汉语的计量特征在语言风格对比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

陈芯莹 李雯雯 王燕 王璐 阚明刚

中国传媒大学 语言学及应用语言学专业 计算语言学方向 100024

E-mail: cici13306@sina.com

摘 要： 提出了将语言计量研究成果应用于语言风格对比及作家判定中的方法。通过对两个 75，000 字的语料中 12 个语言结构特征分布的统计对比，发现了 7 个具有显著分布差异的语言结构特征。并以这 7 个语言结构特征作为文本表示特征对两个 75，000 字的未知作家文本做了相关性分析，并准确判定了未知作家文本的作者。以语言结果的计量特征表示文本的方法加强了语言风格对比及作家判定研究的可解释性，具有较高的理论和应用价值。以语料库和统计方法进行语言结构特征计量研究是汉语语言风格描写研究及作家判定研究的重要方法。

关键词： 语言风格；语言结构特征；三重门；梦里花落知多少

The Application of Chinese Quantitative Characteristics in Comparison of Language Style and Author Judgment — *Triple Gates* of Han Han and *Never Flowers in Never Dreams* of Guo Jingming as Examples

Chen Xinying, Li Wenwen, Wang Yan, Wang Lu, Kan Minggang

Communication University of China 100024

E-mail: cici13306@sina.com

Abstract: the paper proposes the method that applies the results of quantitative language research in comparison of language style and author judgment. The paper discovers 7 language structure characteristics possessing obvious distribution differences through the statistical comparison of 12 language structure characteristics distribution of two corpuses with 75 thousand words. The paper also analyzes two texts with 75 thousand words which are not denoted with authors by regarding the 7 language structure characteristics as text expression characteristics, and accurately judges the authors of the two texts. The method adopting quantitative characteristics of language to denote text can better explain the research of language style and author judgment. The quantitative research of language structure characteristics based on corpus and statistical method is an important method for the research of Chinese language style and author judgment

Key Words: language style; language structure; Triple Gates and Never Flowers in Never Dreams

1 引言

作家在语言表达中所形成的不同言语特征表现在数量上就是统计特征上的差异。换言之，语言风格是由于语言单位使用频率的差异而产生的。语言单位的分布频率是分析作家语言的物质基

础。通过对不同作家语言中语言结构特征的统计可以得出语言风格的一致性或区别性特征，语言结构的分布数据就变成体现作家语言风格的计量特征。反之，如果能够获得陌生文本关于语言结构方面的数据，就有可能以此为基础来判定文本的作者。

通过各种特征对文本进行统计分析的思想最早见于数学家 Morgan 在 1851 年的猜想和建议。到了 20 世纪末，统计方法在比较多种文本的风格特征、判定文本的年代、判定文章的作者、识别同意作者的不同写作风格、判断某一作者的作品先后顺序、推测文章的来源、判定匿名文章作者、辨别文章真伪和判断语言亲属关系等诸多领域内都得到了很广泛和深入的运用（Oakes 1998）。

而在汉语本体研究领域，统计方法的运用主要集中在汉字、词汇的计量研究和风格统计研究中，如常宝儒（1985，1986）、刘源、梁南元（1986）、陈原（1989）、李兆麟（1989）、王德春、陈瑞瑞（2000）、曹聪孙（1988）、吴礼权（2004a，2004b）、曾毅平、朱晓文（2006）。但这些研究没能全面地揭示出不同语言风格中在语言结构上的差异，在语料规模、语言结构的选择和统计方法等方面都有待加强（黄伟 刘海涛 20 09）。

基于语料库和统计方法，获取现代汉语语言结构的计量特征，而后考察这些语言结构计量特征在作家语言风格描写、对比研究及作家判定方面的实际应用。

2 语料与方法

寻求体现作家风格不同的计量特征的具体过程是：选取两个作家的语料样本，对其进行分词并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个样本中的分布是否具有差异。选用作家的其他语料样本，计算其与统计所用语料样本的相关性，测试计量特征在辨别不同作家语言时的有效性。

在选取语料时，考虑到时代等外部因素对语言的影响很难做定性定量分析，我们趋向于选择具有相似语言环境的语料。

郭敬明[1]出生于 1983 年，韩寒[2]出生于 1982 年。两人均成名于新概念作文大赛，被视为 80 后作家的代表。郭敬明的代表作《梦里花落知多少》发表于 2003 年，全文 155，820 字[3]；韩寒的代表作《三重门》发表于 2000 年，全文 158，702 字。两位作家的年龄相仿，成长及写作环境相似，各自代表作的发表时间也相近且篇幅相当。因此，我们认为《梦里花落知多少》与《三重门》具有较大可比性，符合语料选择的标准，是比较理想的实验语料。

确定了语料来源之后，为了更精确地做文本对比，我们从两本书中各自选择了 15 万字（共 30 万字）作为实验文本。其中训练语料文本各 75，000 字，测试语料文本各 75，000 字（训练语料文本共 15 万字，来自于两本小说的前半部分；测试语料文本共 15 万字来自于两本小说的后半部分，训练语料与测试语料无交叉）。之后，我们根据北大的分词体系，采用单词性标注对实验文本进行了自动分词。

[1] http://baike.baidu.com/view/4386.htm?fr=ala0
[2] http://baike.baidu.com/view/5972.htm
[3] 如无提示，文中所列字数之数据均为基于 word 的字数统计结果

3 数据与分析

我们选择的考查对象均为词汇层面和句子层面的语言结构特征。词汇层面的计量信息易于获取，词汇计量研究一直是计量与语言学的研究热点之一。同时，尽管词频仍然是研究的基础，但实词、词性标记、词的位置、词长、词序、单现词（hapax）和N元属性等也都已进入了国内外计量语言学研究的视野。选择了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作为考查对象。（黄伟 刘海涛20 09）

在参考了黄伟、刘海涛（2009）提出的用于文本聚类的汉语计量特征后，我们选择了词长、句长、型例比、副词比例、名词比例、代词比例、助词比例、标点符号比例、陈述句比例、疑问句比例、感叹句比例、单现词等12个语言结构类型作为我们的考察对象。表1列出了两个样本的12个语言结构的分布数据。

表 1　两个训练样本中 12 个语言结构的分布数据[4]

语言结构特征	三重门	梦里花落知多少
词长	1.4054	1.3716
句长	24.7509	33.3967
型例比	6.0304	9.2866
副词比例	0.1068	0.1070
名词比例	0.1751	0.1302
代词比例	0.0646	0.1449
助词比例	0.0706	0.0824
标点符号比例	0.1816	0.1204
陈述句比例	0.7898	0.8585
疑问句比例	0.0791	0.0764
感叹句比例	0.1242	0.0563
单现词（hypax）	0.0875	0.0531

词长=字数（不含标点）/词数；
句长=字数（不含标点）/句数；
型例比=词数/词型数；
副词比例=副词数/词数；
名词比例=名词数/词数；
代词比例=代词词数/词数；
助词比例=助词词数/词数；
标点符号比例=标点符号数量/字数；
陈述句比例=陈述句数量/总句数；
疑问句比例=疑问句数量/总句数；
感叹句比例=感叹句数量/总句数；

[4] 表中数据均四舍五入精确到小数点后4位

单现词（hapax）=文本中仅出现一次的词数；

《三重门》的平均词长（以字数计）比《梦里花落知多少》的平均词长大 2.46%[5]，差距不大。

句长值的研究在统计风格学和作者判别研究方面具有应用价值。根据表 1 的数据，《三重门》的平均句长比《梦里花落知多少》的平均句长少 8.6458，少 25.89%，差距较大。这一数据显示在句子复杂程度方面，《梦里花落知多少》的句子比《三重门》的句子要复杂一些。韩寒曾经评价郭敬明是“小女人”，意指郭敬明的文风较夸浮。而该组数据表明在语言表达上韩寒相较郭敬明确实更加朴实精简一些。

词的型例比可以表示语言中的词汇丰富程度。在这点上《三重门》的型例比大《梦里花落知多少》3.2562，约为 35.06%，差距较大。《三重门》的词汇丰富程度更高一些，《梦里花落知多少》中词的平均使用频率更高一些。

副词比例一项，《三重门》与《梦里花落知多少》低 0.0002，约为 0.19%，几乎没有差别。《三重门》与《梦里花落知多少》在副词使用频率上几乎一致。

《三重门》的名词比例比《梦里花落知多少》的名词比例低 0.0449，约为 34.49%；在代词比例上，《三重门》也比《梦里花落知多少》低了 0.0803，约 55.42%，差距都非常大。数据反应，名词和代词在《梦里花落知多少》中出现的频率明显要高于其在《三重门》中出现的频率。而且《三重门》中名词比例约是代词比例的 271.05%，《梦里花落知多少》中名词比例却是代词比例的 89.86%，说明《三重门》中名词的使用频率要远远高于代词使用的频率，而《梦里花落知多少》中名词的使用频率要低于代词使用的频率，但差距不大。

从助词比例看，《三重门》比《梦里花落知多少》低 0.0118，约 14.32%。《梦里花落知多少》中助词的使用频率要高于《三重门》中助词使用的频率。

从标点符号的比例来看，《三重门》与《梦里花落知多少》高出 0.0612，约 50.83%，差距相当大。这一统计数据符合前面关于句子长度的比较结果。即字数或词数大致相当的文本中，标点符号使用频率高则句子结构相对短小。

陈述句比例，《三重门》比《梦里花落知多少》低 0.0695，约 8.10%。疑问句比例，《三重门》比《梦里花落知多少》高 0.0072，约 3.53%，差距不是特别明显。但在感叹句比例上，《三重门》比《梦里花落知多少》高出 0.0679，约 120.60%，差异非常大。说明《三重门》中感叹句出现的频率要远高于《梦里花落知多少》中感叹句出现的频率。这也符合了人们对韩寒更犀利张狂而郭敬明更温和细腻的印象。

单现词出现比例，《三重门》比《梦里花落知多少》高出 0.0344，约 61.10%。单现词是另一个可以表示语言中词汇丰富程度的数据。单现词越多语言中的词汇丰富程度越高。而此处的数据与型例比显示的结果相符。与《梦里花落知多少》相比，《三重门》的单现词多，型例比低，证明其用词更加丰富。

对比所有 12 组数据，我们发现《三重门》和《梦里花落知多少》在句长、型例比、名词比例、代词比例、标点符号比例、感叹句比例、单现词比例这 7 组数据上的差距较为明显。分析总结可知《三重门》与《梦里花落知多少》相比，词汇使用更加丰富、句子更为简短。在句式选择上《三重门》更多地使用了感叹句。在词汇选择上《梦里花落知多少》更高频率地使用了名词和

[5] 百分比数据均四舍五入精确到小数点后二位。

代词，特别是代词，其使用频率高过了名词；相比之下《三重门》更少使用名词和代词，特别是代词，其使用频率远低于名词。

4 相关性测试

经过统计分析我们发现《三重门》与《梦里花落知多少》的训练文本在考查的 12 组数据中有 7 组数据的差异比较大。我们以这 7 组数据为依据，设计和实施了一个文本聚类实验。

实验用的文本同样取自于《三重门》与《梦里花落知多少》，但于之前统计数据的文本无交叉。待判定作者的这两个实验文本均为 75，000 字的文本，并对其做了如表 2 的 7 组数据的统计：

表 2 《三重门》、《梦里花落之》、未知作家文本 1 和未知作家文本 2 中 7 个语言结构的分布数据

语言结构特征	三重门	梦里花落知多少	未知作家文本 1	未知作家文本 2
词长	24.7509	33.3967	26.5915	28.1548
型例比	6.0304	9.2866	10.7221	6.5499
名词比例	0.1751	0.1302	0.1213	0.1751
代词比例	0.0646	0.1449	0.1599	0.0653
标点符号比例	0.1816	0.1204	0.1280	0.1783
感叹句比例	0.1242	0.0563	0.0644	0.0978
单现词（hypax）	0.0875	0.0531	0.0451	0.0809

在做典型相关分析时，由于典型变量是原始变量的线性组合，具有不同量纲变量的线性组合显然失去了实际意义。不同的数量级别会导致“以大吃小”，即数量级别小的变量的影响会被忽略，从而影响了分析结果的合理性。为了消除量纲和数量级别的影响，必须对数据先做标准化变换处理，然后再做典型相关分析。

（http://wenku.baidu.com/view/c259880d4a7302768e9939ab.html 2010-6-14）。为此，我们以表 2 为基础，对每类语言结构特征的数据（每行数据）进行了标准化处理[6]，得出结果如表 3：

表 3《三重门》、《梦里花落之》、未知作家文本 1 和未知作家文本 2 中 7 个语言结构分布的标准化数据

语言结构特征	三重门	梦里花落知多少	未知作家文本 1	未知作家文本 2
词长	-0.93378	1.391089	-0.43884	-0.01847
型例比	-0.94793	0.510204	1.153025	-0.7153
名词比例	0.859068	-0.70414	-1.014	0.859068
代词比例	-0.86659	0.712248	1.007175	-0.85283
标点符号比例	0.912264	-0.9787	-0.74387	0.810301
感叹句比例	1.229235	-0.93728	-0.67883	0.386878
单现词（hypax）	1.007121	-0.65451	-1.04093	0.68832

[6] 此处所采用的是均值标准差模式．其计算公式为：标准值=（原数据-均值）/标准差

由于相关性能够说明语体的接近程度（桂诗春 20 09），我们可以将四组数据做一个相关性分析，以考察各组文本的语体接近程度。相关系数的公式如下（贾俊平 20 07）：

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \quad (1)$$

根据表 3 所列数据，我们用 excel 统计了各文本之间的数据的相关系数，其统计结果如表 4：

表 4《三重门》、《梦里花落之》、未知作家文本 1 和未知作家文本 2 的相关系数

	三重门	梦里花落知多少	未知作家文本 1	未知作家文本 2
三重门	1			
梦里花落知多少	-0.9515	1		
未知作家文本 1	-0.80887	0.619316	1	
未知作家文本 2	0.87478	-0.76178	-0.95925	1

表 4 所显示结果非常明显，《三重门》与《梦里花落知多少》为负相关，未知作家文本 1 与未知作家文本 2 也为负相关，可知它们分属不同作家的作品。而未知作家文本 1 与《三重门》的相关系数为-0.80887，小于其与《梦里花落知多少》的相关系数 0.619316，且前者为负相关，后者为正相关；相反，未知作家文本 2 与《三重门》的相关系数为 0.87478，大于其与《梦里花落知多少》的相关系数-0.76178，前者为正相关，后者为负相关。所以我们可以由此判定未知作家文本 1 来自于《梦里花落知多少》，作者是郭敬明；而未知作家文本 2 来自于《三重门》，作者是韩寒。通过文本内容验证，证实我们实验所得结果是正确的。

经过实验，发现仅使用上述的 7 个语言结构的分布数据作为文本的表示特征，可以在作家判别问题上取得可信任的结果。可以说，这 7 个结构特征在一定程度上较好地区别了两本小说的语言风格。

5 结语

通过对两个 75，000 字的语料样本进行统计分析，得出了在《三重门》和《梦里花落知多少》中 12 个语言结构特征的数据，并对比分析了这些数据的异同。从中总结出了 7 个具有显著分布差异的语言结构特征，并以这些语言结构特征作为文本的表示特征对 2 个 75，000 字的未知作家文本进行了相关系数统计和分析。以句长、型例比、名词比例、代词比例、标点符号比例、感叹句比例、单现词比例等 7 个语言结构特征作为文本特征，准确地判定了 2 个未知作家文本的作者。

在获取语体计量特征时采用了基于语料库和统计学的方法。黄伟、刘海涛（2009）认为这种方法是对现代汉语语体进行描写研究的重要方法。经过实验证明，它们也是对语言风格描写研究的重要方法。标注体系和工具对统计结果的影响，语言风格在字、词、句等语言结构和语法、语义、语用层面的全面计量描写等，都是今后值得继续和深入研究的课题。

将基于计量语言学研究成果的语言结构分布特征作为语言风格对比和作家判定，实验证明是可行可信的，而且特征选择和对比分析结果都可以从语言学的角度进行分析和解释。而且这种

方法不光可应用于语言风格描写和对比、作家判定，黄伟、刘海涛（2009）曾用这种方法成功地进行了文本聚类和分类实验。这样的方法具有普适性，值得在更多的语言学研究领域进行尝试。

参 考 文 献

[1]Oakes，Michael P.1998.Statistics for Corpus Ling uistics. In Edinburgh Textbooks in Empirical Linguistics. McEndry, Tony and Wilson, Andrew. Edinburgh: Edinburgh University Press.

[2]黄伟，刘海涛.2009.汉语语体的计量特征在文本聚类中的应用.计算机工程与应用.2009,45(29):25-27.

[3]曹聪孙.1988.言语风格统计学试说.天津师大学报,1988（4）：70-75.

[4]常宝儒.1985.现代汉语词汇统计问题的初步研究.语言教学与研究，1985（1）：117-124

[5]陈原主编.1989.现代汉语定量分析.上海：上海教育出版社.

[6]李兆麟.1989.汉语计量研究初探——兼评现代汉语词频词典.辞书研究,1989(1):116-123.

[7]刘源、梁南元.1986.汉语处理的基础工程——现代汉语词频统计.中文信息学报,1(1):17-25.

[8]王德春、陈瑞瑞.2000.语体学.广西：广西教育出版社.

[9]吴礼权.2004a.平淡风格与绚烂风格的计算统计研究.云南师范大学学报,36(2):42-46.

[10] 吴礼权.2004b.庄重风格与幽默风格的计算统计研究.渤海大学学报（哲学社会科学版）,26(5):99-103.

[11]曾毅平、朱晓文.2006.计算方法在汉语风格学研究中的应用.福建师范大学学报（哲学社会科学版）,2006(1):14-17.

[12]桂诗春.2009.基于语料库的英语语言学语体分析,北京：外语教学与研究出版社.

[13]贾俊平.2007.统计学(第三版).北京：中国人民大学出版社.

[14] http://wenku.baidu.com/view/c259880d4a7302768e9939ab.html （2010-6-14）

[15] http://baike.baidu.com/view/4386.htm?fr=ala0 （2009-12-1）

[16] http://baike.baidu.com/view/5972.htm （2009-12-1）