻译文体研究不仅在参数分析上具有可拓宽性，而且通过 R 的可视化功能可以使研究者更容易观察到语言数据的差异，助益于研究者对数据的观察、诠释，此类研究还需要学界进一步的探讨与实践。

参考文献：

[1] 刘青梅(2009). 试论谷崎润一郎作品与《源氏物语》间的关联[J]. 日本问题研究, 23(02): 53-57.

[2] [日] 谷崎润一郎(2017). 细雪[M]. 周逸之, 译. 上海. 译林出版社.

[3] [日] 谷崎润一郎(2011). 细雪[M]. 储元熹, 译. 上海：上海译文出版社.

[4] 林少华(2009). 文体的翻译和翻译的文体[J]. 日语学习与研究(01): 118-123.

[5] BAKER M(2000). Towards a Methodology for Investigating the Style of a Literary Translator[J/OL]. Target. International Journal of Translation Studies, 12(2): 241-266. https://doi.org/10.1075/target.12.2.04bak.

[6] 申丹(2002). 论文学文体学在翻译学科建设中的重要性[J]. 中国翻译(01): 10-14

[7] 黄立波(2009). 翻译研究的文体学视角探索[J/OL]. 外语教学, 30(05): 104-108. https://doi.org/10.16362/j.cnki.cn61-1023/h.2009.05.006.

[8] 胡开宝(2018). 数字人文视域下翻译研究的进展与前景[J]. 中国翻译, 39(06): 24-26.

[9] 胡开宝, 黑黟(2020). 数字人文视域下翻译研究：特征、领域与意义[J]. 中国翻译, 41(02): 5-15+187.

[10] 黄天元(2021). 文本数据挖掘——基于 R 语言[M]. 第 1st 版. 机械工业出版社.

[11] WELBERS K, VAN ATTEVELDT W, BENOIT K. 2017. Text Analysis in R[J/OL]. Communication Methods and Measures, 11(4): 245-265. https://doi.org/10.1080/19312458.2017.1387238.

[12] BENOIT K, WATANABE K, WANG H (2018). quanteda: An R package for the quantitative analysis of textual data[J/OL]. Journal of Open Source Software, 3(30): 774. https://doi.org/10.21105/joss.00774.

[13] 张艳, 陈纪梁(2012). 言语产出中词汇丰富性的定量测量方法[J]. 外语测试与教学(03): 34-40.

[14] SILGE J, ROBINSON D. 2017. Text Mining with R: A Tidy Approach[M]. 1st edition. Beijing ; Boston: O'Reilly Media.

[15] 史维国(2020). 现代汉语构词法中的文言用法研究[M]. 1 版. 北京: 中国社会科学出版社.

[16] 蒋志辉, 张焕香(2014). 论翻译中使用汉语成语的原则[J]. 英语研究, 12(03): 49-52.

[17] 秦洪武(2010). 英译汉翻译语言的结构容量:基于多译本语料库的研究[J]. 外国语(上海外国语大学学报), 33(04): 73-80.

[18] 严丽,严丹.语料库视角下“声声慢”两种英译文比较研究[J].湖北经济学院学报(人文社会科学版),2022,19(02):117-120.

A R-BASED TRANSLATION STYLISTIC STUDY OF TWO VERSIONS OF SASAMEYUKI

Abstract： This paper aims to establish an approach to corpus translation stylistic based on R, an open source programming language. The study observes the differences between the two translation versions of *Sasameyuki* as the research object through statistical analysis of parameter variations on lexical variation, text keyword, idiom, sentence length, etc., underpinning the outstanding capacity of R Language in corpus analysis. The statistics covered in this paper also provide a visualized interpretation of stylistic differences shown between the two translation texts.

Keywords： Translation Style, Corpus−Based translation study, R, stylistic