基于聚类和分类的金庸与古龙小说风格分析*

肖天久，刘颖

（清华大学中国语言文学系，北京，100084）

摘要：本文以金庸与古龙的小说作为语料，从计算风格学的角度考察二人的风格差异。对比了两人小说的文本从众性、句子破碎度，同时，使用文本聚类的方法对词和词类的 N 元文法，标点符号的 N 元文法以及多种特征的总体情况进行了考察，还使用主成分分析和文本分类对八种特征从总体上进行了比较，结果证实金庸与古龙小说风格存在较大差异：金庸小说从众性大于古龙，较多使用俚语方言，口语性更强，同时在语法结构、短语结构、文本节奏以及文本可读性和语言变化程度也有较大的差异。

关键词：计算风格学；N 元文法；聚类；分类；句子破碎度

中图分类号：TP391　　**文献标识码：**A

A Stylistic Analysis of Jin Yong's and Gu Long's Fictions Based on Text Clustering and Classification

Xiao Tianjiu, Liu Ying

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

Abstract: Based on the fictions written by Jin Yong and Gu Long, this paper analyzes the sentence fragmentation and text conformity from the perspective of computational stylistics. The twelve texts are clustered using n-gram of words, n-gram of part of speech, n-gram of punctuations and six multiple features as features. Besides, principal components analysis and classification are conducted with eight multiple features. The results of experiments show that there exist great differences on the styles between Jin Yong's and Gu Long's fictions: Jin Yong's fictions are more colloquial and conform than Gu Long's; Jin Yong use more words and idioms from dialects and slang while the expressions in Gu Long's fictions are more formal. What's more, there are differences between the two authors' fictions on the syntactic structures, phrase structures, rhythms, readabilities and language variation.

Key words: computational stylistics; n-gram; clustering; classification; sentence fragmentation

1 引言

计算风格学是数理语言学的一个分支。不同于传统的风格学研究主要以读者内省为基础，通过对文中的句子、词语，乃至篇章的感悟、归纳来提炼作者和作品的风格,计算风格学主要是以定量的方式利用文本中可以量化的语言结构特征来对文本风格和作者写作习惯进行研究，其理论基础是认为文本的语言结构特征表现了作者个人在写作活动中的言语特征，是作者个人风格不自觉的深刻反映,并且这些特征又可以在一定程度上通过数量特征来进行刻画。

从计算风格学的角度对文本的语言风格进行考察，有两个最为重要的问题：一是语言特征的选择，这些语言特征一般是要求可以量化并且是稳定出现的；二是研究方法的选择，即是统计方法和数学模型的选择。

目前已经提出可以反映文本风格的语言结构特征可以归结为六个层面：字符、词汇、句子、段落、语法、语义等。字符层面主要包括大小写字母[1]、特殊符号（如标点符号[2]）、数字和空格[3]等；词汇方面特征的研究相对是最为成熟的，包括词汇丰富度[4]，功能词[5]，

* **收稿日期：**　　**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）

高频词[6]等有意义的词；句子层面主要有句长、平均句长[7]等特征；段落层面，主要有段落长度[8]；语法层面，包括词类[9]，以及依存语法等等；语义层面，主要有基于 HowNet 的语义类[10]等 。

对特征的研究和分析最早仅简单统计某些特定语言特征的频率、分布，随后引入了 t 检验、χ^2检验[11]等统计方法；后来主成分分析[6]、相关性分析、因子分析等特征分析方法被引入；目前利用文本聚类[12]、文本分类[13]来对文本和作品的风格进行考察比较多。

从金庸与古龙方面来说，金庸与古龙均为新派武侠小说家的代表人物，两人均创作了大量具有深远影响的小说。前人对两人小说的风格对比多从文学方面[14][15][16]，从计算语言学方面，仅有刘颖等[17]从虚词、词类、标点、部分实词等角度对二者进行了比较，认为金庸更关注家国天下的责任，而古龙更关注江湖人的个体感受，并且金庸小说可读性较古龙要弱，古龙用词更具变化；金庸更善于武功招式的描写，而古龙则更倾向于环境气氛的渲染。

从上面可以看出，尽管目前计算风格学的研究在国内外都相对比较成熟，但是其在金庸与古龙的小说的研究上相对是比较少的。同时，作为新派武侠小说家的代表，二人的作品是极好的语言研究的材料。本文从计算风格学的角度出发，继续对金庸和古龙的小说风格进行考察和分析。

2 语料选择

本文选取金庸与古龙各自最具有代表性的六部小说建立语料库，总规模超过 980 万字。从标点符号、句子破碎度、文本从众性、N 元文法等方面对二者进行分析和比较。

选取的古龙与金庸各自六部小说分别为：

古龙：《大旗英雄传》、《武林外史》、《绝代双骄》、《楚留香传奇系列全集》（以下称《楚留香传奇》）、《小李飞刀系列全集》（以下称《小李飞刀》）和《陆小凤传奇系列全集》（以下称《陆小凤传奇》）；

金庸：《射雕英雄传》、《神雕侠侣》、《倚天屠龙记》、《天龙八部》、《笑傲江湖》和《鹿鼎记》。

这十二部小说均是古龙和金庸最具有代表性的著作，并且均是成熟时期的作品，高度代表了金庸与古龙小说语言风格。

对所选十二部小说的总字数、总词数进行统计，得到表 1。

表 1 金庸与古龙所选小说的总字数、总词数统计

古龙	总字数	总词数	金庸	总字数	总词数
大旗英雄传	510 601	382 424	射雕英雄传	758 780	568 804
武林外史	562 900	425 953	神雕侠侣	810 591	614 360
绝代双骄	647 283	488 032	倚天屠龙记	817 568	616 293
楚留香传奇	993 335	728 461	天龙八部	1 021 443	762 806
小李飞刀	1 110 332	819 356	笑傲江湖	827 171	611 048
陆小凤传奇	793 209	575 212	鹿鼎记	1 021 369	741 469
总计	4 617 660	3 419 438	总计	5 256 922	3 914 780

可以发现，在所选的十二部小说中，古龙的《大旗英雄传》、《武林外史》、《绝代双骄》三部的篇幅较短，其余的篇幅均较长。这其中需要指出的是，古龙小说中，《楚留香传奇》、《小李飞刀》、《陆小凤传奇》为系列小说，而其余九部小说为单篇小说。

3 文本从众性

在使用 **ICTCLAS**[1]分词时，词的概念是非常宽泛的，既包括传统意义上的词，也包括传统意义上不被认为是词的一些“标准件”语言素材，包括词组、成语、歇后语、谚语、警句、

[1] http://ictclas.nlpir.org/.

名言、古诗句等等，文本的从众性就是考察作者使用这些语言素材的熟练程度与个人偏好程度。其一般使用“聚类度”的概念来考察文本的从众性。所谓的聚类度，指的就是文本词的成词率与词的平均长度的乘积，二者分别反映文本从众性的广度和强度[19]。

文本成词率，是指文本中所有词的长度（即总字数）与所有字符的长度（即总字符数）的比率[18]。图 1 即为各个文本的成词率。

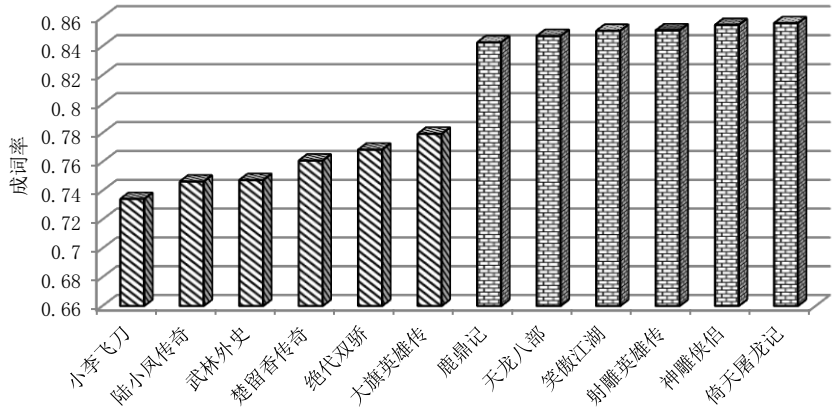

图 1 金庸与古龙小说的成词率

如图 1 所示，横坐标为全部十二部小说，纵坐标为每部小说的成词率。并且对全部十二部小说的成词率按照从小到大的顺序进行排列。可以发现，在全部十二部小说中，成词率最低的是古龙的《小李飞刀》，最高的是金庸的《倚天屠龙记》，并且还可以发现，古龙的小说的成词率均低于金庸。

在此基础上计算各部小说的聚类度，所得结果如图 2 所示。

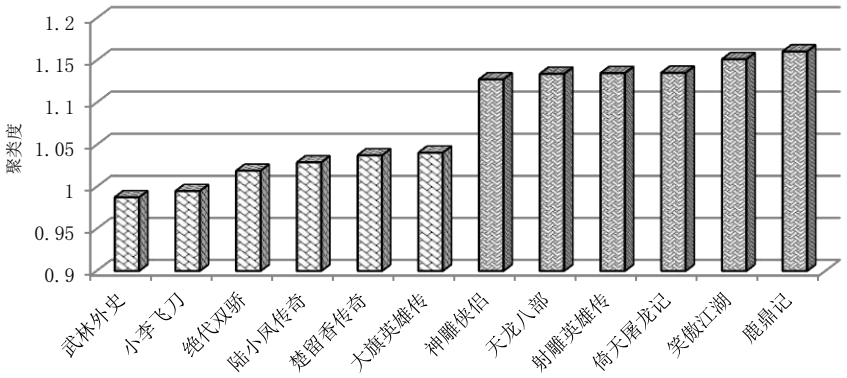

图 2 金庸与古龙小说的聚类度

如图 2 所示，横坐标为全部十二部小说，纵坐标为各部小说的聚类度。全部十二部小说的聚类度按照从小到大顺序进行排列。可以看出，在全部十二部小说中，古龙的《武林外史》聚类度最低，而金庸的《鹿鼎记》的聚类度最高，并且古龙小说的聚类度均低于金庸的小说，反映出古龙小说整体从众性要低于金庸。

这种情况的出现，是与古龙与金庸二人的语言风格和个人背景分不开的。古龙受西方文学尤其是大仲马、毛姆、海明威、杰克·伦敦等人的影响较大，其语言风格一直在追求散文化和诗化，尤其是《天涯·明月·刀》中更是直接用散文诗化的语言采用蒙太奇的手法描写出一个完整的故事，引起了巨大的争议，同时，由于其古典文化修养相较于金庸要单薄，因此，在古龙的小说中，一方面，引经据典，如成语、熟语、诗句、名言等引用较少，另一方

面，为了保证语言的文雅，普罗大众的方言俚语也较少使用，从而导致其成词率较低，也就影响了其文本的从众性。金庸则不同，一方面，深厚的古典文化修养，保证了金庸引经据典的随心所欲、信手拈来，乃至创作诗词的得心应手；另一方面，金庸又并不避讳使用一些街头巷尾的方言俚语，反而为了使人物形象更为突出，大量使用符合其身份的语言，比如，在《鹿鼎记》中，金庸为了塑造韦小宝这个底层小混混的形象，大量使用“乖乖龙的冬，猪油炒大葱”，“辣块妈妈”等俚语，其出身妓院，又对“贼王八”、“路倒尸”、“臭乌龟”、“挨千刀”等粗言鄙语信手拈来。再比如，为了突出人物的地域特点，其广泛使用地方方言，如四川话中的“格老子”、“龟儿子”，广东话中的“班契弟”，苏州话中的“阿是”、“啥事体”等等。

4 句子破碎度

句子的破碎度，是指“一句话中的停顿次数，即一句话的零散程度”[19]。一般认为，书面性越强的文本，语句越流畅，句内停顿较少，破碎度较低；反之，口语性越强的文本，句中的插入语越多，常出现停顿，破碎度越高。破碎度的计算公式如下：

$$\text{句子破碎度} = \frac{\text{句中停顿总次数}}{\text{语料中总句数}} \quad (1)$$

黄伯荣、廖序东认为，“点号主要表示句中的各种停顿”，其将点号分为句末点号和句中点号，如表 2 所示[20]。

表 2 点号列表

类别	句末点号			句中点号			
形状	。	？	！	，	、	；	：

根据表 2，我们统计这 7 种点号在各部小说中出现的总次数，并且计算各部小说中句子的破碎度。结果如图 3 所示。

如图 3，纵坐标为全部十二部小说，横坐标为各部小说的句子破碎度，并对十二部小说的破碎度按照从大到小进行排序。可以发现，古龙《小李飞刀》的句子破碎度最低，而金庸的《神雕侠侣》的句子破碎度最高，虽然由于受到同一种文体的影响，二人小说总体差异不不是非常大，但金庸小说的句子破碎度仍然均高于古龙，可以认为，金庸小说的口语性更强，而古龙小说语言的书面性则较强。这并不奇怪，金庸小说一般段落较长，句中插入成分较多；而古龙则段落较短，一般一句话为一段，因此停顿也少。

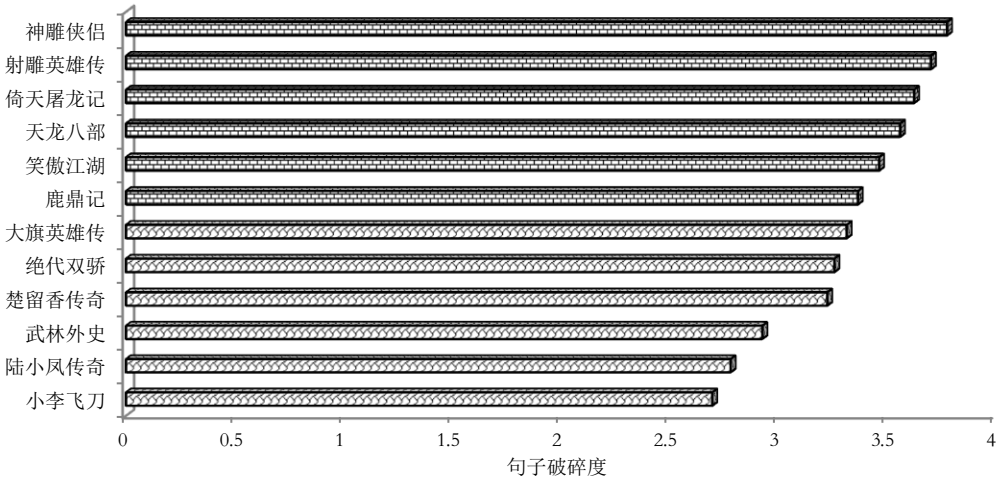

图 3　金庸与古龙小说句子破碎度

5 基于文本聚类的风格分析

文本聚类是将文本集合分组成多个类，使得同一个类中的文本具有较高的相似性，而不同类中的文本内容差异较大。这一过程是无监督的学习过程。

在聚类前，本文使用公式（2）对数据进行归一化处理。

$$X_{ik}^{*} = \frac{X_{ik}}{\sqrt{\sum_{k=1}^{n} X_{ik}^{2}}} \quad (2)$$

X_{ik}^{*}为第 i 个文本在第 k 维空间中的新值，即归一化后的数据，X_{ik}是第 i 个文本本在第 k 维空间中的原始值，$\sum_{k=1}^{n} X_{ik}^{2}$为样本在 n 维空间中所有值的平方和。[21]

本文将使用欧氏距离[18]和 KL 散度两种方式计算文本之间的相似性，采用离差平方和法来合并不同的类，并采用自下而上的凝聚式层次聚类。

5.1 基于欧式距离的层次聚类

5.1.1 基于词的 N 元文法的文本聚类

N 元文法，指的是由 N 个字、词、词类或者特殊符号（如标点符号）组成的序列。对词的 N 元文法来说，当 N=1 时，为一元文法，相当于词表，给出的是文本中使用的所有词；当 N=2 时，为二元文法，给出的是文本中邻接的两个词的使用情况；当 N=3 时，为三元文法，给出的是连续三个词在文本中的使用情况。二元文法和三元文法可以反映文本中短语结构（即邻接词语组合）情况。

本文对词的二元到三元文法进行聚类，分别统计词的二元到三元的前 2000 个词序列在各文本中的出现次数，并进行归一化处理，然后分别进行层次聚类，结果如图 4~图 5 所示。

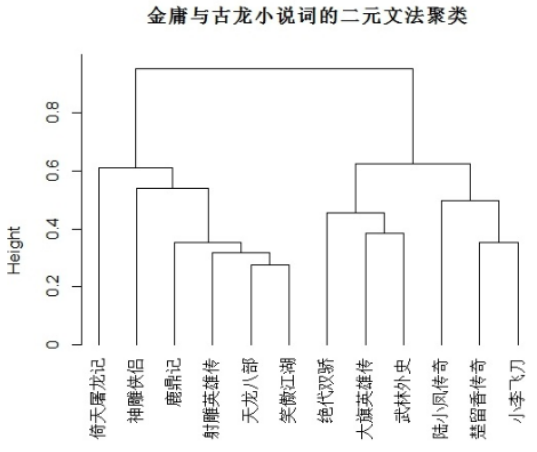

图 4 金庸与古龙小说词的二元文法聚类

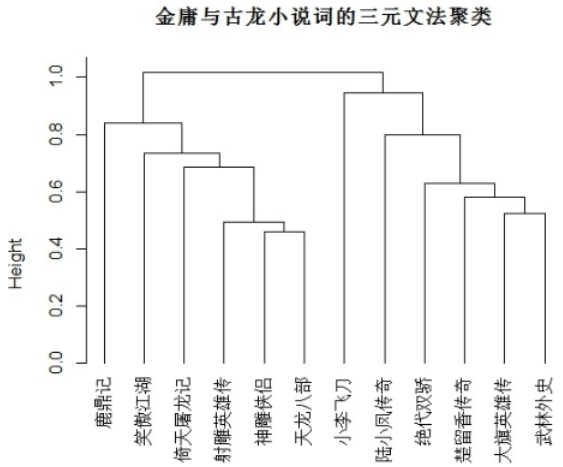

图 5 金庸与古龙小说词的三元文法聚类

如图 4~图 5 所示，横坐标为全部十二部小说，纵坐标为类与类之间的欧氏距离，可以发现，金庸的六部小说始终聚为一类，而古龙六部小说始终聚为另一类，因此，从小说的短语结构上，金庸和古龙具有显著的差异。

5.1.2 基于词类的 N 元文法的文本聚类

词类的 N 元文法模型，指的是以词类为单位的词类组合，当 N=1 时，为一元文法，给出的是文中词类列表，当 N=2 时，为二元文法，给出的是邻接两种词类在文中使用的情况；当$N = k(k = 2，3，4，\cdots)$时，给出的是连续 k 种词性在文中出现的情况。当$N \geq 2$时，其反映的是文本的语法结构。由于词类数目要远远小于词的数目，因而其可靠程度较词的 N 元文法模型要高，更能反映文本的风格。

本文对词类的二元到五元文法进行聚类，分别统计词类二元、三元文法前 500 个词类序列，四元文法前 1000 个词类序列，五元文法前 1500 个词类序列在每个文本中出现的次数，并对统计数据进行归一化处理，然后分别进行层次聚类，得到结果如图 6~图 9 所示。

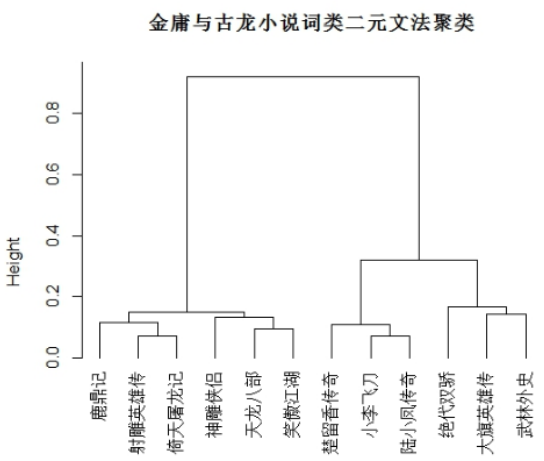

图 6 金庸与古龙小说词类二元文法聚类

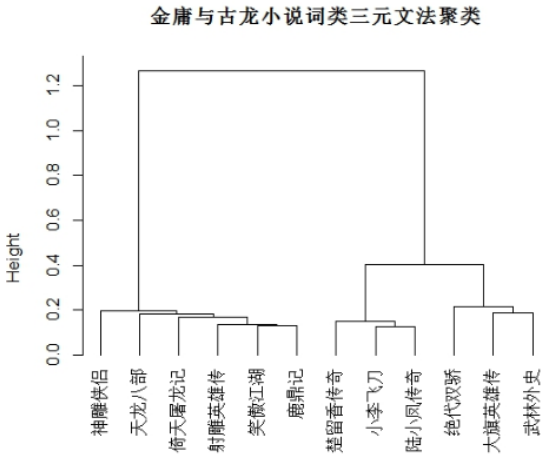

图 7 金庸与古龙小说词类三元文法聚类

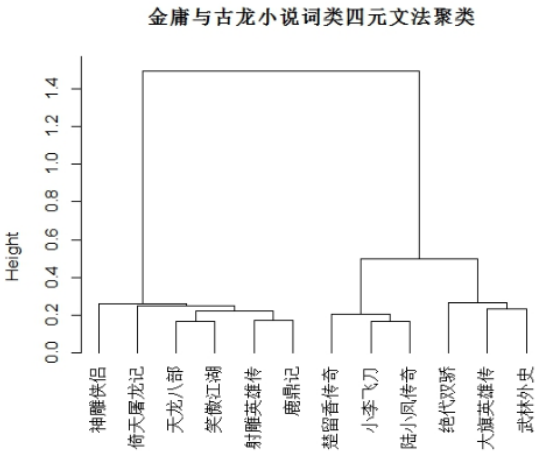

图 8 金庸与古龙小说词类四元文法聚类

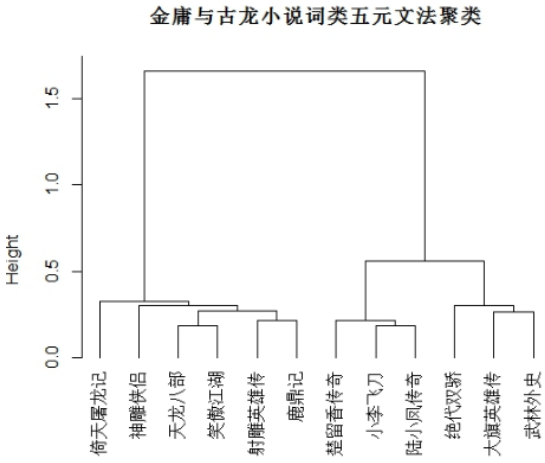

图 9 金庸与古龙小说词类五元文法聚类

如图 6~图 9 所示，横坐标为全部十二部小说，纵坐标为类与类之间的欧氏距离。可以发现，从词类的二元文法到五元文法，金庸小说始终聚为一类，而古龙的六部小说聚为另一类，反映出金庸与古龙小说的语法结构存在较大差异。

进一步我们列出词类的二元文法到五元文法的前十位如表 3 所示：

表 3 金庸与古龙小说中词类 N 元文法的前十位

二元文法		三元文法		四元文法		五元文法	
古龙	金庸	古龙	金庸	古龙	金庸	古龙	金庸
/d/v	/d/v	/d/v/v	/d/v/v	/d/d/v/v	/d/v/d/v	/r/d/d/v/v	/v/u/m/q/n
/v/v	/v/v	/d/d/v	/n/d/v	/r/d/d/v	/n/d/v/v	/n/d/d/v/v	/n/d/v/d/v
/v/r	/n/v	/n/d/v	/v/d/v	/n/d/d/v	/d/v/v/v	/u/n/d/d/v	/d/v/d/v/v
/v/u	/v/n	/r/d/v	/v/v/v	/v/r/d/v	/v/r/d/v	/v/r/d/d/v	/v/d/v/d/v
/v/n	/v/r	/d/v/r	/d/d/v	/d/v/v/r	/v/d/v/v	/d/d/v/v/r	/d/v/r/d/v
/u/n	/v/u	/v/v/r	/n/v/v	/r/d/v/v	/v/v/d/v	/d/v/r/d/v	/v/v/r/d/v
/n/d	/n/d	/v/d/v	/v/r/v	/d/v/d/v	/v/v/v/v	/v/r/d/v/v	/v/r/d/v/v
/n/v	/r/v	/v/r/v	/v/v/r	/n/d/v/v	/d/d/v/v	/r/d/v/v/r	/d/v/v/d/v
/d/d	/v/d	/d/v/u	/v/v/n	/u/n/d/v	/v/n/d/v	/v/v/r/d/v	/n/d/v/v/v
/r/v	/n/n	/v/v/v	/v/n/v	/d/v/v/v	/d/v/v/r	/d/v/v/r/v	/m/q/n/d/v

如表 3 所示，其中，n 表示名词，v 表示动词，d 表示副词，r 表示代词，m 表示数词，

q 表示量词，u 表示助词。

可以发现，在二元到五元文法前十位词类组合中，除了相同组合的频率排序有较大差异外，二元文法有两组是不同的；三元文法有三组不同；四元文法有五组不同，而在五元文法中则有八组是不同的。反映出随着词类元数的增加，金庸与古龙小说的语法结构差异越大。

5.2 基于 KL 散度的层次聚类

KL 散度，是 Kullback-Leibler 散度（Kullback-Leibler Divergence）的简称，也叫做相对熵（Relative Entropy）。它衡量的是相同事件空间里的两个概率分布的差异情况。其意义是：对于归一化后的文本向量$\mathrm{P}(X_1, X_2, \cdots, X_n)$，和 $\mathrm{Q}(Y_1, Y_2, \cdots, Y_n)$，向量特征值的总和均为 1，且对于任何 i 都满足$X_i > 0$及$Y_i > 0(1 \le \mathrm{i} \le \mathrm{n})$。我们利用金明哲改进后的 KL 散度计算公式来计算两个文本之间的相似度，两个文本之间的 KL 散度越小，其相似性越大。[22]

$$\mathrm{KLD}(\mathrm{P},\mathrm{Q}) = \frac{1}{2}\sum_{i=1}^{n}[\mathrm{P}(x_i)\log\frac{2P(x_i)}{Q(x_i)+P(y_i)} + Q(y_i)\log\frac{2Q(y_i)}{Q(x_i)+P(y_i)}] \quad (4)$$

5.2.1.基于标点符号的 N 元文法的文本聚类

标点符号是书面语的有机组成部分，在文本中使用频率很高。每一个标点符号都有自己独特的作用，尤其是语法作用，可以看成是另外一种形式的虚词。[23] 同时，标点符号是句子组织结构的一个重要表现。具有停顿意义的标点符号也是构成文本节奏的重要因素。[24]

标点符号的 N 元文法，是以标点符号为单位的符号组合。当 N=1 时，为一元文法，给出的是文中所有的标点符号；当$\mathrm{N} = \mathrm{k}(\mathrm{k} = 2, 3, 4, \cdots)$时，给出的是连续 k 个标点符号在文中出现的情况。当$\mathrm{N} \ge 2$时，其反映的是文本的节奏和句子的组织结构。

本文利用标点符号二元~五元文法对金庸和古龙小说聚类，结果如图 10~13 所示。

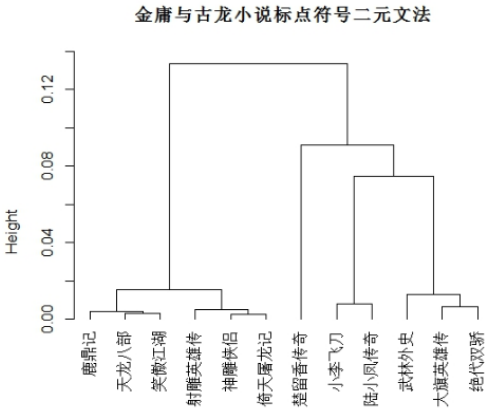

图 10 金庸与古龙小说标点符号二元文法聚类

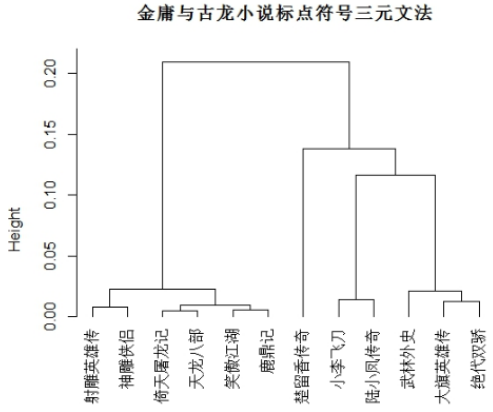

图 11 金庸与古龙小说标点符号三元文法聚类

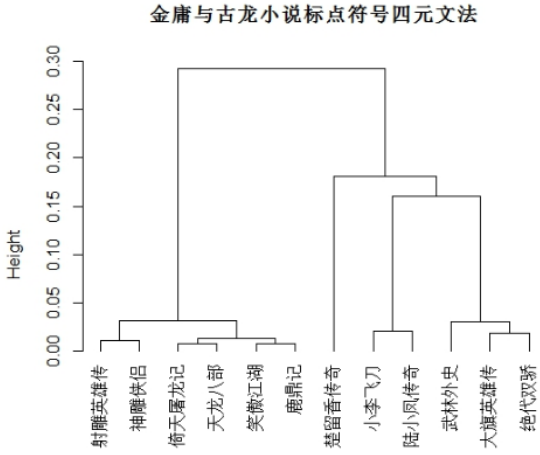

图 12 金庸与古龙小说标点符号四元文法聚类

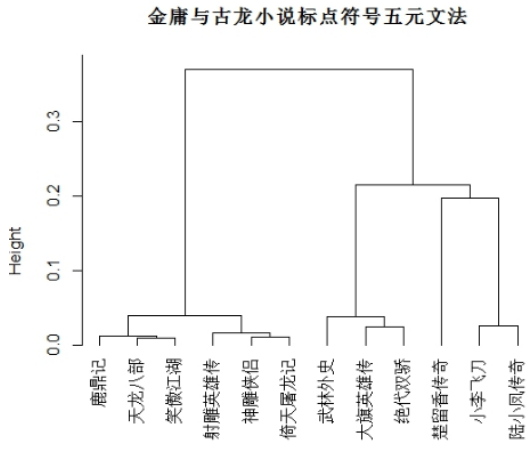

图 13 金庸与古龙小说标点符号五元文法聚类

如图 10~13 所示，横坐标为金庸与古龙全部十二部小说，而纵坐标为文本间的 KL 散度。可以发现，从标点符号的二元文法到五元文法，金庸小说始终聚为一类，而古龙小说则聚为另一类，二者截然分开，可以看出，从标点符号的二元到五元文法来看，金庸与古龙小说是不同的，反映出二者文本节奏的差异。

金庸和古龙小说中标点符号的 N 元文法频率最高的前十位如表 4 所示。

表 4 金庸与古龙小说中标点符号 N 元文法频率最高的前十位

二元		三元		四元		五元	
古龙	金庸	古龙	金庸	古龙	金庸	古龙	金庸
，，	，，	，，，	，，，	，，，，	，，，，	，。”：“	，，，，，
：“	，。	”：“	，，。	，，，。	，，。，	，，，，，	，，。，，
，。	。，	，，。	，。，	。”：“	，。，，	，，，。，	，，，。，
“，	：“	：“，	。，，	”：“，	，，，。	，，，，。	，。，，，
。”	“，	，。，	：“，	，，。，	。，，，	，，。，，	，，，，。
”：	”，	，。”	，。”	：“，，	：“，，	：“，。”	。，，，，
。，	。”	，：“	”：“	，。，，	，：“，	，。，，，	，。”：“
”，	，：	。，，	，：“	，：“，	”：“，	。”：“，	，。，，。
，：	”：	。”：	”，，	，，。”	。”：“	”：“，，	。”：“，
？”	？”	“，，	“，，	，。”：	，，。”	：“，，，	，，：“，

如表 4 所示，我们可以发现：

从标点的二元文法到五元文法，金庸使用最高的都是一逗到底；而古龙则不然，其在五元文法上出现了其他标点；同时除了二元文法中有问号出现以外，在金庸与古龙小说中三元文法到五元文法，出现的标点仅有逗号、冒号、引号、句号四种，反映出在金庸与古龙小说中对话描写都是非常多的。

从标点符号的二元组合来看，金庸与古龙小说前十位的标点组合是一致的，差异在于使用频率排序：金庸前三位标点均是逗号、句号等停顿性标点，而古龙在使用频率第二位便出现了表示对话的冒号和引号，而且出现的两三次前引号相对于金庸使用频率排序都要靠前，反映出，相比于金庸，古龙小说中对白描写较多。事实上，在古龙作品中，常常使用大段的对白推进情节，而且其中不乏是古龙以局外人的身份在自问自答以表达自己的观点或是构成散文诗的结构，这种写作风格导致了文本中引号的大量出现。同时，需要说明的是，古龙小说中的对白很多与书面语相差并不大，甚至故作庄严地使用起散文式的句子，这从一定程度上导致了其虽然对白较多，但是其口语化的程度却不是非常高。

从三元组合来看，前十位中，相同有九组组合，仅出现次序不一致；反映出二者文本节奏和句子结构差异，同时，古龙小说中出现前引号的标点组合相对于金庸的排序更是大大靠前，进一步证实了古龙小说中对白较多。

从四元、五元组合来看，差异逐渐扩大：不仅标点不同的符号组合在五元文法中上升到三组，而且排序的差异也逐渐递增。

5.2.2.基于多特征的文本聚类

我们在前面单一特征的基础上，同时使用多个特征来考察金庸与古龙六部小说的关系。我们选取了标点符号占比、平均词长、平均句长、平均段落长度、句长离散度、词长离散度 6 个特征。其中：

标点符号占比是指文本中标点符号的总数占字符总数的百分比；

句长离散度由公式（5）求得：

$$D_s = \sqrt{\frac{1}{n}\sum(S_i - S_0)^2} \quad (5)$$

其中，S_i为不同的句长，S_0为平均句长。n为文本的句子总数。

词长离散度由公式（6）求得：

$$D_w = \sqrt{\frac{1}{n}\sum(l_i - l_0)^2} \quad (6)$$

其中，D_w为词长离散度，n为总词数，l_i为不同词的词长，l_0为平均词长。

这六个特征中，标点符号占比、平均词长、平均句长、平均段落长度是反映文本可读性的重要指标；而词长离散度可以反映语言变化程度，句长离散度可以反映节奏变化程度。

分别统计并计算这六个特征，并对计算结果进行归一化处理，然后对其进行文本聚类，所得结果如图 14 所示。

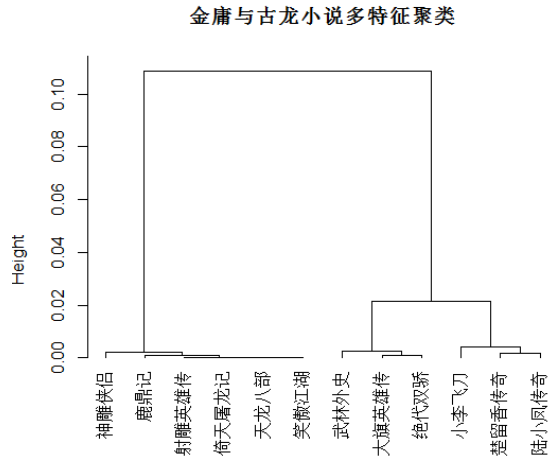

图 14 金庸与古龙多特征聚类结果

如图 14 所示，横坐标为全部十二部小说，纵坐标为不同文本之间的 KL 散度。由图可以看出，在全部十二部小说中，金庸的《射雕英雄传》、《倚天屠龙记》、《天龙八部》、《笑傲江湖》四部小说距离几乎为零，而距《鹿鼎记》稍远，距《神雕侠侣》更远。而古龙小说中，《楚留香传奇》、《陆小凤传奇》距离最近，距《小李飞刀》稍远；《大旗英雄传》、《绝代双骄》距离较近，而距《武林外史》稍远，并且三部单篇小说与三篇系列小说分别自成一类，距离较远。金庸的六部小说聚为一个大类，而古龙的六部小说聚为另一个大类，反映出，从标点符号占比、平均句长、平均词长、平均段落长、词长离散度、句长离散度这六个特征总体来看，金庸与古龙是不一样的。反映出二人在文本可读性和语言、节奏变化程度上有明显差异。

值得注意的是，一是使用多特征进行聚类之时，金庸的聚类结果与使用标点符号的 N 元文法作为特征进行聚类有较大的差异外，古龙小说的聚类结果则与标点符号的 N 元文法的聚类结果一致；二是在使用多特征进行聚类的时候，类与类之间的 KL 散度相较于标点符号的 N 元文法要小很多，反映出随着特征数的增加，两人的风格差异变小。这种情况的出现这是由于用于考察的文体是一样的，均为武侠小说，因此本身存在很大的共性特征，而随着特征数的逐渐增多，其文体的共同性便逐渐增加。

6 基于 SVM 的差异分析

我们使用如下特征作为分类的特征：

句子破碎度、形符类符比、词汇密度、单现词比率、基于字符的平均句长、叠词比率、成词率、聚类度。其中，词汇密度是指文本中词的数量占总字符数的百分比，单现词比率是指仅出现一次的词语数量占词语总数的百分比。句子破碎度、词汇密度、成词率反映的是语

言的正式程度；而形符类符比、单现词比率、叠词比率、聚类度都可以在一定程度上反映文本语言的丰富性。

在进行分类之前，我们先使用主成分分析对其进行降维。主成分是揭示大样本、多变量数据或者样本之间关系的一种方法。其核心目的就是利用降维的思想，将众多的指标转换成为少数几个主要的综合指标，从而降低观测空间的维数，以获取最主要的信息。[25]

在分别统计和计算以上各个特征之后，我们得出各个主成分的特征值和方差贡献率如表5所示。

表5 各主成分的特征值和方差贡献率

主成分	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
特征值	2.062	1.267	1.146	0.856	0.272	0.13	0.073	0.025
方差贡献率	0.531	0.201	0.164	0.092	0.009	0.002	0.001	0
方差累计贡献率	0.531	0.732	0.896	0.988	0.997	0.9992	0.99992	1

根据表5，前3个主成分的方差累计贡献率达到了89.6%，代表了全部变量89.6%的信息，达到了降维的目的。因此，我们选择前三个主成分，用于代表全部8个特征，在此基础上，计算各个文本的主成分得分，并基于此来进行分类。

我们使用SVM作为分类器并使用准确率、召回率、F值来综合评价分类性能。支持向量机（Support Vector Machine，SVM）是在统计学习理的基础上发展起来文本分类方法。其基于结构风险最小化原理，根据有限样本信息在模型的复杂性（即对特定样本的学习精度）和学习能力（即无错误的识别任意样本的能力）之间寻求最佳折中，从而获得更好的泛化能力。

我们分别使用各自五部小说为训练集，剩下一部小说为测试集；以各自四部小说为训练集，剩下两部小说为测试集；以各自三部小说为训练集，剩下三部小说为测试集，其构成如下表6所示。

表6 分类的测试集和训练集的构成：

	训练集5测试集1		训练集4测试集2		训练集3测试集3	
作者	古龙	金庸	古龙	金庸	古龙	金庸
训练集	大旗英雄传 武林外史 绝代双骄 小李飞刀 陆小凤传奇	射雕英雄传 神雕侠侣 倚天屠龙记 天龙八部 笑傲江湖	绝代双骄 小李飞刀 陆小凤传奇 楚留香传奇	倚天屠龙记 天龙八部 笑傲江湖 鹿鼎记	绝代双骄 小李飞刀 陆小凤传奇	倚天屠龙记 天龙八部 笑傲江湖
测试集	楚留香传奇	鹿鼎记	大旗英雄传 武林外史	射雕英雄传 神雕侠侣	大旗英雄传 武林外史 楚留香传奇	射雕英雄传 神雕侠侣 鹿鼎记

分类结果如表7所示。

表7 分类结果

	训练集5测试集1	训练集4测试集2	训练集3测试集3

作者		古龙	金庸	古龙	金庸	古龙	金庸
结果	古龙	1	0	2	0	3	0
	金庸	0	1	0	2	0	3
准确率		100%	100%	100%	100%	100%	100%
召回率		100%	100%	100%	100%	100%	100%
F 值		1	1	1	1	1	1

如表 7 所示，金庸与古龙小说经过主成分分析，提取前三个主成分使用 SVM 进行分类，进行的三次实验中，当训练集分别为 5，测试集分别为 1；训练集分别为 4，测试集分别为 2 以及训练集和测试集均为 3 时，其准确率和召回率均为 100%，F 值也为 1。可以看出，金庸与古龙的小说在语言正式程度和丰富性上还是有较为明显的差异的。

7 结论

本文以金庸与古龙各自六部代表作为语料，从文本从众性、句子破碎度、词和词类的 N 元文法、标点符号的 N 元文法等对二者的风格进行了考察。实验结果证实，二者在这些特征有较大的差异。

从文本从众性来说，金庸小说的从众性要大于古龙，这是由于金庸熟稔古典文化，同时又对各种市井俚语、方言等兼容并蓄，这些因素共同促使了小说名言、诗词、方言、俚语的大量出现；同时，由于金庸小说的口语性更强，句长较长，插入成分较多，导致金庸小说中句子破碎度要高于古龙。

而从词的二元文法和三元文法来看，金庸与古龙小说存在较大差异，反映出二人小说的短语结构不同;而从词类的二元文法到五元文法，金庸小说与古龙小说也是各自被聚为一类，反映出二者小说语法结构的差异，同时，从前十位的词类组合来看，从二元到五元，词类组合相同的越来越少，差异越来越明显；而从标点符号的二元文法到五元文法来看，两人小说也分别被聚为一类，反映出二人文本节奏的差异，同时，从前十位的标点符号组合来看，从二元到五元，标点组合相同的越来越少，同时，古龙小说中引号的使用频率相对非常高，反映出古龙小说中的对白较多。

随后，我们使用六个特征对金庸和古龙文本进行总体上的考察，结果证实二者在文本可读性和语言、节奏变化程度上有较大差异。同时，在使用主成分分析法，对八个特征进行考察，并且利用各个文本的前三个主成分得分对文本进行分类，结果证实，金庸与古龙的在语言的正式程度和丰富性差异是存在的。

本文的不足之处在于主要以词和 N 元文法为特征，未来可以考虑更多的特征。

参考文献

[1] Jack Grieve. Quantitative authorship attribution: an evaluation of techniques[J]. Literary and Linguistic Computing, 2007,,22(3): 251-270.

[2] Baayen, R.H., Van Halteren, H., Neijt, A. et al. An experiment in authorship attribution[C]. In Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data.

[3] de Vel, O., Anderson, A., Corney, M. et al. Mining e-mail content for author identification forensics[J]. SIGMOD Record, 2001,30(4): 55-64.

[4] 陆芸.词汇丰富性测量方法及计算机程序开发:回顾与展望[J].南京工业大学学报:社会科学版,2012,11（2）：104-108.

[5] Binongo, J.N.G., & Smith, M.W.A.. The application of principal component analysis to stylometry[J]. Literary and Linguistic Computing, 1999,14(4): 445-466.

[6]Burrows, J.F.. Word patterns and story shapes: The statistical analysis of narrative style[J]. Literary and Linguistic Computing,1987,2(2), 61-67.

[7]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，(30)：137-139.
[8] Rong Zheng，Jiexun Li, Hsinchun Chen et al. A framework for authorship identification of online messages: Writing-style features and classification techniques[J]. Journal of The American Society For Information Science And Technology, 2006,57(3):378-393.
[9]Stamatatos, E.,et al. Computer-based authorship attribution without lexical measures[J]. Computers and the Humanities, 2001,35(2):193-214.
[10]武晓春，黄萱菁，吴立德. 基于语义分析的作者身份识别方法研究[J]中文信息学报，2006,20((6)：61-68.
[11]李贤平.《红楼梦》成书新说[J].复旦学报:社会科学版，1987，(5)：3-16.
[12] Holmes, D.I. A stylometric analysis of Mormon scripture and related texts[J]. Journal of Royal Statistical Society, 1992,15(5): 91-120.
[13]Ying Zhao,Justin Zobel. Effective and scalable authorship attribution using function words[J].Lecture Notes in Computer Science,2005,2689: 174-189.
[14]曲俐俐.金庸、古龙武侠小说比较论[D].延吉：延边大学，2012.
[15]王开银.金庸、古龙武侠小说语言风格比较研究[D].乌鲁木齐：新疆大学，2008.
[16]陈洁.金庸古龙武侠小说比较论[J].浙江大学学报：人文社会科学版，1999，29（5)：131-138.
[17]刘颖，肖天久. 金庸与古龙小说计量风格学研究[J]. 清华大学学报:哲学社会科学版，2014，29(5)：135-147.
[18]张京楣.基于统计方法的文本风格分析研究[D].济南：山东大学，2010.
[19]阚明刚.几个语体参数的定量对比研究—以新闻报道和访谈对话为例[J].语文学刊，2011，(9)：46-48，54.
[20]黄伯荣，廖序东.现代汉语[M].北京：高等教育出版社，2007.
[21]Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.信息检索导论[M].王斌译.北京：人民邮电出版社，2010.
[22]贺湘情，刘颖.基于文本聚类的语言韵律和节奏风格特征挖掘[J].中文信息学报，2014，28（6)：194-200,207.
[23]丁俊苗.不足与需要：论标点符号的语法功能[J].安徽大学学报：哲学社会科学版，2008，32(4)：83-88.
[24]常淑慧.基于写作风格的中文邮件作者身份识别技术研究[D].保定：河北农业大学，2005.
[25]李惠，刘颖.基于语言模型和文本分类的抄袭判定[J].计算机工程，2013，39（5)：230-233.

作者简介：


肖天久（1990——），男，硕士研究生，主要研究领域为语料库语言学。Email:xtj1990@126.com；


刘颖（1969——），女，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）