文章编号：#106

鲁迅与冰心短篇小说计量风格分析 *

冷婷[1]，刘颖[1]

（1. 清华大学人文学院中国语言文学系，北京 100084）

摘要：本文选用鲁迅的 33 篇小说与冰心的 50 篇小说为语料，通过对小说文本篇幅长度、平均段落长度、句长分布、词汇丰富度、标点使用的统计分析，发现鲁迅的小说篇幅长度变化大，平均段落与句子长度较短，词汇丰富度高；冰心小说反之。通过前 1000 个高频词的层次聚类实验发现，鲁迅小说多以乡土为背景，冰心小说多着眼于家庭。通过基于 SVM 的文本分类实验，发现冰心在小说历时创作的过程中，标点和词类的使用风格发生变化；鲁迅在不同题材小说的创作中，仅标点的使用风格变化较大，词类、二元标点以及二元词类的使用风格较为一致。

关键词：小说风格；统计；层次聚类；文本分类

中图分类号：TP391　　**文献标识码：**A

Quantitative Style Analysis of Lu Xun and Bing Xin’s Short Stories

Ting Leng[1], Ying Liu[1]

(1.Department of Chinese Language and Literature, Tsinghua University, Beijing, 100084, China)

Abstract: Selecting 33 short stories written by Lu Xun and 50 short stories written by Bing Xin as corpus, this paper analyzed the stylistic features of Lu Xun and Bing Xin’s short stories from the perspective of quantitative style. Features include the length of story, the average paragraph length, distribution of sentence length, Yule’s K characteristic and punctuation. Through comparative analysis, the following characteristics are found in Lu Xun’s short stories: the greater change in the length of stories, the shorter average paragraph, more short sentences, richness of vocabulary. The opposite situations are found in Bing Xin’s short stories. Though hierarchical clustering using the first 1000 high frequency words, it shows that Lu Xun’s short stories are often set in rural circumstances and Bing Xin’s short stories focus on families. The text classification experiments based on SVM turn out that Bing Xin changes the language style in the process of writing in terms of punctuation and part of speech. Lu Xun’s short stories maintain relatively the same language style regardless of various themes in terms of part of speech, punctuation bigram and part of speech bigram. Only in the use of punctuation, his language style is different.

Keywords: Short Story Style; Statistics; Hierarchical Clustering; Text Classification

1　引言

五四时期，白话文运动是对与口语越来越脱节的文言文的革命，鲁迅和冰心作为这一时期代表性的男女性作家，都受到了中西方文化的熏陶，具有深厚的文学功底和进步的现代意识。他们的小说均开始使用白话文创作，是现代短篇小说创作初期的优秀作品，其文体形式和语言风格对后来的文学创作都产生了深远的影响。

鲁迅作为白话文运动的领军人物，他在文学上的地位是毋庸置疑的，对鲁迅小说语言风格的研究更是硕果累累。在前人的述评与研究中，沈雁冰[1]评价《狂人日记》一文“冷隽、挺峭、含蓄半吐、风格异样”。郑振铎[2]评论鲁迅的小说语言“讥诮而沉挚”。李长之[3]则针对《呐喊》和《彷徨》中具体的小说作品，从写作技巧的角度对语言风格进行了评判。巴人[4]认为鲁迅文章在风格上受古文影响，“古朴、简劲，不事华饰”。

* **收稿日期：**2015 年 7 月 31 日　　**定稿日期：**

基金项目：清华大学人文社科振兴基金项目“不同文学作品的计量风格比较与研究”（20145081042）；国家自然科学基金重点项目“汉语认知加工机制与计算模型”（61433015）

冰心小说“诗化”、“散文化”的风格特征是以往研究中比较一致的评论。胡云翼[5]在《新著中国文学史》中称冰心“以纯粹的诗人赤子之心，提一枝珊瑚似的笔，来写母亲与孩子的爱，来写海的生活，她的小说几乎就是诗。”郎学初[6]认为冰心的小说在描写的手法和语言的运用上都呈现诗化的特征。林荣松[7]认为冰心的小说“文备众体”，诗化和散文化是其鲜明特征。还有一些研究考察了冰心创作过程中语言风格的变化。如，杨清[8]认为时代对冰心的语言风格产生了影响，在新中国成立之前，其语言风格为柔美，之后逐渐转变为质朴与壮美。李廷姬[9]将冰心的小说创作划分为三个时期，并认为创作中期的风格与前期相比有所转变。

对于鲁迅小说风格的研究，以感性的评判与分析为主，且大多是针对《呐喊》与《彷徨》这两部小说集，而《故事新编》这部小说集与前两部小说集在题材上存在较大的差异。这部小说集是在古代的神话传说以及历史史实的基础上进行的改编，同时穿插了对现代生活语言与细节的描写。鲁迅在不同题材小说的创作中，语言风格是否有所不同这一问题鲜少有人探究。在对冰心小说风格的研究中，虽然有学者已经注意到了其写作风格的历时性变化，但是研究方法同样也偏于感性。因此，本文将使用计量风格学的研究方法来比较与分析鲁迅和冰心的小说语言风格。

计量风格学是数理语言学的一个分支，该学科认为，教育背景、交际范围、思维方式等的不同使得每个个体都有自己独特的语言风格和表达习惯，并会通过某些特征形式不自觉地表现出来。这些特征形式如同个体的“语言指纹”一样，具有一定的稳定性。基于此，将特征形式进行量化，能够更加直观地观察与理性地判断作者的语言风格和表达习惯。对于计量风格学来说，有两个重要的问题：一是定义以及选择可以区别作者风格的语言特征，这些特征一般要求是可以量化并且稳定出现的。二是基于提取出的特征，使用有效的统计方法说明风格特点。

对于前者而言，标点字符、英文大小写字符、空格字符等字符层面[10]的特征，高频词[11]、功能词[12]、词长、词汇丰富度等词汇层面的特征均被认为是有效的语言特征。在短语与句子层面，Stamatatos[13]通过提取句子中名词短语的数目、动词短语的数目、名词短语的长度、动词短语的长度等特征来研究现代希腊语的文本风格。王景丹[14]使用平均句长、不同句型的句频来研究比较不同剧作家的语言风格。Baayen[15]通过将语料库中的句子表示为句法树，统计重写规则的频率作为语言特征进行作者的判别研究。

对于后者而言，频率统计、假设检验[16]、特征分析[17]、文本聚类[18]与文本分类[19]等是语言风格研究中的常见方法。本文中综合使用了基本统计量、假设检验、文本聚类以及文本分类多种方法来探究两者的小说风格。

2 语料库

本文使用的语料包括鲁迅的 33 个短篇小说文本以及冰心的 50 个短篇小说文本。这些电子文本均来自于新浪爱问共享平台[①]的下载，并在此基础上结合《鲁迅全集》[20]与《冰心小说》[21]、《冰心全集》[22]三本纸质出版物进行了较为细致地校对。

为了保证小说篇章故事的完整性，本文将每一篇短篇小说作为一个文本进行处理，并对每个文本进行编号，以字母 Z 开头的 01 至 33 的文本代码分别表示鲁迅的 33 个小说文本，字母 Z 取自鲁迅的原名（周树人）中姓的首字母；以字母 X 开头的 01 至 50 的文本代码分别表示冰心的 50 个小说文本，字母 X 取自冰心的原名（谢婉莹）中姓的首字母。校对后的文本以 UTF-8 的格式进行存储，使用中科院 ICTCLAS 汉语分词系统[②]进行分词与词性标注，使用日本同志社大学金明哲的 MLTP（Multi Lingual Text Processor）软件进行词频统计，使用自主编写的程序进行句长的统计。

① 该平台自 2014 年 5 月 5 日关闭整合。
② http://ictclas.nlpir.org/

经统计，鲁迅小说的语料规模为 191807 个字，冰心小说的语料规模为 195718 个字。在所使用的冰心小说语料中，创作于 1919 至 1925 年的有 31 篇小说（文本代码为 X01 至 X31），这也是冰心创作小说的密集时段，本文中将其设定为前期。之后，在 1929 年至 1988 年这 60 年的时间里又比较分散的创作了 19 篇小说（文本代码为 X32 至 X50）。本文将这一时间段设定为后期。

与冰心前期的同时段内，鲁迅有 25 篇短篇小说（文本代码为 Z02 至 Z26），其中 24 篇（文本代码为 Z02 至 Z25）现均收录在《呐喊》与《彷徨》这两本小说集内。除了这 24 篇小说外，文本代码为 Z01 的《狂人日记》创作于 1918 年，也收录在《呐喊》中。《补天》（又名《不周山》，文本代码 Z26）原收录在《呐喊》中，现收录在《故事新编》小说集内。

3 基本统计

3.1 篇幅长度

每个文本的字符数，反映的是每个短篇小说的篇幅长度。经统计，鲁迅最长的小说《阿 Q 正传》长达 21314 个字，最短的小说《一件小事》为 1027 个字。计算鲁迅文本字符数的标准差为 3803.8，远大于冰心文本字数的标准差 2861.5。因此，鲁迅小说作品与作品之间离散性较大。此外，从总体上来说，冰心小说的篇幅长度较短。

3.2 平均段落长

廖秋忠[23]认为，段落是居于句子或话轮之上建立篇章层级结构的中间单位，用来说明句子之间语义联系或功能联系疏密的不同。一般来说，篇章愈长，内容愈复杂，篇章的结构层次愈多。在视觉上，段落以换行为标志，使得整篇文章的条理清晰，眉目清楚，便于停顿、回味与阅读。统计每个文本的段落数，并根据文本的总字符数计算出每个文本的平均段落长可以反映鲁迅与冰心在小说创作中对篇章层级结构的把握。

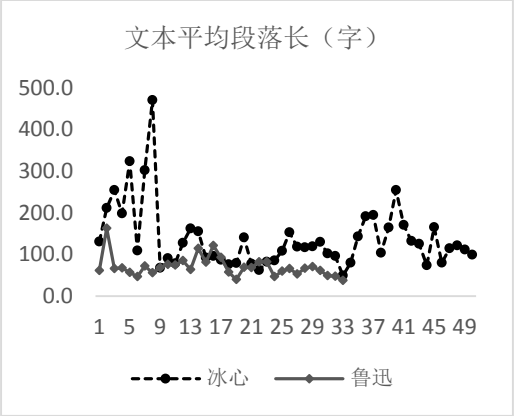

图 1 冰心与鲁迅文本平均段落长

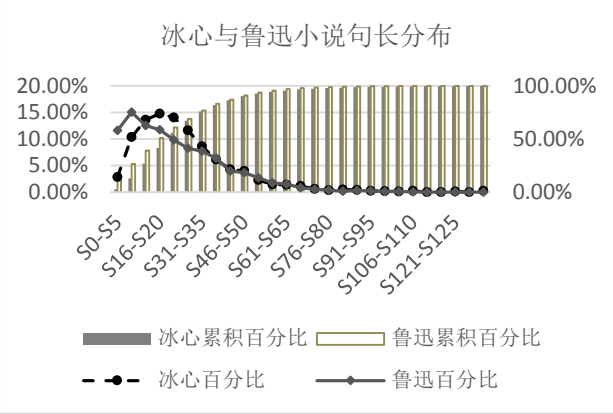

图 2 冰心与鲁迅小说句长分布

图 1 中横坐标表示每个文本，纵坐标表示每个文本的平均段落长，即每段平均的字符数。与文章篇幅相反的是，冰心文本平均段落长变化差异大，其离散度明显要高于鲁迅。

五四时期，对短篇小说这一文体仍处于探索阶段。但总的来说，小说被认为是一种集抒情、叙事、描写为一体的文体。冰心的小说以抒情见长，对故事的情景与片段进行描写。鲁迅同样借助小说展现社会问题，抒发个人情感。不同的是，鲁迅在运用对话展现故事的情节发展时，经常让话轮单独成为一个段落，呈现出类似戏剧的人物对话描写效果。冰心与鲁迅这两种描写上的差异部分体现在了平均段落长这一数据上。因此，即使鲁迅小说的篇幅长于冰心，但是其段落数也远多于冰心，如此，鲁迅小说的平均段落长明显低于冰心小说的平均段落长。

3.3 句长分布

句长一直是语言风格研究中重要的一个特征。将标点作为判断句子的依据，分别统计鲁

迅与冰心作品句长。句长统计所依据的标点包括：句号、问号和叹号。

为避免数据的过度稀疏，在对句长的统计中，以 5 为基准进行了分类汇总，即句长为 1，2，3，4，5 的句子个数均合计后记录在 S0-S5 这一特征数据下，句长为 6，7，8，9，10 的句子个数均合计后记录在 S6-S10 这一特征数据下，其余以此类推。特别地，句长大于 130 时，不再以 5 为基准合计，而是直接合并为一项，记录在 S>=131 这一特征项之下。

图 2 横坐标表示句长特征，左纵坐标为句长使用的百分比，右纵坐标为累积百分比。从句长数量的累积百分比上看，当句子长度在 20 字以内时，冰心句子数的累积百分比为 41.59%，而鲁迅句子数的累积百分比为 51.02%。当句子长度在 35 字以内时，冰心句子数的累积百分比为 75.92%，鲁迅句子数的累积百分比为 76.78%。当句子长度在 50 字以内时，冰心句子数的累计百分比与鲁迅句子数的累积百分比大致相当，分别为 90.39%和 90.90%。这说明不论是冰心的小说还是鲁迅的小说，大部分句子的长度不超过 50 个字。这与文体学研究中学者们对于鲁迅与冰心的小说简练这一语言风格的界定相契合。

从句长分布的百分比上来看，鲁迅短句使用数量要明显多于冰心短句的数量。其中，5 个字以内句子数百分比，冰心仅为 2.84%，而鲁迅的则高达 11.65%。当句子的长度在 6 个字到 10 个字之间时，鲁迅的句子数达到了峰值，冰心的句子数仅与鲁迅 5 个字以内的句子数相当，甚至还要低上 1.3 个百分点。冰心句子数的峰值出现在 16 个字到 20 个字之间，与鲁迅的峰值相比，总体增加 10 个字。句长在 11 个字以上 35 个字以内的句子数，冰心均高于鲁迅。

综合以上的数据说明，冰心小说中的句子总体上要长于鲁迅小说中的句子。

3.4 词汇丰富度

词汇丰富度是度量文本词汇复杂度的一项重要指标。1944 年 Yule[24]最早提出了 K 特征值。该特征值考虑出现频率的指数，不依赖于文本的长度。

$$\mathrm{K} = \frac{10^4 \times (M_2 - M_1)}{M_1 \times M_1} \tag{1}$$

其中，

$$M_2 = \sum_{i=1}^{v} i^2 V(i) \qquad\qquad M_1 = \sum_{i=1}^{v} iV(i)$$

V(i)为文本中出现了 i 次的词汇数，i 的最大值 v 为最高频使用的词语的词频，M_1即文本的总词频数。当 K 值越小，词汇越丰富；K 值越大，词汇的丰富度越低，文本词汇的复杂性越低。

将鲁迅与冰心的小说文本语料作为两个集合，分别计算其 K 特征值，得到鲁迅语料的 K 特征值为 70.78，冰心语料的 K 特征值为 81.80。

从总的数值上来看，鲁迅的 K 特征值要小于冰心的 K 特征值，因此，鲁迅语料总体的词汇丰富度要高于冰心。

分别计算 83 个文本的 K 特征值，采用两个总体平均数之差的参数假设检验 Z 检验[25]来验证上面整体数值所呈现的结果。

$$\mathrm{Z} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2}$$

其中，$\bar{X}_1$与$\bar{X}_2$分别为样本平均数，S_1^2与S_2^2分别是这两个样本的方差，样本数为n_1和n_2。则，

$$Z = \frac{(102.77 - 80.46) - 0}{\sqrt{\frac{19.12^2}{50} + \frac{12.67^2}{33}}} \approx 6.39$$

当设定的显著水平 α=0.05 时，临界值$Z_{0.05} \approx 1.65$，6.39 大于 1.65，因此接受原假设，即在 95%的置信度水平上认为冰心的词汇丰富度要低于鲁迅的词汇丰富度。

3.5 标点符号

为了判断鲁迅和冰心在标点使用上是否存在差异，统计每个文本中各个标点使用频次，然后利用卡方统计来检验[26]。表 1 给出鲁迅和冰心在标点符号使用上的卡方检验结果。以冒号为例，其卡方统计量(χ^2)为 442.29，相关联的 P 值为 0。由于 P 值小于显著水平 α=0.05，所以有 95%的把握拒绝原假设，即认为鲁迅与冰心在冒号的使用上存在显著性差异。

表 1　标点使用百分比及卡方检验

	鲁迅（%）	冰心（%）	χ^2	P			鲁迅（%）	冰心（%）	χ^2	P
：	1.54	4.6	442.29	0		！	2.64	3.52	36.43	0
；	2.79	0.79	316.98	0		《》	0.27	0.13	14.68	0.000
……	4.95	2.69	194.35	0		“”	7.29	6.51	12.99	0.000
，	46.74	52.17	165.53	0		？	2.67	2.43	2.95	0.086
。	19.49	16.77	69.69	0		——	2.56	2.34	2.85	0.091
、	0.01	0.27	64.78	0		‘’	0.58	0.63	0.42	0.515
（）	0.3	0.03	62.78	0						

通过表 1 可以发现，在显著水平 α=0.05 时，鲁迅与冰心在冒号、分号、省略号、逗号、句号、顿号、括号、感叹号、书名号、双引号的使用上均存在显著性差异，而在问号、破折号、单引号的使用上不存在显著性差异。

其次，从各个标点使用的百分比数据来看，逗号和句号均为两者最高频使用的标点，这与一般文章的写作规律相吻合。差别在于冰心使用逗号的频率远远高于鲁迅，而鲁迅使用句号的频率略高于冰心。句号表示句与句之间的停顿，逗号表示句内的停顿。同样表示句内停顿的还有分号，它主要用于复句内部分句与分句之间的停顿[27]。鲁迅在分号的使用上则要高于冰心。因此，鲁迅在句内的停顿相对来说多选用分号。这也体现了鲁迅在叙述描写上的风格。

同样地，对标点的二元特征使用χ^2检验可以发现，鲁迅与冰心在“冒号_前双引号”、“句号_前双引号”、“后双引号_前双引号”等特征的使用上存在显著性差异。其中，冰心“冒号_前双引号”的使用百分比为 4.16%，远高于鲁迅的 0.62%。而“句号_前双引号”、“后双引号_前双引号”的使用，鲁迅则要高于冰心。产生这种差异主要源于两位作者在人物对话的描写上采用了不同的手法。冰心在对话的描写中多使用冒号加前双引号。鲁迅则多直接引述，不使用冒号进行提示。

4　基于高频词的文本聚类

文本的聚类分析指的是，通过将文本转换为包含若干个特征的数据向量，并依据数据向量的近邻度将文本分类到不同类或者簇中，同时使得同一类中的文本类似度最高，不同类间的文本有很大的相异性。

在文本聚类的近邻度测算中，本文中主要使用的是 KL 距离[28]（Kullback-Leibler divergence）。KL 距离是对两个概率分布 P 和 Q 差别的非对称性的度量。对于一个离散随机变量的两个概率分布 P 和 Q 来说，他们的概率分布差异越大，KL 散度越大，其计算公式为：

$$\mathrm{KL}(P \,||\, Q) = \sum_{i=1}^{n} P(x) \log \frac{P(x)}{Q(y)} \qquad (3)$$

常见的聚类算法包括层次方法和划分方法。本文主要使用的是层次聚类的方法，使用离差平方和的方法[29]计算类间距离并自底向上进行聚类。

选择语料中前 1000 个高频词作为语言特征，计算词特征的 TF-IDF 权重[30]，对 83 个小说文本进行层次聚类实验。

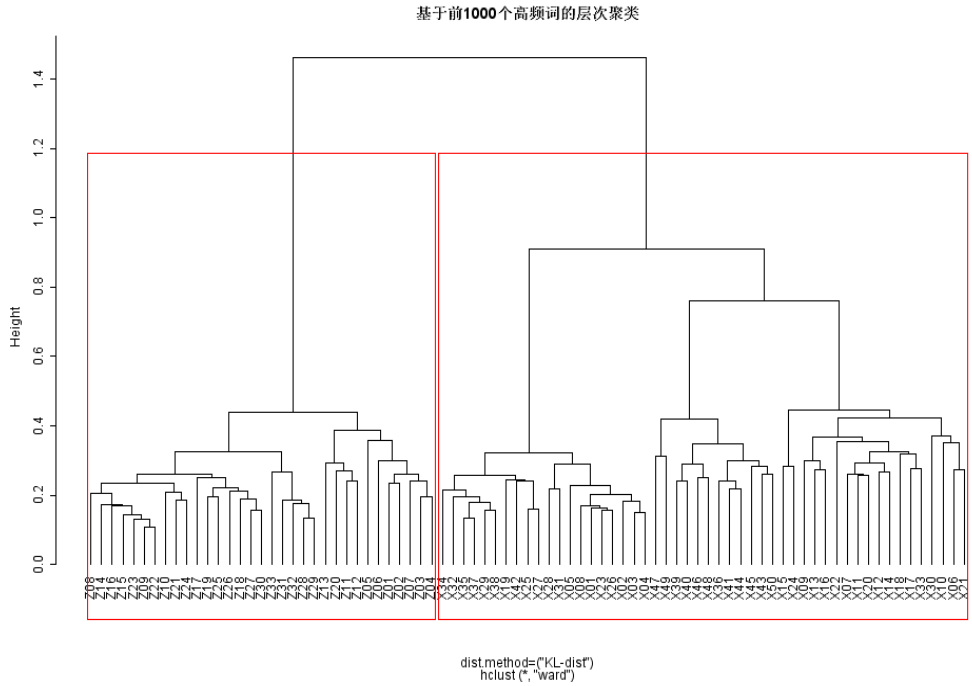

图 3　基于前 1000 个高频词的文本聚类

由图 3 的聚类结果可以看出，鲁迅的 33 个小说文本之间的距离较近，被划分为一类；冰心的 50 个小说文本之间距离也较近，被划分为另外一类；与此同时，两者文本的类间距离较远。因此，前 1000 个高频词对于区分两者的小说风格具有很高的代表性。这反映了鲁迅与冰心在小说创作中使用的语言和关注点的不同。

表 2　前 1000 个高频词中鲁迅与冰心的使用情况（部分）

	鲁迅	**冰心**
形容词	大，老，胖，寂静等	快乐，远，活泼，低，黑暗等
名词	老人，辫子，庄子，太爷，老爷，秀才，嫂子，老头子，女人，庄，村，祖母，男人，酒，船等	母亲，姊姊，妹妹，女士，同学，父亲，儿，弟弟，太太，信，诗人，爸爸，小姐，教授，哥哥，朋友，夫人，叔叔，先生，学校，墙，椅子，车等
代词	他，这，那里，大家等	她，我，我们，你，这时，她们等
动词	骂，闹，捏，咬，摸等	笑，说，听见，微笑，爱，谈话，想起等

表 2 所列举的是在前 1000 个高频词中两位作者相比于对方来说更高频使用的词语。由表可以发现，冰心小说中的亲属类称谓词出现尤为频繁，这反映了冰心的小说以家庭为立足点，描写在各种家庭关系中的冲突与矛盾。而鲁迅的小说则多以乡土为背景，描写封建社会下人物的愚昧以及其悲惨命运。

5 文本分类

基于机器学习的文本分类通常划分训练文本与测试文本，经过预处理、特征提取得到文本的特征信息形成特征集，并用计算机能够识别的形式进行表示（一般为向量空间模型）。计算机会根据训练文本特征集进行机器学习，构造分类器。生成的分类器将对测试文本进行测试，并输出测试的结果[31]。通过分类，我们将得到待分类文本的所属类别。

在分类的算法方面，常见的有：K 最近邻，决策树，随机森林，朴素贝叶斯，支持向量机等。支持向量机是由 Cortes 和 Vapnik 提出的一种可针对小样本学习的分类算法[32]。其本质是一种二类分类模型，寻找在线性可分情况下，使得两类样本在特征空间上间隔最大的超平面，其学习策略便是间隔的最大化[33]。支持向量机分类器的分类精度高、速度快，是应用最多的分类器之一。

文本分类实验主要探究两个问题：一是冰心在小说历时创作的过程中语言风格是否发生变化，二是鲁迅创作的不同题材的小说语言风格是否不同。

将冰心和鲁迅的小说语料划分为 4 个不同的数据集。其中，冰心的按照时间划分，鲁迅的按照题材划分。

表 3　文本数据集

文本数据集名称	**A**	**B**	**C**	**D**
划分依据	按时期		按题材	
	冰心前期	冰心后期	鲁迅《呐喊》《彷徨》	鲁迅《故事新编》
创作时间	1919 至 1925	1929 至 1988	1918 至 1925	1922 至 1935
包含文本数	31	19	25	8
文本代码	X01-X31	X32-X50	Z01-Z25	Z26-Z33

在 A、B、C、D 四个文本数据集中，从创作时间上来说，文本数据集 A 与 C 大致处于同时期；从题材上来说，文本数据集 A、B、C 均以现实生活为题材。

实验中，为防止出现过度拟合[34][35]，构造了文本数据集 A 的子集 E。文本数据集 E 包括了 1919 至 1925 年间冰心的 16 个小说文本，各个年份的文本数大致等同。

此外，为了保证实验的可比性，还分别构造了文本数据集 A 的子集 M 与 N，文本数据集 B 的子集 P 与 Q。

$E=\{X01，X02，X03，X07，X08，X11，X19，X20，X21，X25，X26，X27，X28，X29，X30，X31\} \subset A$

$M=\{X01，X02，X04，X07，X10，X13，X16，X19，X22，X25，X28，X31\} \subset A$

$N=\{X03，X06，X09，X12，X15，X18，X21，X24，X27，X30\} \subset A$

$P=\{X32，X35，X38，X41，X44，X47，X50\} \subset B$

$Q=\{X33，X36，X39，X42，X45，X48\} \subset B$

除了集合 M 中的元素 X02 为随机抽取所得，集合 M、P、N、Q 中其余的元素均是采用等距抽样[25]的方式得到。

以下的两组文本分类实验使用了不同的特征。每一组实验，根据训练集、测试集以及特征的不同，又分别进行了 6 个实验。

实验 1、实验 4、实验 7 和实验 10 使用冰心后期 19 篇与鲁迅的同样以现实生活为题材的 25 篇小说（$B \cup C$）作为训练集，使用冰心前期的 16 个文本（E）作为测试集。其目的在于：测试样本与训练样本中 B 数据集为同作者但不同时期，测试样本与训练样本中 C 数据集

为同时期但不同作者，若冰心小说语言风格历时变化较小，那么测试样本的分类结果将划分到冰心的那一类中。

实验 2、实验 5、实验 8 和实验 11 分别是实验 1、实验 4、实验 7 和实验 10 的参照对比实验，训练样本和测试样本均包括了冰心前期和后期的文本，目的在于观察冰心的小说在不同的语言特征集上风格是否一致。

实验 3、实验 6、实验 9 和实验 12 均使用冰心与鲁迅同时期以现实生活为题材的文本作为训练样本，构造出的分类器按照两者语言风格的不同将其划分为两类。使用鲁迅以神话史实为题材的文本作为测试样本进行测试，其预设在于：测试样本与训练样本的题材均不一致，若在不同题材的文本中鲁迅的语言风格具有一致性，那么测试样本的分类结果将划分到鲁迅的那一类中。

5.1 基于标点特征的文本分类

利用支持向量机分类算法，分别使用标点（共 13 个标点符号，见表 1）和二元标点（表示连续两个句子的标点符号）作为特征，对不同的训练集和测试集分别进行实验，得到实验结果见表 4。

表 4　基于标点符号的测试集分类结果

	实验 1	实验 2	实验 3	实验 4	实验 5	实验 6
特征	标点			二元标点		
训练集	B∪C	M∪P∪C	A∪C	B∪C	M∪P∪C	A∪C
测试集	E	N∪Q	D	E	N∪Q	D
作者	冰心	冰心	鲁迅	冰心	冰心	鲁迅
正确率 P	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
召回率 R	56.3%	87.5%	75.0%	87.5%	93.8%	87.5%

在实验 1 至实验 6 训练集的分类结果中，冰心和鲁迅的文本在超平面上得到了完全二类划分。测试集的分类结果中，由于在各个实验中测试样本只有一类（即实验 1、2、4、5 只有属于冰心的分类文本，实验 3 与实验 6 中只有属于鲁迅的分类文本），正确率 P 恒为 100%。而召回率对于分类结果的意义更为重要。

从分类的结果来看：

（1）实验 1 与实验 4 是对冰心历时风格的考查，其对比实验 2 与实验 5 的召回率均达到了 85%以上，而实验 1 的文本分类召回率只有 56.3%，说明冰心小说在标点这一语言特征上的风格变化较大。

（2）实验 3 与实验 6 是对鲁迅不同题材文本语言风格的考查，实验 3 的分类结果为 75%，可以说鲁迅不同题材小说在标点这一语言特征上的风格有所不同。

（3）实验 1 与实验 4、实验 2 与实验 5、实验 3 与实验 6 的三组实验中，基于二元标点特征的分类结果相对基于标点特征的分类结果均有所提升。这表明二元标点的特征比标点特征更能体现上下文信息与作者的语言风格。而实验 4 与实验 6 均获得了 87%左右的召回率，在很大程度上可以认为冰心在二元标点特征的使用上历时变化较小，鲁迅不同题材的小说在二元标点特征的使用上差异较小。

对原始文本中语言特征使用情况进行分析，能够进一步说明分类器判定结果的合理性。以实验 4 与实验 6 基于二元标点特征的分类判定情况为例，实验 4 将冰心的 X11（《一个奇异的梦》）与 X21（《爱的实现》）两个文本判定为鲁迅的作品，实验 6 将鲁迅的 Z33 文本（《起死》）判定为冰心的作品。《一个奇异的梦》这篇小说以主人翁与主人翁臆想出来的“社会”这个人物的对话为主线，描写青年人内心的疑惑与挣扎。其中对话描写部分仅使用了逗号加前双引号，而冒号加前双引号使用频率为 0。《爱的实现》这篇小说多为对事件发生与进展的叙述，仅有一处对主人公心理活动的描写使用了引号，且前接标点为逗号。这两

篇小说均未体现出冰心小说二元标点使用的重要特点。而鲁迅的小说《起死》与以往创作手法也有很大的不同。从形式到内容，它更像是一部短剧。赵光亚[36]认为《起死》是一篇现代戏剧体的小说。文中以人物对话为主体，分别由“‘说话人’——”引出，场景、人物动作、神情均使用括号标出。

5.2 基于词类特征的文本分类

利用支持向量机分类算法，分别使用词类和二元词类（表示连续两个词的词类）作为特征，对不同的训练集和测试集分别进行实验，得到实验结果见表 5。

表 5　基于词类的测试集分类结果

	实验 7	实验 8	实验 9	实验 10	实验 11	实验 12
特征	词类			二元词类		
训练集	B∪C	M∪P∪C	A∪C	B∪C	M∪P∪C	A∪C
测试集	E	N∪Q	D	E	N∪Q	D
作者	冰心	冰心	鲁迅	冰心	冰心	鲁迅
正确率 P	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
召回率 R	68.8%	87.5%	87.5%	87.5%	100.0%	100.0%

实验 7 至实验 12 训练样本的分类结果显示，冰心与鲁迅的文本被完全二分。

测试结果中：

（1）分别对比实验 7 与实验 10、实验 8 与实验 11、实验 9 与实验 12 发现，从词类到二元词类，冰心文本与鲁迅文本分类的召回率均有所提升。

（2）基于词类的实验，实验 7 冰心文本分类的召回率低于 70%，其对比实验 8 的召回率为 87.5%，说明冰心小说在词类这一语言特征上的写作风格历时变化较大；鲁迅文本分类的召回率高于 85%，说明鲁迅不同题材的小说中词类使用风格基本保持一致。

（3）基于二元词类特征，冰心与鲁迅文本分类的召回率都比较高，说明两位作者在这一特征上的写作风格基本稳定。

6　结论

本文以鲁迅和冰心的短篇小说文本为语料，从计量风格学的角度出发，使用基本统计、层次聚类以及基于 SVM 的文本分类方法，定量与定性相结合地分析了鲁迅与冰心短篇小说的语言风格。

通过基本统计发现，鲁迅小说的篇幅长度变化较大；段落平均长度较小；短句居多，充满力量和起伏；多使用分号进行表示句内停顿。冰心小说篇幅相比与鲁迅的较短；段落平均长度较大；句子偏长，多使用逗号表示句内停顿；对人物对话的描写较为中规中矩，多使用冒号加引号的结构。

通过计算 K 特征值以及假设检验发现，鲁迅小说的词汇丰富度要高于冰心小说。前 1000 个高频词的层次聚类结果表明两者在词语的选用上存在差异性，体现了两位作者对社会现实的不同关注点。

通过文本分类实验发现，基于二元标点和二元词类特征的分类召回率普遍要高于标点和词类特征。冰心在小说历时创作的过程中，标点以及词类的使用风格发生转变，二元标点以及二元词类的使用风格变化较小。鲁迅在不同题材的小说创作中，仅标点的使用风格转变略大，词类、二元标点、二元词类的使用风格基本保持平稳。

参考文献：

[1] 沈雁冰.《读<呐喊>一文》[A].《鲁迅研究学术论著资料汇编》（第一卷）[C].北京：中国文联出版公司,1989.

[2] 郑振铎.《呐喊》[A].《鲁迅研究学术论著资料汇编》（第一卷）[C].北京：中国文联出版公司,1989.

[3] 李长之. 鲁迅批判[M]. 北京：北京出版社, 2003.
[4] 巴人.《鲁迅的创作方法》[A].《六十年来鲁迅研究论文选（上）》[C]. 北京：知识产权出版社,2010:251-268.
[5] 胡云翼.新著中国文学史[M].上海：华东师范大学出版社,2004.
[6] 郎学初.冰心小说的诗化特征[J].广西社会科学, 2007 (1): 100-103.
[7] 林荣松. 冰心小说文体新论[J]. 宁德师专学报: 哲学社会科学版, 2002 (2): 33-38.
[8] 杨清. 冰心作品语言风格变化探析[J]. 福建工程学院学报, 2010, 8(2): 168-172.
[9] 李廷姬. 论冰心的小说创作[D]. 西北大学, 2000.
[10] De Vel O, Anderson A, Corney M, et al. Mining e-mail content for author identification forensics[J]. ACM Sigmod Record, 2001, 30(4): 55-64.
[11] Burrows J F. Word-patterns and story-shapes: The statistical analysis of narrative style[J]. Literary and linguistic Computing, 1987, 2(2): 61-70.
[12] Argamon S, Levitan S. Measuring the usefulness of function words for authorship attribution[C]//ACH/ALLC. 2005.
[13] Stamatatos E, Fakotakis N, Kokkinakis G. Automatic text categorization in terms of genre and author[J]. Computational linguistics, 2000, 26(4): 471-495.
[14] 王景丹. 从句频分析看八位剧作家的风格异同[J]. 修辞学习, 2004 (4): 28-29.
[15] Baayen H, Van Halteren H, Tweedie F. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution [J]. Literary and Linguistic Computing, 1996, 11(3): 121-132.
[16] Carole E. Chaski. Empirical evaluations of language-based author identification techniques. Forensic Linguistics. 2001, 8(1),1-65.
[17] Biber, D. Dimensions of register variation: A cross-linguistic com-parison. Cambridge, UK: Cambridge University Press,1995.
[18] 施建军. 关于以《红楼梦》120 回为样本进行其作者聚类分析的可信度问题研究[J]. 红楼梦学刊,2010(5).
[19] Diederich J, Kindermann J, Leopold E, et al. Authorship attribution with support vector machines[J]. Applied intelligence, 2003, 19(1-2): 109-123.
[20] 鲁迅.鲁迅全集[M]. 北京：人民文学出版社, 2005.
[21] 冰心, 乐齐著. 冰心小说[M]. 浙江：浙江文艺出版社，2009.06.
[22] 冰心, 卓如.冰心全集[M]. 福建：海峡文艺出版社, 1994.
[23] 廖秋忠. 篇章与语用和句法研究[J]. 语言教学与研究, 1991 (4): 16-44.
[24] Yule, G.U. The Statistical Study of Literary Vocabulary. Cambridge University Press, 1944.
[25] 张卫国, 语言学. 汉语研究基本数理统计方法[M]. 中国书籍出版社, 2002.
[26] 张敏强. 教育与心理统计学[M]. 北京：人民教育出版社, 2010.
[27] 陈亚丽. 基础写作教程[M]. 北京：北京大学出版社,2008.07:312-315.
[28] Roberts S J, Everson R, Rezek I. Maximum certainty data partitioning[J]. Pattern Recognition, 2000, 33(5): 833-839.
[29] 方开泰. 聚类分析（Ⅰ)[J]. 数学的实践与认识, 1978, 1: 66-80.
[30] Salton G, Yu C T. On the construction of effective vocabularies for information retrieval[C]//ACM SIGPLAN Notices. ACM, 1973, 10(1): 48-60.
[31] 刘颖. 统计语言学[M]. 北京:清华大学出版社, 2014:153.
[32] 刘晓亮, 丁世飞, 朱红, 等. SVM 用于文本分类的适用性[J]. 计算机工程与科学, 2010, 32(6): 106-108.
[33] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程, 2004, 30(10): 6-9.
[34] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[J]. 2003.
[35] 孙微微, 刘才兴, 田绪红. 训练集容量对决策树分类错误率的影响研究[J]. 计算机工程与应用, 2005, 41(10): 159-161.
[36] 赵光亚. 鲁迅小说《起死》的文体选择与重构[J]. 南京师范大学文学院学报, 2012 (1): 62-67.
作者简介：冷婷（1990——），女，硕士研究生，主要研究领域为语料库语言学。Email: lengting1990@163.com；刘颖（1969——），女，博士，教授，主要研究领域为语料库语言学、计算语言学、自然语言处理和机器翻译。Email: yingliu@mail.tsinghua.edu.cn（通讯作者）。

照片：

 （冷婷）

 （刘颖）