汉语语体特征的计量与分类研究

邰沁清，饶高琦
北京语言大学
raogaoqi@blcu.edu.cn

摘要

本文运用语料库和统计方法对汉语语体进行特征的计量研究，并进一步实现自动分类任务。首先通过单因素方差分析描述语体特征区别不同语体的作用和功能。其次，选取其中具有区分度的语言要素拟合逻辑回归模型，量化语体表达形式并观察特征对语体构成的重要性，并通过聚类计算得到了语体的范畴分类体系。最后，以具有代表性的机器学习模型为分类器，挖掘不同组合特征的结构对于语体自动分类的影响。得出在“词2n+词类2n+标点符号2n+语言特征”的组合特征上，取得了最好的分类结果，随机森林模型达到97.25%的准确率。

关键词： 语言结构 ； 语体 ； 计量 ； 方差分析 ； 机器学习

A study on the measurement and classification of Chinese stylistic features

Tai Qinqing, Rao Gaoqi
Beijing Language and Culture University
raogaoqi@blcu.edu.cn

Abstract

In this paper, corpus and statistical methods are used to conduct quantitative research on Chinese stylistic features, and further automatic classification is achieved. Firstly, the function and function of different stylistic features are described through one-way ANOVA analysis. Secondly, the linguistic elements with distinguishing degree are selected to fit the logistic regression model, and the stylistic expression forms are quantified and the differentiation degree and importance of the features to the stylistic composition are observed. Finally, a representative machine learning model is used as a classifier to explore the influence of the structure of different combination features on the automatic classification of language styles. It is concluded that the best classification result is achieved on the combination feature of "word2n + part of speech2n + punctuation2n + language feature", and the accuracy rate of random forest model reaches 97.25%.

Keywords: Linguistic structure , Style , Metrological characteristics , Analysis of variance , Machine learning

第二十届中国计算语言学大会论文集，第398页-第412页，呼和浩特，中国，2021年8月13日至15日。

1 引言

目前,自然语言处理的应用已在语言形式和语义上取得长足的进步，面向语用的实现是未来的发展方向。语体是从语言形式迈向语用的入口，加快语体学与自然语言处理两大领域结合的研究是学者共同的期望，对语体性质和计算研究均有重要意义。

随着功能语法引入语体研究，学者们开始关注语体在表现形式、表达方法上存在着的系统性差异。语体由各种语言要素构成，要素的比例不同，因而形成了不同的语体。很多研究均已证实语言结构在不同语体中的分布差异是客观存在的，冯胜利（2011）指出，“不同的语体（正式、典雅）有不同的语法（亦即语音、词汇、句法等不同法则），不同的语法反映了不同语体（不同对象、场合、内容等）的需要”。金立鑫（2012）假设“语言机制中存在词汇、句法、语篇和语音的语体模块，它们监控语篇的生成，制约着说写者根据特定语体要求选择相应的词汇、句法结构、语篇衔接形式以及韵律形式”。通过对语言成分分布、语言成分之间的关系进行计量统计，便能够发现某些语言指标在区别、构成语体时的作用和重要程度，进而探究语体的组成和特点。可见，语体特征是反映语体风格和功能最主要的因素，也是度量语体之间相似性与差异性的重要媒介。在现阶段的语体计量研究中，主要从以下四个层面选择特征，代表性研究如下：

一、语音层面。由于语言运用时需要表达不同的语气和感情色彩，因此在使用标点方面提出了不同的要求。林毓霞（1987）较早从标点符号的运用中指出语体间形式的差异；Stamatatos（2001）通过使用文本的高频词语作为特征项，认为标点符号具有区分语体的重要作用；冯胜利（2010；2017）考虑到了语体和韵律之间的重要关联，以人工形式标注了韵律信息和语体信息，发现了单双音节可以区别书面语体和口语语体。

二、词汇层面，包括词类、词和短语特征。Douglas Biber（1988）首创语域变异的多维向分析模式（简称MD），采用因子分析算法，对不同的语域（主要是书面语体与口语语体）进行全面、多维度的描写和解释研究。对国内的诸多研究有启示和指引作用，如刘艳春（2016、2017、2019）使用汉语文本对科技、辩论与演讲、小说等语体进行功能解释；范楚琳、刘颖（2019）采用随机森林和k-means聚类算法筛选出区别鲁迅书信、小说和杂文的语言特征。此外，已有研究证明，仅依靠语言的词汇结构也可以区分文本内容（Pustylnikov 2006；Lindemann,C.&Littig,L. 2006；Germany.Fang 2015）。在英语文本中，Karlgren and Cutting（1994）以词性的频次（如动词、名词、介词）作为区分语体的特征，表明词性信息优于词语和衍生特征。在汉语文本中，侯仁奎（2016）通过主成分分析法以词类特征构建文本向量，通过层次聚类法表明词性特征可以作为汉语不同语体的分类特征。

不少学者也将计量语言学中的词频计量指标引入语体差异的研究。张聪（2018）、黄伟（2017、2018）的研究均表明词频分布参数可以体现出语体的差异，体现语体演化；陈蕊娜（2016）以熵值揭示不同语体的语言差异；侯仁奎（2019、2020）分析汉语句子和从句长度的频率分布、以小句长度计算平均词长的分布，发现不同语体中的参数有所不同，可以作为区分不同语体的语言特征。

三、句法层面：句法成分之间的关系往往也是影响语体的重要因素。刘炳丽（2012、2013）利用依存句法树库，探究了各词类在语体中充当的语法功能。陈芯莹（2013）以句法复杂网络的边数、节点数、聚类系数、平均最短路径长度、网络中心势等语言特征，应用于语体分类和聚类。Mingyu Wan（2019）则借助ICE-GB国际英语语料库英国部分进行内部句法结构的语体差异研究。近年来，基于机器学习和神经网络的算法语体研究中也有不小的进展。周浩（2017）提出了基于层次化神经网络的句法分析模型，能有效提取区别语体的句法结构。吴海燕（2020）则结合注意力机制和多层感知机，挖掘区别不同语体的语言特征。

四、语篇层面。除了语音、词汇、句型等实体性的语言成分要素以外，也需要注重要素之间的联结关系。话题在汉语研究中具有极为重要的价值，徐烈炯、刘丹青（1998）提出“话题正是结构和功能的一个交汇点”，话题能够充当语体统计分析的参项，打破以往句子层面研究话题的传统，提升到语篇的层面。乐耀（2007）通过口语和书面语语篇的对比，分析了不同语体语篇话题的各种表现差异，如名词性话题成分、代词性话题成分的回指、有形回指、无形回指的使用情况等并给出合理的解释。尚英、宋柔（2014）则以标点句为基础，从广义话题的角度对

教育部人文社科基金（20YJC740050）

比了工作报告语体和小说语体的差异，涉及到了命名实体话题、状行话题、谓性话题、逻辑话题和关系话题等5种类型的话题结构，丰富了语体特征的维度，也为计算机自动分析话题结构工作打下了基础。

因此，通过选取不同层次的语体特征，说明其区别度和重要性，能够探索语体的构成及本质，反哺语体学研究。在此基础上，本文将提取具有代表性的语体特征并开展计量和分类研究。主要讨论两个问题：1）不同的语体特征体现了语体何种风格和功能，诸多特征对语体的区别度如何？2）语体能否以量化形式表征并通过特征取得较好的聚类、分类效果。

2 研究方法

2.1 语体语料库建设

语体的概念体系十分庞杂，语料的种类、内容又是无限的，因此，在处理问题时应以简化的手段才能使研究具有可行性。本文以具有典型性的八种语体为分析对象，尽可能纳入丰富多元的语料，达到广度和深度的平衡。在后文中，希望通过代表性的语体特征对语体聚合，形成层次化的语体范畴的分类体系。语料库内容展示如表1：

语体	篇数	类别	总字符数	总词数	备注
微博	200	4	222875	141376	
歌词	200	1	208499	143455	
学术	200	7	219789	126292	涵盖地质、农业等54个学科
新闻	200	6	209231	120248	新华社新闻
公文	200	2	305164	168182	《国务院公报》中的行政法规、决定、命令等
政论	200	8	316592	169730	历年政府工作报告、总书记讲话
小说	200	5	214543	154714	近现代文学作品，鲁迅、巴金、金庸、茅盾等
散文	200	3	228458	155172	朱自清、林清玄、余秋雨、季羡林等人代表作

表 1. 语体语料库建设

在语料加工程度的问题上，由于研究语料均来自真实世界中的实际应用场景，原文中存在结构、标点方面的错误，尤其是在口语化强的语体中。针对微博语料，将无意义和多余的字符删除，只留下个人所发的信息并集合在一起；对于歌词语料，文本中大多没有标点符号，因此预先处理为：若每行末尾没有标点，则加上句号，并将文本中的空格断句改为逗号断句。但对于原始内容中可能存在的词汇句法上的错误，并未加以校对和整理，以还原该语体本身的内容信息和风格特点。所获取的均是未加工过的生语料，以jieba进行分词和词性标注。在语料涵盖的范围问题上，考虑到了文本的同质性的问题，尽可能缩小语料之间的时间跨度。对于小说和散文等语料，存在作者写作风格差异的因素，但也能够代表文学创作语体的总体特点。另外，每小类语体下的样本篇数尽量保持一致，均为200篇（类别为每类语体编号，便于下文对比分析）。参考了其他构建语料库的文本长度后（冯胜利2017、ICE-GB），每篇语料字数定为900-2000字之间，以保证文本篇幅的完整性和大小相当，便于对比。至此，构建了一个总规模为192万字符的语体语料库。

2.2 语体特征的选取

在选取语言特征时，遵循以下两方面的原则：1）语体不同于体裁，并非依据文本内容，而是由于语言风格差异产生的。因此，选取体现“语体”体式结构的形式特征，不以词语的语义内容或文本主题作为语体分类的标准。同时，诸如“把”字句、“被”字句等由于数据分布稀疏而不纳入选取范围内。2）基于以往的研究成果，选取已验证过的具有较好区别不同类别语体作用的语言特征，在同一水平上比较区别度和重要性。

根据上文调研的汉语语体特征研究的进展发现：1）近年来开始关注语音韵律的信息（冯胜利2017），但数量仍较少。因此，本文将标点符号、单双音节动词和形容词比例纳入特征范围内。2）有关句型结构的语体特征研究减少，因此不纳入特征范围内。3）词汇层面的特征仍旧是刻画汉语语体差异最主要的内容，特别是体现词汇丰富程度的特征。因此选取计量语言学中具有典型性的TTR、重复率、基尼系数、indicator-a。4）在语法层面的研究中，由于精标注语料的限制，本文主要将体现语言风格差异的文本可读性特征——句子破碎度、离散度、平均词长等纳入特征范围内。由此总结了17个语言特征，之后将结合具体数据说明语体特征的计算方式和体现的语体功能。

2.3 研究及分析过程

1、在17项语体特征上，使用SPSS中的单因素方差分析和事后检验的方法，一一描述语言特征区别语体的作用和体现的语体功能。进一步以逻辑回归模型表征语体形式，观察各语体特征构成语体时发挥的作用和重要性，从语言学角度阐释定量的数据。

2、结合1的研究，使用Minitab工具，选取区别性的语言特征对语体加以聚类。

3、以机器学习模型随机森林为分类器。除17项语体特征外，还采取N元结构为形式特征。其表示方法是：以整个语体语料库代表汉语语体，提取语言要素Top300的N元结构，再根据每篇文本中该N元结构出现的频率为值，构造1600篇文本300维的特征项。按照通常做法[0]，在实验过程中以25%的数据样本作为测试集，剩余75%的数据用作训练，通过不同层次的语言特征对语体进行分类，挖掘不同组合特征的结构对于语体自动分类的影响。

3 语体特征的计量和聚类研究

3.1 特征的方差分析

方差分析是研究各随机变量Y对某个变量X是否有显著变化的统计方法，即检验变量X与Y之间的相关性。我们以各语言特征为自变量，八大语体种类为因变量。具体流程如图1所示：

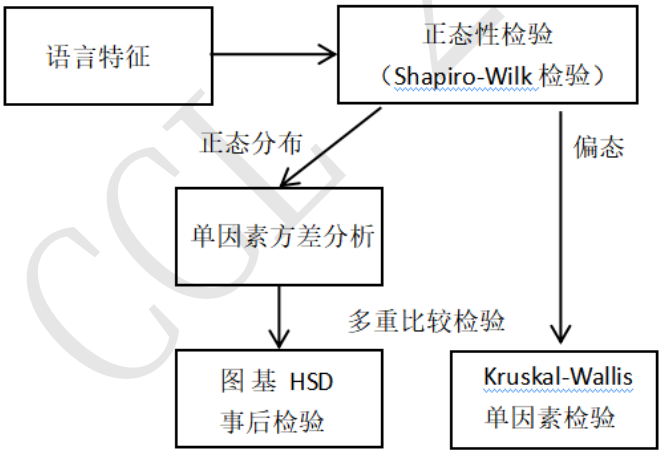

图1.统计研究流程

从统计结果看，表2中的前12项特征指标能够作为因变量进行方差分析，p=.000＜0.05，均对区分语体具有显著差异性。在确定某一因素确实对语体具有显著性的区别和影响后，继续通过图基HSD事后检验的方法对文本分别进行多重比较检验，这是对每两个组的均值进行两两比较，确定具体差异。当两两语体间差异检验的p值小于0.05时，原假设不成立，即两种语体差异显著；当p值大于0.05时，原假设成立，即两种语体具有相似性。p值越大，则表示两种语体越相似（庞双子2019）。例如，在“标点符号比例”的语体特征上，歌词（1）、散文（3）、微博（4）三种语体之间的p值接近1（表格中的数字为语体编号），说明从标点符号的形式特征看，它们在功能和风格上相似。为了方便查看，主要寻找p值＞0.8时两两语体的情况。根据语体特征对语体间的风格和功能归纳为以下三种类别：

[0]范淼，李超.Python机器学习及实践[M].清华大学出版社，2016

		歌词（1）		公文（2）		散文（3）		微博（4）		小说（5）		新闻（6）		学术（7）		政论（8）	
图基HSD	标点符号比例	**3**	1.00	**6**	0.05	**1**	1.00	**1**	0.99	**1**	0.13	**1**	0.62	**1**	0.54	**2**	0.90
		4	0.99	**8**	0.90	**4**	0.98	**3**	0.98	**3**	0.11	**2**	0.05	**3**	0.49	**6**	0.69
		5	0.13			**5**	0.11	**5**	0.61	**4**	0.61	**3**	0.67	**4**	0.96		
		6	0.62			**6**	0.67	**6**	0.14	**7**	1.00	**4**	0.14	**5**	1.00		
		7	0.54			**7**	0.49	**7**	0.96			**8**	0.69				
	词汇密度					**5**	0.08	**7**	1.00	**3**	0.08	**8**	0.97	**4**	1.00	**6**	0.97
	MVR	**4**	1.00					**1**	1.00					**4**	0.05		
	TTR	**2**	0.79	**1**	0.79	**6**	0.97	**7**	0.05			**3**	0.97	**1**	0.99	**3**	1.00
		7	0.99	**7**	0.27	**8**	1.00					**8**	1.00	**2**	0.27	**6**	1.00
	平均词长	**5**	0.75	**8**	0.77					**1**	0.75					**2**	0.77
	单音节形容词比例	**3**	0.29			**1**	0.29	**1**	0.78	**6**	0.85	**5**	0.85	**4**	0.59	**1**	1.00
		4	0.78			**8**	0.75	**7**	0.59							**3**	0.75
		8	1.00					**8**	0.31							**4**	0.31
	单音节动词比例					**6**	0.95	**5**	0.24	**4**	0.24	**3**	0.95				
										6	0.39	**5**	0.39				
	双音节动词比例	**4**	0.36			**5**	0.99	**1**	0.36	**3**	0.99	**7**	0.38	**6**	0.38	**2**	1.00
	indicator-a	**2**	0.85	**1**	0.85	**8**	0.72			**6**	1.00	**5**	1.00			**3**	0.72
	R1			**7**	0.96	**5**	0.32	**8**	0.46	**3**	0.32	**3**	0.30	**2**	0.96	**4**	0.46
						6	0.30					**8**	0.32			**6**	0.32
	RRmc			**7**	0.11	**5**	0.76	**8**	0.60	**3**	0.76	**8**	0.86	**2**	0.11	**4**	0.60
						7	1.00			**7**	0.30			**3**	1.00	**6**	0.86
														5	0.30		
	G	**2**	0.33	**1**	0.33	**6**	1.00					**3**	1.00	**1**	0.11	**3**	1.00
		7	0.11	**7**	1.00	**8**	1.00					**8**	1.00	**2**	1.00	**6**	1.00
非参数Kruskal-Wallis	破碎度			**3**	1.00	**6**	1.00										
				6	1.00												
	双音节形容词比例	**4**	1.00	**6**	0.61	**1**	1.00			**2**	1.00	**3**	0.94			**7**	1.00
						4	0.07			**6**	0.26						
	平均句长（字）					**8**	0.10			**4**	1.00	**2**	0.53				
	离散度（词）					**4**	1.00	**8**	0.68								
	平均句长（字）									**4**	1.00	**2**	1.00			**3**	1.00
	离散度（词）					**4**	1.00									**3**	1.00
																4	

表 2. 多重比较检验的p值结果

（一）篇章韵律、语气特点

A.标点符号比例：所有标点数量与词语数量的比例。结合附录A统计的各语言特征的均值来看，在篇章韵律、语气特点方面，各语体按照标点符号比例均值的排序从大到小依次为：小说＞学术＞歌词＝散文＝微博＞新闻＞政论＞公文。此外，在表2中也能看到，基本上可以将语体分为两大类：小说、学术、歌词、散文、微博，公文、政论、新闻。标点符号比例高的文本往往具有篇章结构、句子语气丰富的特点。在歌词、微博中，多有句间停顿，因而标点符号的比例也高；在小说、散文语体中，作者一般是叙事或者抒情，往往需要借用标点符号的停顿表达丰富的内容和感情，因此在这方面，能够聚为一类相似性的语体。从表格中发现，学术语体的标点占比也较高。从根本上来说，学术文献是要准确、系统地描述解释自然社会以及思维现象，在论证这些现象规律的基础上，理性地反映客体，因而不能够带有谈话语体的随意性、艺术语体的形象性的风格特征。选取部分语料得知，摘要中常常使用英文字符与标点，如“LA-ICPMS”“Co/Ni”等标记符号，使话语表述更加通用、简洁；或者在展示结果数据时，多使用“%”“:”等；也有使用特别序号如“①”的，均被标记为标点符号，因此标点的停顿也较多。公文、政论、新闻由于是国家机关、社会团体等上传下达的一类表达方式，具有断字断句

严谨、正规的特点。由上述分析可以得知，除了限于分词技术，未能区分出学术语体，标点停顿仍然是一个比较好的区别性特征。

B.单双音节占比：一般用来考察文本的正式程度，双音节词占比高则正式程度高；反之，则常用单音节词。在单双音节占比方面看，各语体的比值中，动词占比总是高于形容词占比，且单音节词语的比例总是高于双音节词语的比例。以双音节动词为例，语体的比例均值排序为政论＞公文＞新闻＞学术＞微博＞歌词＞散文＞小说，结合表2，语体主要可以聚集为四类，分别是：政论、公文，歌词、微博，散文、小说，新闻、学术。而单音节的区分作用并不明显，由于分词准确程度所限，并非较好的语言特征。这也从一定程度上说明，双音节动词比例或可继续应用到其他语体的分类中，对广泛的语体都有一定鉴别性。

（二）描写性分析

C.词的型例比（TTR）：词语的数量与词语种类数量的比例，值越大，用词也越丰富。

D.indicator-a：将文本的词语（x）按照其频次f（x）降序排列，排列后词语的次序为秩，记为r（x）。Popescu（2009）提出，h点与文本长度N之间存在$N=ah\hat{2}$的关系，参数a（indicator-a）与实词、虚词使用频率有关。indicator-a不受文本长度的影响，具有较好的区别书面语体、口语语体的作用。

E.重复率（RR）:是描述集中程度的统计量，重复率越高即意味着有更多的词重复出现，词汇丰富度小。

F.基尼系数（G）：Popescu（2011b）研究证实基尼系数越小，文本中词的使用就越不均匀，词汇丰富度越低；反之，文本中的词的使用就越趋于平均。

在描写性分析层面，首先选取了四种不同的词汇丰富度特征。根据以上四类指标在不同语体上的均值来看，可以得到相对较为一致的排序，其中：微博的词汇丰富度总是最高的，其次是新闻和政论，再次是散文与小说，而学术、公文和歌词的词汇使用相对均匀，丰富性较低。仔细查看显著性值，基于词汇丰富性的特征“重复率”“词的型例比”“基尼系数”，新闻与政论语体的用词涉及社会生活的方方面面，因而具有相似性，可聚为一类。但在其他情况下，词汇丰富度的特征体现在不同语体上存在冲突，如在“重复率”的特征下，散文和学术语体具有相似性，但在“基尼系数”特征下，公文语体和学术语体具有相似性，与理想状态不符合，因而不是很好的区别性特征。

G.MVR：是文本中具有修饰性的词语数量与动词数量的比例。从在不同语体上的均值看，从大到小依次为：散文＞小说＞歌词＞微博＞学术＞新闻＞政论＞公文。不同语体对应于不同的交际目的，因而在话语的叙述上具有不同程度的描写性、议论性的风格。MVR值高，体现了文本的描写性和修饰性作用强，语言更加生动，表述灵活多变，具有文学性。反之则更加体现了政论与公文文本的说明性和指令性意味，话语表达可能较为平淡单一。在图凯检测中，将歌词与微博两类语体聚合，体现了口语化的风格，但却未能够将其余语体聚合，或许是由于不同于歌词和微博语体的句子普遍较短，其他文本的数值分布不够均衡，因而较难归类。

H.词汇密度：指文本中的实词数量与词汇总数量的比例。学者Ure（1971）的研究表明词汇密度越高，实词数量越多，一定程度上能够反映文本的书面化与正式的程度。词汇密度高，表明文本中出现的含有实际意义的词语更多。从词汇密度上看，均值的排序为：公文＞政论＞新闻＞学术＞微博＞歌词＞小说＞散文，大致符合我们对于语体的印象。前四类具备书面语体的正式性特点，具有传达信息、指导的性质；微博与歌词作为以网络为媒介的非传统形式上的书面语，介于中间。从词汇密度的角度也发现，语体是一个连续统，处于正式语体/非正式语体的连续体中。小说、散文中，前者叙述故事刻画人物，后者议论抒情，可以认为是书面语中的非正式语体。从显著性值观察，语体主要聚集为两类，分别是：政论、公文，歌词、小说。

（三）句子可读性分析

I.平均词长：总词数数量与总字数数量的比值，能够代表语言单位的复杂程度，反映了文本用词的平均长度。从平均词长的均值看，可知政论＞公文＞学术＞新闻＞微博＞散文＞小说＞歌词。平均词长越短，文本更加简单易读。图凯检测中将歌词与小说，政论与公文两种语体聚集在一起，分别代表了文本阅读难易程度的两极。

J.句子破碎度：指句子中间因断句而停顿的次数，也即一句话的零散程度。一般来说，句中标点符号越多，停顿也就越多，破碎度也就越高，语言口语性越强，反之，则句子流畅通达，书面性越强。从破碎度的均值可知，学术、新闻、公文、散文、政论语体的句子流畅通达，书面性较强，而微博、小说与歌词三种语体偏向于口语化，也印证了小说在正式语体/非正

式语体的连续体中，偏向于口语化表达的语体。

K.句子离散度和平均句长：离散度指文本中句子句长偏离平均句长的程度，反映了文本节奏上的变化，离散度小则表明句子富有韵律性，平稳有秩；反之则跌宕起伏、富于变化。其中，平均句长指文本中的所有句子的长度之总和与句子总数之比。两者可用汉字或词语进行相应的统计。从离散度及句子长度的均值比较八类语体，排序从大到小为：学术＞公文＞新闻＞政论＞散文＞微博＞小说＞歌词。可以看到语体依据文本的韵律节奏分为了两大类，学术与公文都有明确的写作规范和指导，对于语言的规整性和书面化的程度约束性极强，因而文本节奏趋于稳定；歌词的文本一般篇幅相对较短，句子与句子之间的长度变化不一，音乐旋律的节奏对其有约束性；从另一方面来说，歌曲也属于诗歌，是富有韵律性的语体，在节奏乃至押韵方面都有特色。根据Kruskal-Wallis检验的成对比较可知，散文、微博、政论语体之间差异性不显著，这三种语体在一定程度上可以说都具有评述性和议论性，因而文本的节奏较为相似。

结合单因素方差分析，以上的部分计量指标能够较好地区分出不同的语体，分别是：标点符号、TTR、MVR、平均词长、词汇密度、平均句长（词）、平均句长（字）、破碎度、离散度（词）、离散度（字）、双音节动词等11个语言特征。将上述分析体现的语体的风格特点与功能归纳为表3：

	歌词	公文	散文	微博	小说	新闻	学术	政论
语气丰富	+		+	+	+		+	
正式性		+				+	+	+
词汇丰富				+		+		+
描写性	+		+	+	+			
议论、抒情性	+		+	+	+			
文本可读性	+		+	+	+			
口语性强	+			+	+			
文本节奏韵律性	+		+	+	+			

表 3. 语体风格汇总表

3.2 基于逻辑回归模型的语体表达

不同的语体能够因不同的语言成分聚集在一起，印证了语体与语体之间并非绝对互斥，而是由于语言特征的分布和比例各有差异，因此形成了具有独特语言风格的语体。那么语体是否能够通过语言特征及其各自的系数进行量化表达呢？我们借助SPSS进行逻辑回归分析，能够探讨对因变量产生影响的自变量，以及这些自变量的权重。

在逻辑回归分析中，自变量的选择应遵循少而精的原则，需剔除显著性不高的变量，以从诸多变量因子中找到更具影响的要素。因此，先采取逐步回归法确认主要变量，能够确保加入新变量之前，回归方程中始终都包含显著的变量。将上述的11种特征传入后，逐步回归进行了10步，其中离散度（字）被自动除去。表4是逐步回归最终模型的拟合优度信息表，显示P＝0.000，拒绝原假设，因而认为回归模型显著有效。其中，-2 对数似然的值越小，拟合效果越好。从结果中看出，加入了自变量后的模型比仅有常数项的模型拟合更优秀（6654.213＞950.1），显示出具有统计学意义。

模型拟合信息				
模型	模型拟合条件	似然比检验		
	-2 对数似然	卡方	自由度	显著性
仅截距	6654.213			
最终	950.100	5704.113	70	.000

表 4. 多分类逻辑回归拟合优度信息表

据此，得到了每类语体的线性回归函数表达式（具体数值详见附录B），初步认为每种语体由以下的语言特征和其回归系数组合构成。例如，Y为小说语体，X1X2.....X10为语言特征，构成如下表达式：Y=70.644X1-5.44X2+2.222X3-50.041X4-0.375X5-0.139X6+0.045X7+0.051X8-99.596X9-16.560X10+117.758

其中，Exp（B）是逻辑回归中的一个重要概念，用以度量某个自变量对因变量影响的程度。在理想条件下，值越高，则对应的自变量越为重要。将每种语体下语言特征的P值依次进行排序，整理于表5。发现标点符号、词汇密度对于大部分语体都十分重要，排于前列，而双音节动词稍弱。例如歌词语体，体现文本节奏和韵律的语言特征——标点符号和双音节动词，对于区别该语体更为重要；而小说语体中，TTR显著，更能突出与其他语体的相异之处。

歌词	标点符号、双音节动词、破碎度、平均句长（字）、MVR、离散度（词）、平均句长（词）、词汇密度、TTR、平均词长
公文	标点符号、词汇密度、平均句长（字）、离散度（词）、平均句长（词）、破碎度、平均词长、MVR、双音节动词、TTR
散文	标点符号、破碎度、平均句长（词）、离散度（词）、TTR、平均句长（字）、词汇密度、MVR、平均词长、双音节动词
微博	标点符号、词汇密度、TTR、破碎度、离散度（词）、平均句长（字）、平均句长（词）、MVR、平均词长、双音节动词
小说	标点符号、TTR、离散度（词）、平均句长（词）、平均句长（字）、破碎度、词汇密度、MVR、平均词长、双音节动词
新闻	标点符号、词汇密度、TTR、平均句长（字）、离散度（词）、破碎度、平均句长（词）、MVR、平均词长、双音节动词
学术	平均词长、词汇密度、平均句长（词）、破碎度、离散度（词）、平均句长（字）、标点符号、MVR、TTR、双音节动词

表 5. 各语体特征重要性排序

3.3 语体的聚类

在前两部分中，通过单因素方差分析与逻辑回归模型，得到了具有较好区别性的11 类语体特征。为了验证计量特征区别语体的可行性，将在此基础上，进一步对语体进行聚类研究，并从计算的角度观察和分析语体的范畴分类体系。筛选后的11个特征分别为：标点符号比例、双音节v比例、TTR、MVR、词汇密度、平均词长、破碎度、平均句长（字）、离散度（词）、平均句长（字）、离散度（词）。使用的聚类方法原理是：以每种特征的平均值为依据，根据值的相似程度来合并不同的语体，适用于暂不确定语体应当分为几大类别时的情况。设定统一参数为：相似性水平（I）为95%，联结法为最长距离，距离度量为Euclidean平方，标准化变量。聚类结果如下图所示：

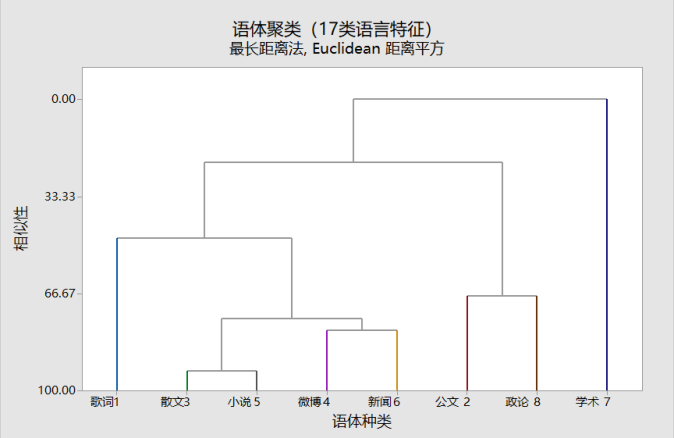

图2.语体聚类（17类语言特征）

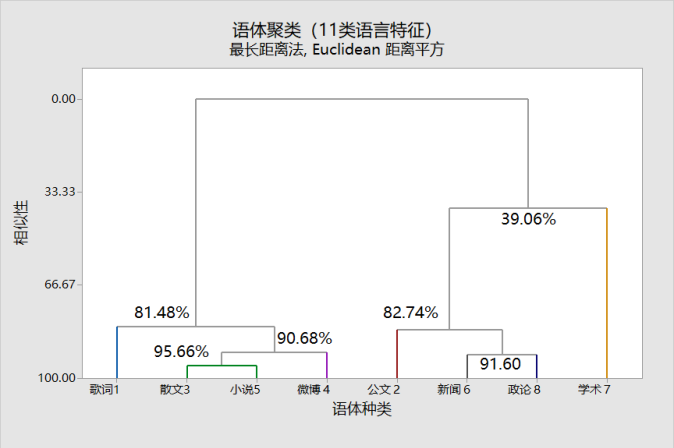

图3.语体聚类（11类语言特征）

通过对比发现，在11类语言特征的基础上明显具有较好的聚类效果，分为了三个层次。首先，八类语体共分为两大类别：第一个聚类的集合中有歌词、散文、小说、微博。其中散文和小说语体的文本相似度为95.66%，微博与其相似度为90.68%，歌词与其相似度为81.48%——也即第一大聚类的文本间相似度；第二个聚类的集合中有公文、新闻、政论和学术，新闻和政论语体的文本的相似度为91.60%，公文与其相似度为82.74%，学术与之相似度为39.06%。据此，我们对语体体系进行了归纳总结，构建了基于计算的语体范畴分类体系。如下图所示，语体自上而下，口语性逐渐增强，非正式性程度加深。在正式性的程度上：公文＞学术文献＞政论＞新闻报道＞小说＞散文＞微博＞歌词＞谈话＞问答。

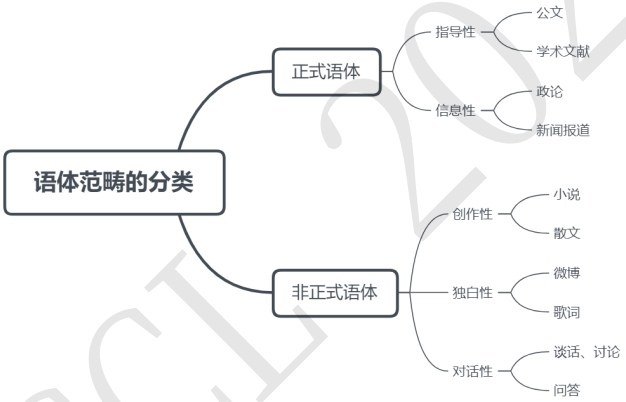

图4.语体范畴分类

通过计算获得的聚类结果符合传统语体学对于语体范畴分类的认知，并且在前文相关性分析的章节中也有体现，但也有与经验认知相异之处。首先，在第一层次上，语体二分为“正式语体和非正式语体”。这里的命名方式借鉴了冯胜利（2010）的语体体系，正式度是最基本、最原始的语体范畴，是话语的本质属性。我们不以媒介载体（书面语/口语）为分类依据，而是以特征体现出来的风格为标准，这也是语体形成的内在语言学根据。其次，从聚类的结果看，内部还有更进一步、更深层次的合并。在讨论语体的语言属性时，语体所体现出来的风格归根结底是由于其交际性来定义的。因此在第二层次上，从交际目的出发进行了抽象的概括。命名方式上借鉴了国际英语语料库ICE-GB的名称。

正式性的语体下,分为指导性和信息性语体，前者包含公文、学术等具有指导、规范性质的语体，一般是给具有同等专业水平的人员查阅，因而专业性、正式性强，读者在阅读时较为晦涩、困难；后者包含政论、新闻等由上至下传达信息的语体，这类语体由于需要对更加广泛的群体进行信息的传播，一定程度上要求通顺易读，所以正式性相比指导性语体稍弱。非正式语体包含创作性语体、独白性语体和对话性语体。其中，散文和小说虽为书面语形式，但从前文的分析中能够看到，特征所显示的风格是有口语化倾向的。尤其是小说的破碎度、离散度高，口语性强，节奏韵律强。不同于指导、信息性语体具有抽象性和说明性，创作性语体多叙述，多刻画描写细节。为了表达上的通俗易懂，语言可以说是经过一定修整和加工的口语化表达。

新的分类体系帮助我们认识了语体之间的连续性，在进行多种语体分类时，提供了新的角

度和手段。相比于以往的分类体系，进一步摈弃了从内容上对语体进行区分的刻板印象，以语言特征的分布和功能上重新调整，更加具有客观性。

4 基于机器学习模型的语体分类

不同于上一小结探讨语体特征与语体之间的关系，强调的是语体风格的分析，这一部分开展语体的自动分类，希望纳入更加丰富的语言信息，以查看这些语言要素的组合在分类中的作用，提高分类的准确率。

首先，以17种语体特征为特征项，取得随机森林（RFC）总体准确率为91.25%。可见从语言结构的计量特征上已经能够较好地区分各类语体。观察分类结果，在八类语体上，歌词的F1值始终是最高的，政论语体次之。说明所选择的语言特征已经有较好的区别和解释作用。特别是歌词语体，能够全部被分类正确。观察语料，这是由于歌词语体本身的形式十分有特点：

那些年错过的大雨
那些年错过的爱情
我们俩的回忆
就像一本青春手册
有过天真有过痛
也有很多快乐快乐
.....

在格式上，每一小句构成一行，极为工整；每一句小句的字数、词语数也基本相同，因而正确率高。体现了这两类语体的结构稳固，形式化的特征最为明显；而散文的分类结果较低，这是由于作者的写作风格差异，且创作性语体也不需要遵循过多的格式规范，因而在语体分类上表现稍差。

观察错例发现，公文语体主要是被错误归类到了新闻、学术和政论语体中，政论和学术语体也是一样，说明这四种书面、具有正式性语体的相似度强。新闻语体分类的正确率最低，错例分布范围最广，这是因为其种类多样，一般有消息、通讯、新闻特写、调查报告、专访、社论、述评、思想评论、理论文章、副刊等。其中，思想、理论性强的社论、理论文章易与政论混淆；而报道人物、具有文艺色彩的专访、副刊也易与创作性语体混淆。下两例分别被错误划分至了小说和散文，例4具有小说的叙事性质，口语化较强；例5具有散文的议论性质，均具有文艺色彩：

例1：这一聊，竟聊出了感情。“他天天来我家附近找活做，空闲的时候就和我聊天，有时有人想用他的残疾摩托车，他就送人家过去，回来后又跟我接着聊天，渐渐就有了感情。”方老说。

例2：安倍“拜鬼”之日，日本外相玄叶光一郎正在欧洲兜售日本在历史和领土问题上的“独特”主张。但安倍在靖国神社为军国主义招魂的深深一躬，使得玄叶的一切外交言辞都加倍苍白，也加倍讽刺。倘若心障不除，纵然舌绽莲花，也无法把“魔影”打扮成“魅影”。

微博也被错误归类到不同语体。微博是社交平台，涵盖了官方、自媒体、个人的发布的各类消息。如下方例6被分至新闻语体，例7被归类到散文语体：

例3：2010年12月，湖南农民李清及妻子李红英被带至内蒙古鄂尔多斯看守所。2011年9月当地法院以“犯假冒注册商标罪”，判处李清有期徒刑5年，并处罚金2151万元。

例4：征税是一项艺术，如同拔鹅毛，把鹅都快拔裸体了，鹅还舒服的一声不叫，那真叫顶级艺术家。

从上述错误分类的内容看，难以从外在格式线索辨别差异，需进一步挖掘其他语言特征作为分类依据。因此，除了主观地选取语言成分外，还采取N元结构为形式特征，这是将文本中连续N个语言要素组成的字符串作为特征，可以体现语体中语言成分联结的紧密程度和搭配关系，并弥补词和词、句与句之间关系省略的问题。理论上认为，N值越大，越贴近真实的语言组成特点，但实际上N值的增加会带来数据的稀疏。例如，本研究中由于歌词语体格式具有特殊性，大部分文本每行只有1-2句句子，如果以3元结构及以上作为特征，则可能产生较多空值，对于文本分类意义弱。因此主要选择了词、词类、标点符号及其2元结构为比较研究的对象，并选择语言要素Top300的特征项，理由如下：

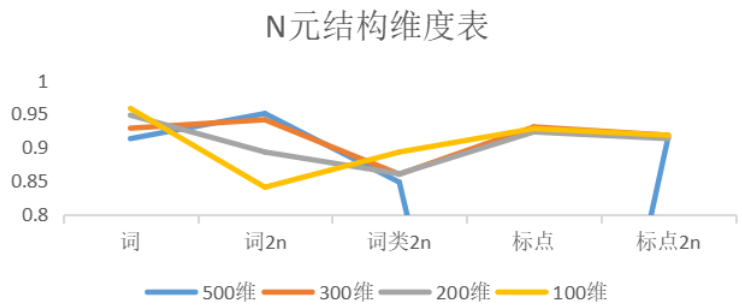

图5.N元结构组合准确率

上图中，横坐标为N元结构，纵坐标为随机森林分类器的F1值，不同颜色的折线代表了不同的维度在此特征上的准确率。由于标点符号没有500维的数据，因此数据为0（词类仅有32维，同理）。可以看到，当选取Top100时，分类的准确率不稳定；当选取Top200时，准确率偏低。因此，最终选取Top300维，将每一种语言特征都尽可能保持在300维度左右。

结合多种语言层面的特征，观察分类效果，所选择的语体特征及对应的准确率列表如下：

语体特征	准确率（%）	语体特征	准确率（%）
词	92.81	语体特征	91.25
词类	84.06	词2n+词类2n	93.75
标点符号	87.19	词2n+标点符号2n	95.00
词2n	94.06	词2n+词类2n+标点符号2n	96.00
词类2n	85.94	词2n+词类2n+标点符号2n+语体特征	97.25
标点符号2n	88.59		

表 6. 语言特征组合准确率（%）

第一、二列数据展示了词、词类及标点符号的一元和二元的分类效果，比较来看，准确率由高到低依次是：词的2元结构＞词＞语言特征＞标点符号的2元结构＞标点符号＞词类的2元结构＞词类。从分类结果可以明显看到，首先，这几类单从形式上表征文本的特征均具有一定的分类效果，但区别的效果不同。在词层面的区别效果最佳，词是组成语句、篇章的具有一定意义的最小结构单元，关于语体定义的“词语类别”说（乐秀拔1959）认为，语体是指在运用上受各种范围所限制的词语类别，根据文章的性质和内容、根据接受的对象，选择适合于我们所需要的词语。虽然不足以定义语体的概念，但通过分类实验，我们验证了在单一的形式表征层面，词语的区分效果最好。语言特征略次于词及词的二元结构，高于其他的形式结构，再次说明了其在语体分类中的重要作用。标点符号是语篇中辅助词语表达的一种手段，可以表示句子的停顿和连贯、句子疑问、感叹、陈述等语气，体现了语句组织连贯性的特点；词类是词按照句法功能进行的分类，表示词语的属性类别，相比于标点符号，其在文章中的作用较弱，因而分类准确率较低。其次，从词、词类和标点符号来说，2元结构的分类作用均高于单一的结构特征，一般认为，2元体现了语言单位前后组合的相互关系，包含了更多信息，因而分类结果更优。

第三、四列数据体现了语言的组合结构，我们发现，在词2n的基础上依次增加一种形式特征，分类的准确率就有逐步地提升。词2n+词类2n+标点符号2n+语体特征＞词2n+词类2n+标点符号2n＞词2n+标点符号2n＞词2n＞词2n+词类2n。“词2n+词类2n+标点符号2n+语体特征”的组合具有目前最优的分类效果，准确率达97.25%。根据分类结果查看组合特征的有效性，发现在语体的构成中，每一种语言层级都是构筑成语篇的一种建筑材料，具有不可或缺的重要性。如微博文本（截取片段）在“词2n+词类2n”的特征上被错分为了新闻语体：

例5：10月4日21时许，甘肃省庆城县三十铺镇阜城村一刚满2个月男婴在家中被盗。经当地公安机关工作，12日抓获王家彦等4名犯罪嫌疑人，安全解救被盗婴儿。昼夜奋战三天，刚刚回到北京。圆满完成抓捕解救任务，心情舒畅。送上鲜花，向辛苦奔波的全体打拐民警致以诚挚的慰问和衷心的感谢！今晚工作进展顺利，可以睡个安稳觉了。诸位晚安。

但在“词2n+标点符号2n”和“词2n+词类2n+标点符号2n”等组合特征下均能够被分类正确，可能是由于标点符号的使用体现了微博文本的特点。又如小说《幸福的家庭》被错分为微博，说明仅从词语、标点2n的组合关系难以区分这类口语化程度较高的文本：

例6：他的笔立刻停滞了；他仰了头，两眼瞪着房顶，正在安排那安置这“幸福的家庭”的地方。他想：“北京？不行，死气沉沉，连空气也是死的。假如在这家庭的周围筑一道高墙，难道空气也就隔断了么？简直不行！江苏浙江天天防要开仗；福建更无须说。四川，广东？都正在打。

从上述的分类结果，能看到组合特征的重要性和区别度。部分文本即使在人工判断时也可能出错，但在客观的语言规律和统计数据的支持下，语体分类系统能够较好地平衡经验与理性的不一致问题，从而为语体的计量和分类研究提供了一种量化的方法和参考。

5 结论和展望

本文开展了面向自然语言处理的语体研究，兼顾语体学的基本理论和统计机器学习的方法。使用了量化的方法对语体加以形式化的表达，并探索语体间的差异和语体的基本构成，使研究反哺语体学理论的发展。主要可以总结为下列两个方面：1、语言特征的组合构成了语体，其权重的差异表现为不同语体的风格和功能有差异。2、语体又因语言特征的差异加以明显区分，从分类的角度来看，不同层次的语言特征具有不同的区别度，说明对文本的重要性不同。

限于对语体学理论知识和分类算法的知识，目前仅考虑到从语言特征层面、分类的角度来研究自然语言处理中的语体学，仍较为浅薄。在未来的研究中，有两个可以探索的方向。首先，目前选取的语言特征还较有限。语体监控着语篇整体的形成，每一种层面都是不可或缺的。虽然从语音韵律、词语、词类层面进行了不同程度的说明，但未涉及到句法结构层面以及语篇层面。其次，在选取代表性的语言特征时，主要是基于前人的研究和自身的知识结构，较为主观。是否可以运用深度学习自动挖掘语言特征，突破人工选取的经验性，形成新的发现。

但文本能够在“词2n+词类2n+标点符号2n+语体特征”上被分类正确，可能是由于虽然文本的破碎度较高，但其离散程度和平均句长具有小说的特点，能够体现小说语体的韵律性和节奏性。从上述的分类结果，能看到组合特征的重要性和区别度。部分文本即使在人工判断时也可能出错，但在客观的语言规律和统计数据的支持下，语体分类系统能够较好地平衡经验与理性的不一致问题，从而为语体的计量和分类研究提供了一种量化的方法和参考。

参考文献

Biber D. 1988. *Variation across speech and writing*. Cambridge University Press.

Jussi Karlgren, Douglass Cutting. 1994. *Recognizing text genres with simple mertics using discriminant analysis*. In:Proceedings of the 15th conference on Computational linguisitics

Mingyu Wan, Alex Chengyu Fang, Chu-Ren Huang. 2019. *The Discriminativeness of Fine-grained Internal Syntactic Representations in Automatic Genre Classification*. In Journal of Quantitative Linguistics.

Hou Renkui, Huang Chu-Ren, Liu Hongchao. 2019. *A Study on Chinese Register Characteristic Based on Regression Analysis and Text Clustering*. Corpus Linguistics and Linguistics Theory.

Hou Renkui, Jiang Minghu. 2016. *Analysis on Chinese quantitative stylistics features based on text mining*. Digital Scholarship in the Humanities.

Alex Chengyu Fang, Jing Cao. 2015. *Text genres and registers: The computation of linguistic features*. New York: Springer, Heidelberg.

刘艳春. 2019. 小说等四语体在语体变异模式中的定位与特征——基于*17*个语体的语体变异多维度考察. 江汉学术.

刘艳春,赵艺. 2018. 专门科技语体和通俗科技语体多特征对比研究. 江汉学术.

刘艳春,胡凤国,赵艺. 2016. 辩论与演讲语体多维度、多特征对比研究. 语言教学与研究.

冯胜利. 2018. 汉语语体语法概论. 北京语言大学出版社.

范楚琳,刘颖. 2020. 基于多维度分析法的鲁迅三种文体比较研究. 中文信息学报.

吴海燕,刘颖. 2020. 基于注意力网络的语体多元特征挖掘. 计算机应用.

冯胜利,王永娜. 2017. 语体标注对语体语法和叙事、论说体的考察与发现. 汉语应用语言学研究.

周浩. 2017. 基于神经网络的句法分析研究. 南京大学.

张聪,刘海涛. 2018. 词频计量指标与汉语语体演化. 外语教学.

霍四通. 2000. 语体研究和自然语言处理. 修辞学习.

霍四通. 2004. 面向自然语言处理的语体学理论. 福州：海风出版社.

刘丙丽,牛雅娴,刘海涛. 2013. 汉语词类句法功能的语体差异研究. 语言教学与研究.

陈芯莹,刘海涛. 2013. 句法复杂网络作为语体分类的知识源研究. 计算机工程与应用.

刘丙丽,牛雅娴,刘海涛. 2012. 基于依存句法标注树库的汉语语体差异研究. 语言文字应用.

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.

陶红印. 2010. 从语体差异到语法差异*(上)*——以自然会话与影视对白中的把字句、被动结构、光杆动词句、否定反问句为例. 当代修辞学.

陶红印. 2010. 从语体差异到语法差异*(下)*——以自然会话与影视对白中的把字句、被动结构、光杆动词句、否定反问句为例. 当代修辞学.

尚英,宋柔. 2014. 基于广义话题结构语料库的语体对比研究——以报告体与小说体为例. 计算机工程与应用.

金立鑫,白水振. 2012. 语体学在语言学中的地位及其研究方法. 当代修辞学.

冯胜利. 2011. 语体语法及其文学功能. 当代修辞学.

冯胜利. 2010. 论语体的机制及其语法属性. 中国语文.

方鸷飞,林鸿飞,杨志豪,赵晶. 2006. 中文文本体裁的自动分类机制. 中文信息学报.

王永娜. 2010. 汉语书面正式语体的语法手段. 北京语言大学博士学位论文.

乐耀. 2007. 现代汉语语篇话题探微. 华中师范大学硕士学位论文.

刘海涛. 2018. 计量语言学研究进展. 杭州：浙江大学出版社.

附录A.各语体特征均值

	歌词	公文	散文	微博	小说	新闻	学术	政论
标点符号比例	0.24	0.20	0.24	0.24	0.26	0.22	0.25	0.21
单音节a比例	0.04	0.02	0.04	0.04	0.03	0.03	0.05	0.04
单音节v比例	0.27	0.29	0.22	0.24	0.23	0.23	0.20	0.30
双音节a比例	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.03
双音节v比例	0.14	0.26	0.12	0.15	0.11	0.18	0.17	0.27
TTR	0.28	0.28	0.34	0.37	0.33	0.35	0.28	0.34
indicator-a	10.03	9.57	13.98	16.30	12.43	12.52	8.07	14.52
RR	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01
G	0.45	0.46	0.37	0.31	0.39	0.37	0.47	0.36
MVR	0.55	0.22	0.67	0.54	0.61	0.42	0.49	0.32
词汇密度	0.73	0.80	0.68	0.75	0.70	0.77	0.75	0.78
平均词长	1.80	2.27	1.86	1.96	1.82	2.10	2.18	2.29
破碎度	1.22	3.38	3.32	2.48	2.14	3.53	6.14	2.90
平均句长(字)	8.65	57.54	30.10	22.00	19.67	48.62	91.16	35.36
离散度(字)	2.76	39.21	19.24	20.06	16.28	28.20	60.59	24.44
平均句长(词)	5.55	27.90	18.54	12.79	12.58	26.11	48.78	17.09
离散度(词)	2.07	21.30	12.92	12.58	11.04	16.63	35.56	13.22

附录B.多分类逻辑回归参数表

种类a		B	显著性	Exp(B)	种类a		B	显著性	Exp(B)
歌词	截距	838.847	.998		小说	截距	117.758	.000	
	标点	539.898	.997	2.98E+234		标点	70.644	.001	4.78806E+30
	词汇密度	-77.366	.999	2.52E-34		词汇密度	-5.440	.649	0.004
	TTR	-161.041	.999	1.15E-70		TTR	2.222	.880	9.222
	平均词长	-405.274	.998	9.81E-177		平均词长	-50.041	.001	1.85E-22
	破碎度	22.785	.999	7856581236		破碎度	-.375	.643	0.688
	平均句长（字）	22.546	.998	6187005439		平均句长（字）	-.139	.867	0.87
	平均句长（词）	-39.472	.999	7.21E-18		平均句长（词）	.045	.975	1.046
	离散度（词）	-26.569	.996	2.89E-12		离散度（词）	.051	.643	1.052
	双音节动词比例	214.817	.	1.97E+93		双音节动词比例	-99.596	.000	5.57E-44
	MVR	7.629	1.000	2057.091		MVR	-16.580	.000	6.30E-08
公文	截距	31.436	.086		新闻	截距	45.072	.013	
	标点	33.033	.008	2.21912E+14		标点	57.665	.000	1.10591E+25
	词汇密度	26.278	.000	2.58529E+11		词汇密度	25.635	.000	1.35842E+11
	TTR	-62.397	.000	7.97E-28		TTR	4.312	.632	74.603
	平均词长	-13.516	.120	1.35E-06		平均词长	-29.978	.000	9.57E-14
	破碎度	-1.301	.005	0.272		破碎度	-.357	.449	0.7
	平均句长（字）	.522	.127	1.686		平均句长（字）	.463	.167	1.589
	平均句长（词）	-.565	.427	0.568		平均句长（词）	-.414	.553	0.661
	离散度（词）	-.089	.126	0.915		离散度（词）	-.253	.000	0.776
	双音节动词比例	-30.098	.000	8.49E-14		双音节动词比例	-60.590	.000	4.86E-27
	MVR	-19.704	.000	2.77E-09		MVR	-12.762	.000	2.87E-06
散文	截距	66.910	.007		学术	截距	-11.335	.594	
	标点	39.452	.052	1.36055E+17		标点	-11.010	.475	1.65E-05
	词汇密度	-14.571	.219	4.70E-07		词汇密度	2.960	.684	19.29
	TTR	-1.023	.942	0.36		TTR	-51.649	.000	3.71E-23
	平均词长	-18.416	.173	1.01E-08		平均词长	16.668	.106	17326202.1
	破碎度	3.444	.000	31.3		破碎度	.280	.585	1.323
	平均句长（字）	-1.422	.028	0.241		平均句长（字）	-.786	.037	0.455
	平均句长（词）	2.328	.039	10.257		平均句长（词）	2.092	.007	8.1
	离散度（词）	-.344	.002	0.709		离散度（词）	-.192	.002	0.825
	双音节动词比例	-111.554	.000	3.57E-49		双音节动词比例	-75.466	.000	1.68E-33
	MVR	-16.812	.000	5.00E-08		MVR	-11.754	.001	7.86E-06
微博	截距	64.105	.006						
	标点	40.473	.026	3.77588E+17					
	词汇密度	31.600	.001	5.29487E+13					
	TTR	24.525	.054	44800022660					
	平均词长	-38.633	.002	1.67E-17					
	破碎度	2.188	.005	8.914					
	平均句长（字）	.227	.697	1.254					
	平均句长（词）	-1.273	.235	0.28					
	离散度（词）	.354	.000	1.425					
	双音节动词比例	-63.155	.000	3.73E-28					
	MVR	-15.881	.000	1.27E-07					