

汉语二语学习者作文质量影响因素研究

刘志芳,马廷芳

(河南师范大学 文学院,河南 新乡 453007)

摘要:以 HSK 动态作文语料库中 56 篇蒙古国汉语学习者的作文语料为研究对象,使用语料库调查法,数据分析法和多元线性回归法,考察了词汇、文字、语法 3 个特征对蒙古国汉语学习者作文质量的影响.研究发现:词汇、文字、语法都能影响学习者的作文质量,其中词汇特征对作文质量影响最大;不同分数段的作文中各特征的影响不均衡;经过回归诊断,选取词数、词种数、中级词数、高级词数 4 个指标构建多元线性回归模型,检验出对自动评分最显著有效的参考指标是词汇特征的下辖指标词数.

关键词:汉语二语学习者;蒙古国汉语学习者;作文质量;影响因素

中图分类号:O212;H030

文献标志码:A

文章编号:1000-2367(2025)03-0112-09

学界对汉语二语学习者书面语输出质量影响因素的研究较多,主要涉及语言特征和内容质量两部分^[1],其中语言特征包括作文长度、语法、词汇特征等内容^[2],词汇特征指标主要涉及词种数和高级词种数^[3],词汇复杂性和词汇准确性^[4]等内容.针对蒙古国汉语学习者作文质量影响因素的研究多集中在单个语言特征研究方面,将不同语言特征综合对比的研究不多.本文选取了词汇、文字、语法 3 个一级指标及其下辖的 16 个二级指标,分别考察这些语言特征对蒙古国汉语学习者作文质量的影响.

1 研究设计

1.1 研究问题

研究问题包括以下几个方面:1)词汇、文字、语法是否影响蒙古国汉语学习者作文质量? 2)3 个特征中哪个特征对蒙古国汉语学习者总体作文质量影响最大? 3)不同分数段的作文,其质量受词汇、文字、语法的影响是否不同? 4)16 个二级指标中,哪个是最显著的评分参考指标?

1.2 语料来源及处理步骤

本文从北京语言大学 HSK 动态作文语料库中检索到 56 篇蒙古国汉语学习者作文语料,按照 56 篇的作文分数划分为 3 个阶段,即 A[80,90]、B(60,80)、C[50,60].

语料处理步骤如下:首先将作文语料导出并整理成电子文本,使用 Corpus Word Parser 语料库分词和词性标注程序^①统计词数、词种数等指标;其次利用汉语考试服务网^②中的《国际中文教育中文水平等级标准》(2021)进行词汇的等级统计和词性统计;最后将统计数据按照作文分数从高到低排列,导入 SPSS 数据软件

收稿日期:2024-05-22;**修回日期:**2024-06-17.

基金项目:国家自然科学基金(32402398);河南省研究生教育改革与质量提升工程项目(YJS2023SZ22);河南省教师教育课程改革研究项目(2023-JSJJYB-018);河南省哲学社会科学规划项目(2023BYY007).

作者简介(通信作者):刘志芳(1980—),女,河南三门峡人,河南师范大学副教授,博士,研究方向为汉语词汇和词汇教学, E-mail:lzf211@163.com.

引用本文:刘志芳,马廷芳.汉语二语学习者作文质量影响因素研究[J].河南师范大学学报(自然科学版),2025,53(3): 112-120.(Liu Zhifang, Ma Tingfang. A study on the influencing factors of the composition quality learners of Chinese second language[J]. Journal of Henan Normal University (Natural Science Edition), 2025, 53(3): 112-120. DOI:10.16366/j.cnki.1000-2367.2024.05.22.0002.)

① website:www.cncorpus.org Corpus Word Parser 语料库分词和词性标注程序.

② 汉语考试服务网(www.chinesetest.cn)《国际中文教育中文水平等级标准》查询系统.

分析结果.

1.3 研究指标的选取与分析

本研究首先建立作文分数为因变量,自变量为词数^①、词种数、初级词数、中级词数、高级词数、总字数^②、错字数、别字数、漏字数、多字数、关联词语数、人称代词数、连词数、副词数、助词数、介词数的数据集.被试者均为蒙古国汉语学习者,按照作文分数划分为3个分数段,分别是高分段[80,90]、中分段(60,80)和低分段[50,60],高分段作文11篇,中分段作文30篇,低分段作文15篇,另外高分段和低分段的分数都是10分的差距,只有中分段的作文有20分的差距.

其次,在词汇、文字、语法3个特征下,细分为16个具体指标.运用数理统计法,统计作文语料的自变量,将蒙古国汉语学习者的56篇作文语料按照词汇、文字和语法3个特征统计为3个工作表,并对每个特征下所含的具体指标进行详细分析,方便导入SPSS数据分析软件研究数据结果.

词汇特征按照词汇量、词汇等级和词汇丰富性划分为词数、词种数、初级词数、中级词数和高级词数5个指标;文字特征按照HSK字处理^③划分指标进行分类,即[C]错字数、[B]别字数、[L]漏字数、[D]多字数和总字数5个指标;语法特征在前人研究基础上选取在语篇中最值得关注的指标,即关联词语数、人称代词数、连词数、助词数、副词数和介词数6个指标.连词是关联词语的一种,但本研究发现蒙古国汉语学习者作文语料中除连词之外,关联词语的使用较为显著,因此分开考察这两个指标.

最后,通过语料库分词和词性标注程序将作文语料的电子文本进行切分,统计作文的总字数、总词数和词汇种类数量,在汉语考试服务网中查询切分出来的词汇的等级并统计.通过词性标注统计关联词语、人称代词、连词、助词、副词、介词的数量.

2 数据分析

2.1 描述性统计结果

2.1.1 数据集各变量描述分析

本文对数据集中各变量的分布情况进行了数据统计,蒙古国汉语学习者的56篇作文语料的作文分数均值为68.13分,表明作文分数整体在合格以上,作文质量有待整体提高.字数的均值为349.64字,词数的均值在236左右.HSK(高等)写作评分标准^[5]中对篇幅长度的要求为400~600字,因此蒙古国汉语学习者56篇作文语料的篇幅长度为中等水平.由作文分数和篇幅长度的均值可知,蒙古国汉语学习者整体上的作文质量为中等,有很大的提升空间.词汇等级中初级词数远大于中级词数和高级词数,人称代词数、助词数、副词数的分布均值相差不大,由此可知,教师应在写作教学中着重强调中级词和高级词.

2.1.2 高分段语料描述性分析

本研究共收集蒙古国汉语学习者高分段作文语料11篇,其中最高分为90分,最低分为80分,作文字数均值达到420字,符合HSK高等作文评分标准中对篇幅长度的要求.词种数在3个分数段中分布最多.但是,与中分段和低分段作文语料相比,高分段作文中初级词数均值高于中分段和低分段;中级词数均值高于低分段,低于中分段;高级词数均值低于中分段和低分段.词汇等级分布结果显示词汇特征在3个分数段中的影响程度有待进一步探究.

2.1.3 中分段语料描述性分析

中分段作文语料数量较多,一共有30篇.中分段作文最高分为75分,最低分为65分,作文总字数均值约为368字.与高分段相比,中分段作文错字数、别字数、多字数均值较高,而漏字数均值远低于高分段作文;与低分段相比,中分段作文的漏字数均值与其相等,错字数和别字数均值较低,反而多字数均值高于低分段

① 由于语篇中有超纲词出现,本文选取自变量初级词、中级词、高级词,不统计超纲词的数量,因此还需单列出词数这一总指标进行统计.

② 本研究总字数指标作为语料篇幅长度的统计,与文字特征下辖二级指标并列统计说明.

③ “HSK动态作文语料库”<http://hsk.blcu.edu.cn/>语料标注及代码说明.

作文.由此可知,教师在教学中要关注高分段学习者作文写作中存在的漏字现象;注重引导中分段学生写作中多字的写作偏误;对低分段学生应打好汉字识记基础,避免错别字的出现.

2.1.4 低分段语料描述性分析

低分段作文共有 15 篇,其中最高分为 60 分,最低分为 50 分,作文总字数均值约为 268 字,与中、高分数段作文篇幅长度相比,低分段作文篇幅长度预测作用较强.低分段作文与高分段作文和中分段作文在词汇特征和文字特征的各个指标上呈现不同程度的差别,在语法特征的 6 个指标中,低分段作文的均值都低于中分段和高分段作文,由此可见,低分段作文学习者的写作弱势体现在语法特征方面.特别是人称代词和助词方面,在教学中,应该着重提醒学生人称代词的使用和助词的区分.

2.2 相关与回归分析结果

2.2.1 3 个特征与作文质量回归分析

词汇、文字、语法 3 个特征对作文质量的预测作用归纳在回归分析中,结果如表 1 所示.

表 1 3 个特征与作文质量的回归分析汇总表

模型	r	r^2	调整 r^2	标准估计 的误差	更改统计量					Durbin- Watson
					r^2 更改	F 更改	df1	df2	Sig. F 更改	
1	0.748	0.560	0.516	7.069	0.560	12.728	5	50	0.000	1.033
2	0.783	0.613	0.574	6.629	0.613	15.843	5	50	0.000	1.446
3	0.695	0.484	0.420	7.736	0.484	7.649	6	49	0.000	0.721

注:模型 1.词汇特征预测变量为(常量),高级词数,初级词数,词种数,中级词数,词数.
模型 2.文字特征预测变量为(常量),多字数 D,漏字数 L,总字数,错字数 C,别字数 B.
模型 3.语法特征预测变量为(常量),介词数,关联词语数,人称代词数,助词数,副词数,连词数.
因变量为作文分数.

词汇特征与作文质量回归模型显示,词汇特征 5 个指标能够解释作文质量 56.0%的变异($r^2=0.560$),即本研究选取的词数、词种数、初级词数、中级词数、高级词数 5 个指标对作文质量的影响程度为 56.0%.对于一篇作文来说,词汇特征的影响占到一半以上的比例,说明词汇指标对蒙古国汉语学习者作文质量的影响是显著的.

文字特征与作文质量回归模型结果显示,文字特征能够解释作文质量 61.3%的变异($r^2=0.613$).本文按照 HSK 动态作文语料库字处理方法选取了文字特征的 5 个指标,即总字数、错字数、别字数、漏字数和多字数,结果表明,这些指标对作文质量的影响是显著的,按照效应量比较,词汇和文字特征对作文质量的解释力都在 50.0%以上,占到预测作用的一半,由此可见,词汇和文字特征对作文质量的预测贡献最大.

语法特征与作文质量回归模型显示,语法特征中关联词语数、人称代词数、连词数、助词数、副词数和介词数 6 个指标能够解释作文质量 48.4%的变异($r^2=0.484$).结果表明,与词汇特征和文字特征相比,语法特征对作文质量的影响程度次之.

2.2.2 3 个特征下辖的指标与作文质量相关分析

本研究对词汇、文字、语法下辖的各指标与作文质量进行了相关分析^①,数据结果^②见表 2.

由表 2 可知,词汇特征内部的 5 个指标中,词数和初级词数与作文质量相关性强($r=0.758;r=0.686;p<0.01$).词种数、高级词数和中级词数与作文质量的相关性不明显,且依次递减.

在文字特征内部,只有总字数即篇幅长度这一个指标与作文质量相关性强($r=0.703;p<0.01$).而错字数与作文质量呈较弱的负相关,剩下 3 个指标与作文质量的相关性不明显,别字、漏字、多字现象的出现有时不会影响语篇的整体阅读和意思的表达,因此相关系数非常小且可能为正相关.

① 注: * * 在 0.01 水平(双侧)上显著相关, * 在 0.05 水平(双侧)上显著相关.
② 相关系数 r 满足:当 $|r| \geq 0.5$ 时,认为 A 和 B 有强的相关性;当 $0.3 \leq |r| < 0.5$ 时,可以认为有弱的相关性;当 $|r| < 0.3$ 时,认为没有相关性.数值的正负代表相关是正向相关或者负向相关.

在语法特征内部 6 个指标中,其中 5 个指标与作文质量相关性较为均衡且一般,相关性从强到弱依次为助词数、人称代词数、副词数、关联词语数、介词数($r=0.640$; $r=0.550$; $r=0.497$; $r=0.451$; $r=0.444$; $p<0.01$).

表 2 3 个特征指标与作文分数相关分析结果

Tab. 2 Analysis results of correlation between three characteristic indexes and essay scores

词汇特征	词数	词种数	初级词数	中级词数	高级词数	
作文分数	0.758**	0.205	0.686**	0.073	0.116	
文字特征	总字数	错字数	别字数	漏字数	多字数	
作文分数	0.703**	-0.209	0.019	0.135	0.017	
语法特征	关联词语数	人称代词数	连词数	助词数	副词数	介词数
作文分数	0.451**	0.550**	0.183	0.640**	0.497**	0.444**

由以上分析可知,词汇、文字、语法 3 个特征下辖的二级指标对作文质量的影响显著性不同,其中词汇特征中词数、初级词数与作文质量强相关($r=0.758$; $r=0.686$; $p<0.01$),文字特征中总字数与作文质量强相关($r=0.703$; $p<0.01$),语法特征中助词数、人称代词数、副词数、关联词语数、介词数与作文质量呈现由强到弱的相关性($r=0.640$; $r=0.550$; $r=0.497$; $r=0.451$; $r=0.444$; $p<0.01$).其余指标与蒙古国汉语学习者作文质量的相关性不强.

以上相关与回归分析结果显示,词汇、文字、语法对蒙古国汉语学习者作文质量存在影响,但词汇、文字、语法 3 个特征下辖的二级指标对作文质量的影响程度不均衡,其中词汇特征下辖的词数指标因其系数最大,所以相关性最强.

2.2.3 3 个分数段语料各变量相关分析

本研究对 3 个分数段的作文语料进行区分后,分析不同分数段作文语料各变量的相关性,变量间 Pearson 相关系数有显著的变化,凸显了 3 个分数段中各二级指标对作文质量的影响有明显的不同,呈现出不同程度的预测作用.

高分段 11 篇语料中,变量间 Pearson 相关系数有明显的变化.高分段作文中各特征具体指标与作文质量呈强相关的排列为助词数、别字数、多字数、高级词数($r=0.981$; $r=0.979$; $r=0.944$; $r=0.858$);相关性次之的排列为词种数、初级词数、中级词数($r=0.765$; $r=0.675$; $r=0.667$),总字数、人称代词数、词数($r=0.648$; $r=0.640$; $r=0.544$).其余指标与作文质量显著相关性较弱.

中分段 30 篇语料中变量间相关分析结果表明,Pearson 相关系数有明显的变化.各特征具体指标与作文质量呈强相关的排列为中级词数、错字数、高级词数、别字数、副词数和连词数($r=0.897$; $r=0.726$; $r=0.654$; $r=0.652$; $r=0.518$; $r=0.514$).

低分段 15 篇语料变量间 Pearson 相关系数有明显的变化.各特征具体指标与作文质量呈强相关性的排列为人称代词数、助词数、词种数($r=0.989$; $r=0.971$; $r=0.818$),显著相关性次之的排列为中级词数、高级词数、别字数、错字数($r=0.672$; $r=0.640$; $r=0.587$; $r=0.453$).

由此可知,不同分数段的作文质量,其影响因素相关性有显著的差异.高分段中词汇特征中的 5 个指标都与作文质量强相关,语法特征中的助词数与作文质量的相关性最强,其次是文字特征中的别字数和多字数.中分段中词汇、文字、语法特征中均有两个指标与作文质量强相关,其中词汇特征中的中级词数相关系数最高,文字特征中的错字数次之.低分段中语法特征中的人称代词数和助词数指标与作文质量强相关,词汇特征中的词种数与作文质量同样强相关.

2.3 回归诊断分析结果

2.3.1 变量间多重共线性检验

在回归分析结果中,词汇、文字和语法 3 个特征都能够影响作文质量,但要解决 16 个二级指标中哪个是最显著的评分参考因素的问题,还需要使用多元线性回归方法构建评分模型,为确保模型解释变量间不存在高度共线性,研究计算了所有因子的方差膨胀系数(variance inflation factor, VIF),当因子 VIF 值大于 10 时,认为因子间存在严重共线性,这些因子不被同时加入模型中.结果见表 3.

表 3 变量间共线性分析结果汇总表

Tab. 3 Summary table of colinearity analysis results between variables

模型	共线性统计量 1		共线性统计量 2	
	容差	VIF 值	容差	VIF 值
词数	0.017	58.025	0.116	8.612
初级词数	0.019	52.704		
总字数	0.026	37.817		

由表 3 可知,词数 VIF 值是 58.025,初级词数 VIF 值是 52.704,总字数 VIF 值是 37.817,反映了 3 个变量间存在多重共线性问题.由于总字数和初级词数相关性较强($r=0.703$; $r=0.686$),因此逐一删除存在严重共线性问题的自变量总字数和初级词数之后,剩余 14 个自变量之间不存在多重共线性问题,排除了干扰回归模型构建结果不稳定的因素,可构建线性模型.对自变量中异常值和离群点的处理,本文采用均值代替的方法,以确保作文语篇样本数量的不变.共线性问题解决之后对自变量进行主成分因子分析,剔除不重要的变量,以确保回归模型的准确性和可靠性.

2.3.2 变量间主成分因子分析

本研究的降维方法主要采用主成分因子分析,将不存在共线性问题的 14 个自变量带入主成分分析,采取最大方差法进行因子旋转,进一步明晰因子结构并且找出因子和 14 个自变量的对应关系,方便后续提取显著性的指标进行回归模型的构建,一般来说特征值大于 1 的因子都是值得关注的指标.因子分析一共提取出 3 个因子,特征根值均大于 1,旋转后的方差解释率分别是 29.067%、12.417%、12.000%,旋转后累积方差解释率为 53.483%.最大方差旋转法提取出的 3 个因子与 14 个自变量有一定的对应关系,进一步研究因子对于自变量的提取情况,结果如表 4.

表 4 旋转后因子载荷系数表

Tab. 4 Table of factor load coefficients after rotation

类别	成分			公因子方差		类别	成分			公因子方差
	1	2	3				1	2	3	
词数	0.931	0.027	0.065	0.872		多字数 D	0.130	-0.216	-0.489	0.302
词种数	0.061	-0.394	0.530	0.440		关联词语数	0.766	-0.201	-0.235	0.683
中级词数	0.227	0.789	-0.169	0.703		人称代词数	0.732	-0.203	0.321	0.681
高级词数	0.152	0.751	0.183	0.621		连词数	0.621	-0.149	-0.193	0.444
错字数 C	-0.070	0.509	0.379	0.408		助词数	0.792	0.107	0.042	0.640
别字数 B	0.339	0.096	-0.488	0.363		副词数	0.795	-0.074	0.040	0.639
漏字数 L	-0.068	-0.347	0.365	0.258		介词数	0.502	0.164	0.394	0.434

从表 4 可知,除了别字数、漏字数和多字数 3 个自变量,剩余自变量的公因子方差均高于 0.4,表明剩余 11 个自变量与 3 个因子之间有着较强的关联性,因子可以有效提取出信息,由上可以得出成分(C)公式如下:

$$C = (0.291/0.535) \times C_1 + (0.124/0.535) \times C_2 + (0.12/0.535) \times C_3 = 0.544 \times C_1 + 0.232 \times C_2 + 0.224 \times C_3.$$

通过上述公式可知, C_1 (成分 1)的权重为 54.4%、 C_2 (成分 2)的权重为 23.2%、 C_3 (成分 3)的权重为 22.4%,由此本文提取的显著性变量为成分 1 中的词数、成分 2 中的中级词数和高级词数、成分 3 中的词种数,共 4 个变量作为建立回归模型的指标.

3 评分模型构建与检验

经过回归诊断分析后,本文选取词数、词种数、中级词数和高级词数 4 个指标构建评分预测模型.由于因变量作文分数与自变量词汇特征间存在关系显著的线性关系,因此本研究探讨自变量与因变量之间的关系

时选择多元线性回归模型,多元线性回归方法^[6]可以通过确认因变量和自变量的关系,拟合出预测模型,目前广泛用于作文自动评分系统中,如作文局部连贯评价模型的构建^[7].本文参考自动评分系统模型,选择多元逐步回归法,将提取出的与作文质量有显著相关的自变量进行显著性检验,运用逐步法逐个引入或剔除自变量,得到最优的回归方程.

3.1 逐步线性回归模型构建

研究将提取出的词数、词种数、中级词数和高级词数 4 个变量中存在的异常值和离群点进行检验和处理,之后将其导入 SPSS 进行逐步回归模型构建,使用逐步线性回归对变量进行分析,将不显著的指标剔除在模型外.全模型公式如下:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon,$$
 (1)

其中, y 是因变量作文分数, β_0 为截距(表示当自变量为 0 时,因变量的值), X_1 、 X_2 、 X_3 和 X_4 分别为词数、词种数、中级词数和高级词数 4 个特征变量指标, $\beta_i (i = 1, 2, 3, 4)$ 均为常数且代表各特征变量对分数 y 的正或负效应量, ϵ 为随机误差.

剔除不显著的指标后,词数这一具体指标与作文质量显著相关.模型的复相关系数 $r = 0.707$, 决定系数 $r^2 = 0.500$, 调整后决定系数 $r^2 = 0.491$, 说明进入模型的变量能解释人工评分的 50.0% 的变异, 即词数指标能够解释作文质量 50.0% 的变异, 具有统计学意义.此模型对作文质量的解释程度较明显, 但显著性指标存在数量较少, 这与本研究选取指标有限和作文语料偏少有关, 有待增加变量进一步分析.确认上述模型具有统计学意义之后, 本文对逐步线性回归模型进行显著性检验, 结果显示各指标的 Sig 值均小于 0.05, 可以认为模拟出的回归方程中, 自变量与因变量存在线性关系, 方程中的系数皆具有显著性, 是有效的参考指标.

根据线性回归预测模型, 因变量作文分数与自变量词数的对应的最优回归方程公式如下:

$$y = 40.365 + 0.117 X_1,$$
 (2)

其中, y 为因变量作文分数, 40.365 为常量, 0.117 是常数且代表词数这一特征变量对作文分数的正效应量, X_1 为作文语篇中的词数指标.

从表 5 可知, Sig 值小于 0.05, 说明该模型自变量与因变量之间存在显著的线性关系.为检验词数与作文分数间是否存在非线性关系, 向线性模型中加入词数的二次项.该二次项结果不显著(系数为 -1.0528 , $p = 0.181$), 因此词数与作文质量间非线性关系不显著.

表 5 逐步线性回归预测模型

Tab. 5 Stepwise linear regression prediction model

类别	非标准化系数		<i>t</i>	Sig.	共线性统计量	
	<i>B</i>	标准误差			容差	VIF 值
模型常量	40.365	3.902	10.346	0.000		
模型词数	0.117	0.016	7.345	0.000	1.000	1.000

3.2 预测模型效度检验

预测模型有效性分析主要通过观察回归预测模型的残差图和散点图, 一方面观察残差图变量的分布是否正态、是否左右偏离和存在异常值点, 可用以协助观察模型是否拟合原成绩.另一方面观察残差正态 P-P 图, 分析预测模型是否呈正态分布, 若 P-P 图的数据点排列趋向直线, 则表明正态性较强, 否则表明正态性较弱^[8].

由图 1 可知, 预测模型生成的直方图与 P-P 图显示, 该模型中直方图高峰只有一个, 峰值较为居中, 标准化残差绝对值小于等于 3, 并未观测到异常值点.观察 P-P 图, 可以发现散点较为接近直线, 拟合度较高.因此该模型经过检验对于解释作文质量的变异有效值较高.

3.3 预测模型精度检验

对于模型精度的检验, 本文考虑到作文语篇数量的有限性, 以及本研究不具备采用人工评分进行辅助验证作文分数的条件, 因此采用 SPSS 数据分析软件随机抽取数据集的部分数据作为验证集, 即按照随机原则从总体中抽取一部分单位进行观察, 并运用数理统计的原理, 以被抽取的那部分单位的数量特征为代表, 代入模型中确保对模型性能的准确评估^[9].魏培文等^[10]将 220 项数据中的 190 项即 86.0% 的数据作为训练样

本,剩余的 30 项即约为 14.0%的数据作为测试样本进行仿真测试,验证预测模型分类结果.因此为避免误差过大,本研究从已知的 56 篇作文语料中随机抽取 15.0%的作文语料作为数据验证集,代入预测模型回归方程式观察作文分数的分布情况,并检测原始作文分数与预测作文分数之间是否存在 Pearson 相关性,并对 Pearson 相关系数进行解释说明,检验预测模型能够多大程度解释作文的质量.检验结果表明,原始作文分数与预测作文分数 Pearson 相关性系数为 0.852,大于 0.5,且显著性(双侧)小于 0.05,说明测试集数据与验证集数据高度相关,相关性达到 85.2%,本研究的预测模型通过精度检验,拟合度较高.

通过构建多元线性回归模型,本文选取的 16 个指标中对作文评分最显著有效的参考标准为词数变量,本研究选取的指标有数据支撑,并能够证明其有效性.

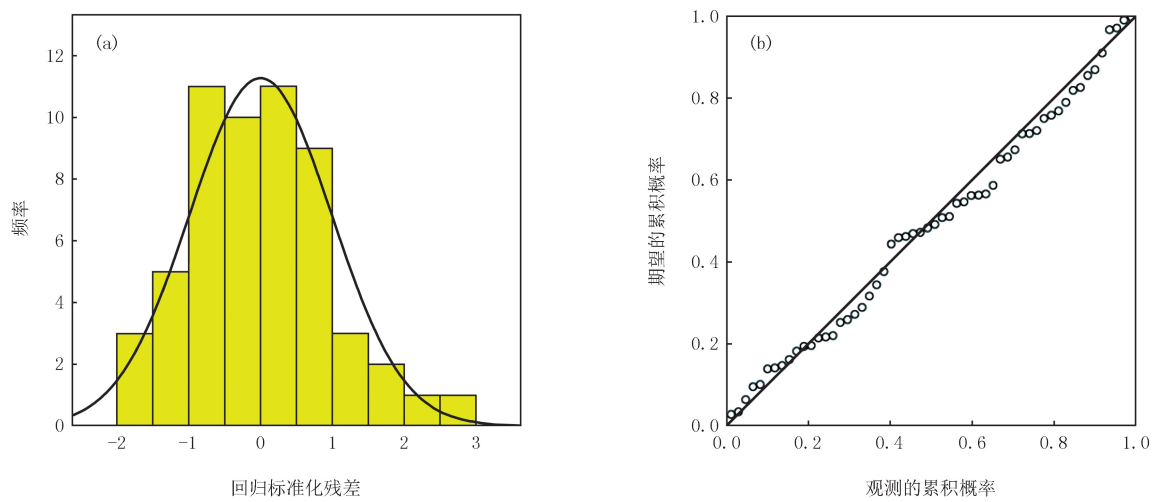


图1 回归方程标准化残差直方图(a)和P-P图(b)

Fig.1 Regression equation standardized residual histogram(a) and P-P diagram(b)

4 研究结果

本研究从词汇、文字、语法 3 个角度考察了蒙古国汉语学习者作文质量的影响因素,研究表明:词汇、文字、语法 3 个特征均可以在一定程度上预测蒙古国汉语学习者的作文质量,其中词汇特征对作文质量影响最大,3 个特征及其下辖的二级指标对不同分数段的作文质量影响程度不同,16 个二级指标中最显著的评分参考指标是词数.

4.1 词汇、文字、语法 3 个特征均能影响学习者的作文质量

词汇、文字、语法都对蒙古国汉语学习者的作文质量有影响,但影响程度不均衡.蒙古国汉语学习者整体上的作文质量为中等.按照词汇、文字、语法 3 个特征与作文质量的回归分析结果显示,词汇特征下辖的 5 个指标对蒙古国汉语学习者作文质量的预测能够达到 56.0%,文字特征下辖的 5 个指标的预测率能够达到 61.3%,而语法特征下辖的 6 个指标的预测率只有 48.4%.由此可见,词汇、文字、语法 3 个特征都能够影响作文质量,且影响的程度不均衡.通过将 3 个特征及其下辖的 16 个二级指标与作文质量进行相关分析,结果显示文字特征中只有总字数与作文质量强相关,这主要是因为 HSK 评分标准中作文字数是评定作文等级的重要参考因素.

4.2 词汇特征对蒙古国汉语学习者作文质量影响最大

虽然词汇、文字、语法 3 个特征与作文质量的回归分析结果显示文字特征对蒙古国汉语学习者作文质量的预测效应量最大,但随着研究的深入,在对 3 个特征及其下辖的 16 个二级指标与作文质量进行相关性分析时,研究发现,词汇特征下辖的 5 个指标中,词数和初级词数与作文质量相关性强($r=0.758$; $r=0.686$; $p<0.01$),文字特征下辖的 5 个指标中,只有总字数这 1 个指标与作文质量相关性强($r=0.703$; $p<0.01$),语法特征下辖的 6 个指标与作文质量相关性一般.由此可见,文字特征中与作文质量强相关的二级指标数最少,而且相关系数比词汇特征下辖的二级指标词数的相关系数小,文字特征和语法特征下辖的各指标与作文

质量的相关性不如词汇特征下辖的二级指标,这3个特征对作文质量的影响是不均衡的,其中词汇特征对蒙古国汉语学习者作文质量影响最大。

4.3 3个特征及其下辖的二级指标对不同分数段的作文质量影响程度不同

按照作文成绩将蒙古国汉语学习者的作文分为高分段、中分段和低分段3个分数段。研究结果表明,3个特征及其下辖的二级指标在不同分数段中的影响程度不同。具体来说,与高分段、中分段和低分段作文质量强相关的指标分别是语法特征下辖的助词数,词汇特征下辖的中级词数和语法特征下辖的人称代词数。与不同分数段作文质量强相关的二级指标不同,主要集中在词汇特征和语法特征,这也再次印证了文字特征不是最为显著的作文评分变量。不同分数段的作文质量,其影响因素相关性有显著的差异。高分段中词汇特征下辖的5个指标都与作文质量强相关,语法特征中的助词数与作文质量的相关性最强,文字特征中的别字数和多字数与作文质量相关性较强。中分段中词汇、文字、语法特征中均有两个指标与作文质量强相关,其中词汇特征中的中级词数相关系数最高,文字特征中的错字数次之。低分段中语法特征中的人称代词数和助词数指标与作文质量强相关,词汇特征中的词种数与作文质量强相关。针对不同分数段的作文,3个特征下辖的各二级指标具有可分析性和显著性,在写作教学中可有针对性地扬长避短,提高作文质量。

4.4 词汇特征下辖指标词数是最显著的评分参考因素

在描述性分析中,词数的均值在236个词左右,因此蒙古国汉语学习者56篇作文语料的篇幅长度为中等水平。在相关性分析中,词汇特征下辖指标词数和初级词数与作文质量相关性强($r=0.758$; $r=0.686$; $p<0.01$)。在回归性分析中,词汇特征5个指标能够解释作文质量56.0%的变异($r^2=0.560$),即词数、词种数、初级词数、中级词数、高级词数5个指标对作文质量的影响程度为56.0%。不同分数段中,词数这一指标同样具有显著性。运用多元线性逐步回归法构建评分模型,数据显示词数是进入模型的唯一变量,能够解释人工评分50.0%的变异。经过回归诊断后提取出影响显著的4个指标构建回归模型,通过分析多元线性回归方程的残差图和散点图,分析了回归模型的有效性,通过验证集和原始测试集的相关系数分析,检验了回归模型的精度,结果显示词数指标是最显著的评分参考因素。

总的来说,蒙古国汉语学习者作文中3个特征的具体指标均有预测作用,在评分时应该全面考虑,写作教学时应针对不同分数段的学习者有针对性地调整教学重难点。

5 结论与建议

本研究发现,蒙古国汉语学习者的作文质量受词汇、文字、语法3个客观指标的影响是可预测的。据多元线性回归模型,检验出对自动评分最显著有效的参考指标是词汇特征的二级指标词数,但词汇、文字、语法3个特征在评分标准中应受到同等重视。

基于以上研究结果,对汉语二语学习者写作教学的建议是,教师应在确保学生书写文字正确性的基础上,强调词汇学习的重要性,采用创新性教学方法,扩大学生中、高级词在写作时的产出量,另外还要引导学习者注重不同体裁作文语篇连贯^[1],在语篇衔接和内容逻辑方面加强训练,提高连词成句的准确度及成段表达的熟练度。总之,在写作教学时,教师应注重语言特征词汇、文字、语法的综合运用,三者是相互关联的,词汇更是最显著的影响因素,在提高文字特征正确率的基础上,运用丰富的词汇,形成正确的语法表达,提高作文整体质量。

参 考 文 献

- [1] 吴继峰,周蔚,卢达威.韩语母语者汉语二语写作质量评估研究:以语言特征和内容质量为测量维度[J].世界汉语教学,2019,33(1):130-144.
WU J F,ZHOU W,LU D W.Assessing Chinese L2 writing quality on basis of language features and content quality[J].Chinese Teaching in the World,2019,33(1):130-144.
- [2] 任春艳.HSK作文评分客观化探讨[J].汉语学习,2004(6):58-67.
REN C Y.Exploratory research on objective scoring of HSK composition[J].Chinese Language Learning,2004(6):58-67.
- [3] 王浩学,程勇,胡晓清.基于词汇特征与语法模式的汉语二语写作质量动态发展研究[J].华文教学与研究,2022(2):20-31.
WANG H X,CHENG Y,HU X Q.A dynamic development study of CSL writing quality based on lexical features and grammatical pat-

terns[J].TCSOL Studies,2022(2):20-31.

[4] 吴佩,邢红兵.内容、词汇、篇章特征对汉语学习者二语作文质量的影响研究[J].语言教学与研究,2020(2):24-32.
WU P,XING H B.The influence of content,lexical and discourse features on the quality of CSL learners' L2 writing[J].Language Teaching and Linguistic Studies,2020(2):24-32.

[5] 聂丹.汉语水平考试(HSK)写作评分标准发展概述[J].云南师范大学学报(对外汉语教学与研究版),2009,7(6):15-20.
NIE D.A historical account of the assessment criteria for HSK writing[J].Journal of Yunnan Normal University(Teaching and Research on Chinese as a Foreign Language Edition),2009,7(6):15-20.

[6] 黄晓玉,王兰会.SPSS 24.0 统计分析:在语言研究中的应用[M].北京:中国人民大学出版社,2021.

[7] 刘国兵.基于 WordNet 语义知识库的英语学习者作文局部连贯自动评价[J].河南师范大学学报(自然科学版),2016,44(6):149-158.
LIU G B.Automatic evaluation of local coherence in Chinese EFL learners' essays with WordNet[J].Journal of Henan Normal University (Natural Science Edition),2016,44(6):149-158.

[8] 付强,杨壮,董锁成,等.河南省国家级传统村落空间可达性及影响因素研究[J].河南师范大学学报(自然科学版),2021,49(6):82-90.
FU Q,YANG Z,DONG S C,et al.Research on spatial accessibility and influencing factors of national traditional villages in Henan Province[J].Journal of Henan Normal University(Natural Science Edition),2021,49(6):82-90.

[9] 李洁明,祁新娥.统计学原理[M].6 版.上海:复旦大学出版社,2014:244-248.

[10] 魏培文,朱珂,叶海智,等.基于 BP 神经网络的高校教师精准教学能力评价模型构建[J].河南师范大学学报(自然科学版),2024,52(5):108-116.
WEI P W,ZHU K,YE H Z,et al.Construction of precision teaching ability evaluation model for college teachers based on BP neural network[J].Journal of Henan Normal University(Natural Science Edition),2024,52(5):108-116.

[11] 刘国兵,常芳玲.向心理论应用于英语学习者书面语篇局部连贯评价的有效性研究[J].河南师范大学学报(自然科学版),2020,48(4):19-28.
LIU G B,CHANG F L.A study on the efficiency of centering theory in evaluating local coherence of English learners'written language[J].Journal of Henan Normal University(Natural Science Edition),2020,48(4):19-28.

A study on the influencing factors of the composition quality learners of Chinese second language

Liu Zhifang, Ma Tingfang

(College of Liberal Arts, Henan Normal University, Xinxiang 453007, China)

Abstract: Taking 56 compositions of Chinese language learners in Mongolia from the HSK dynamic composition corpus as the research object, this paper uses corpus investigation method, data analysis method and multiple linear regression method to investigate the influence of three features of vocabulary, characters and grammar on the composition quality of Chinese language learners in Mongolia. The results show that the three features of vocabulary, characters and grammar can affect the composition quality of learners, among which vocabulary has the greatest impact on composition quality. The influence of various features in the compositions with different score segments is uneven. After the regression diagnosis, the multiple linear regression model is constructed by selecting four indexes:word number, word type number, intermediate word number and advanced word number, and the most significant and effective reference index for automatic scoring is the index word number under lexical features.

Keywords: Chinese second language learner; Chinese language learners in Mongolia; composition quality; influence factor

[责任编辑 刘洋 杨浦]