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

吴继峰[1]　　周　蔚[2]　　卢达威[3]

[1]首都师范大学国际文化学院　[2]首都师范大学心理学院　[3]中国人民大学文学院

提要　本文以210名不同汉语水平的韩语母语者为研究对象，采用相关和多重回归的统计方法，首先从总体上考察词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言特征与写作成绩、内容质量评分的关系，然后考察内容质量评分和写作成绩的相关关系，最后考察语言水平对6个语言测量指标和写作成绩关系的影响效应。研究发现，从总体上看，6个语言特征均与写作成绩、内容质量评分显著相关，并且都能显著预测写作成绩，其中汉字正确性、语法正确性和词汇正确性贡献程度最大，且前两者达到中等效应量。除词汇多样性外，其他5个语言特征都能显著预测内容质量评分，其中词汇复杂性贡献程度最大，达到中等效应量。内容质量评分和写作成绩达到高相关。另外，研究还发现语言水平与汉字正确性存在交互效应，与中高级水平相比，初级水平汉字正确性对写作成绩的预测作用更大。与此同时，本文结合学生作文表现、对评分员的访谈以及对测量指标内部构念的分析，对造成语言各维度对写作成绩和内容质量分数不同贡献程度的原因以及写作质量和内容质量的关系进行了相应的质性分析。

关键词　汉语二语　写作质量　内容质量　区别性特征　评估

一　引言

在第二语言评估研究中，学习者具体语言的使用及所体现的特征已成为口语测试和写作测试评分的热点话题(Crossley et al.，2015；金檀等，2016；Plakans et al.，2016)。其中，热门研究议题之一是区别性特征的选择及使用，具体包含以下两个研究问题：如何选择最具代表性的区别性特征来构建评分标准(Jin & Mak，2013)，是否应在评分标准每一等级中都使用相同的特征(Humphry & Heldsinger，2014)，这两个问题亟待解决(参看金檀等，2016)。有关研究结果具有广泛的应用价值，可以量化部分区别性特征并用于机器自动评分。只有

* 本研究得到教育部人文社会科学研究青年项目“汉语二语者语言产出的任务复杂度效应研究”(18YJC740114)、北京市委组织部青年骨干个人项目(2017000020124G153)和北京市教委社科一般项目(SM201710028014)的资助。《世界汉语教学》匿名审稿专家和江新老师及陆小飞老师、金檀老师为本文提出了宝贵的修改意见，谨此一并致谢！

在区别性特征研究清楚的基础上，才能在机器自动评分中选择特征和设置权重。可见，评分标准中的区别性特征研究是一项十分重要的基础研究，具有很大的研究价值。

综观 GRE、TOEFL、AP 语言测试、英语四六级考试、HSK 考试的作文评分标准，其作文评分涉及的区别性特征可以分为语言特征和内容质量两部分。其中，语言特征指应试者在词汇、语法、汉字等语言维度上的表现，词汇维度又可分为词汇复杂性、词汇多样性、词汇正确性等子维度，语法维度分为句法复杂度、语法正确性等子维度（Banerjee et al.，2015）。内容质量是指学习者表达的内容达到交际目标的程度，在国际上不同研究者使用的术语有所不同，例如使用较多的有功能充分性（functional adequacy）、交际充分性（communicative adequacy）。目前内容质量也是国际二语评估研究中较热的研究领域，因为之前大部分研究是围绕语言区别性特征进行的，而关注内容质量对语言产出质量评估影响的研究却少之又少。近些年逐渐有研究者（辛平，2007；Pallotti，2009；De Jong et al.，2012）提出，只使用语言特征的测量来评估学习者表现成功与否是不够的，必须考虑学习者产出的内容质量。关于二语写作内容质量评估的探讨，Kuiken、Vedder 等进行了一系列研究，呼吁建立一个新的写作评分量表，关注二语写作的功能充分性。他们在实证研究的基础上，设计了功能充分性的评分量表，具体包含内容（content）、任务需求（task requirements）、可理解性（comprehensibility）、衔接和连贯（cohesion and coherence）四项，每一项包括 6 个等级（Kuiken et al.，2010；Kuiken & Vedder，2014、2017）。这个量表的效度如何，有待实证研究的进一步验证。

目前在汉语作为第二语言的评分研究中，考察区别性特征对汉语学习者产出质量影响的研究数量很少。在口语评分方面，王佶旻（2002）和朱世芳（2009）是较早利用语言特征来分析外国学生口语产出质量的研究。王佶旻（2002）考察了发音、语法、流利性三个语言区别性特征与教师口语评估的关系，研究发现这三个语言特征测量指标均与教师总体评价具有显著的相关关系。朱世芳（2009）分别选取了初、中、高汉语水平的韩国被试各 9 名，从词汇、语法、流利性和语篇连贯性等方面考察了韩国学生的汉语口语语篇特点，研究发现在不同水平上，学习者语言表现的区别性特征存在显著差异。Jin & Mak（2013）以 66 名高级水平的汉语二语学习者为研究对象，从发音、流利性、词汇和语法 4 个维度入手，选择 7 个语言区别性特征，其中，发音维度包含一个区别性特征——类似目标语的音节（target-like syllables），流利度维度包含两个区别性特征——语速、停顿时间，词汇维度包含两个区别性特征——词型数（word tokens）、词种数（word types），语法维度包含两个区别性特征——语法正确性和句法复杂性。他们考察这 7 个语言区别性特征对口语评分的影响。研究发现，7 个语言特征的组合能解释口语评分总变异的 77%－79%，其中，词型数的贡献最大，其次是词种数，标准化系数 Beta 值分别达到 0.60 和 0.52，远远大于其他 5 个区别性特征。简言之，该研究认为词汇维度对高级水平汉语学习者口语评分的影响最大。

在作文评分方面，吴继峰（2016）和王艺璇（2017）考察了词汇维度对汉语二语者写作成绩的影响。吴继峰（2016）以 46 名英语母语者为研究对象，考察词汇变化性、词汇复杂性、词汇密度、词汇错误和写作质量的关系，研究发现，词汇错误、词汇复杂性与写作质量的关系更为密切，四个自变量构成的组合能解释写作成绩总变异的 46.2%。王艺璇（2017）基于北京语言大学“HSK 动态作文语料库”中的 360 篇作文，考察了词汇各维度对写作成绩的影响，研究发现，词种数、词汇错误比重和常用词数三个参项能解释写作总成绩总变异的92.8%。

另外，吴继峰(2018a)以50名中级汉语水平的英语母语者为研究对象，从词汇和语法两个维度，考察词汇多样性、词汇复杂性、词汇正确性、语法正确性和句法复杂性5个语言区别性特征对汉语二语写作成绩的影响，研究发现，5个语言特征均与写作成绩显著相关，但是只有词汇复杂性、词汇正确性和语法正确性能显著预测写作质量，其中语法正确性和词汇正确性的预测能力更强。

以上研究为汉语二语区别性特征与产出质量关系的考察做出了探索性贡献，但是还存在以下研究空间：第一，这几项研究考察的都是语言区别性特征对汉语二语者产出质量评估的影响，均未涉及内容质量这一维度；第二，除了朱世芳(2009)之外，其他几项研究考察的都是某一水平阶段的学习者，没有同时考察初、中、高不同水平的学习者；第三，虽然朱世芳(2009)考察了初、中、高不同汉语水平的学习者，但其被试总共只有27名，数量较少，无法进行多重回归分析；第四，王艺璇(2017)发现仅词汇的三个子维度——词种数、词汇错误比重和常用词数就能解释写作成绩总变异的92.8%，如果该结论成立的话，那么语法、内容、结构等其他维度只能解释写作成绩总变异的7.2%，这是否符合写作评估事实需要进一步验证；第五，吴继峰(2018a)根据新HSK作文评分标准，将汉字错误作为词形错误纳入词汇层面进行统计，没有单独考察汉字维度在作文评分中的贡献程度。故本文以初、中、高不同汉语水平的韩语母语者为研究对象，综合考察词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言区别性特征与写作质量、内容质量的关系，考察写作质量与内容质量的关系，以及汉语水平是否会影响6个语言区别性特征对写作质量的预测作用。另外，通过访谈考察评分员在对写作质量和内容质量评分时主要考虑哪些因素，质性访谈和量化分析的结果是否一致。

二 研究设计

2.1 研究问题

(1)词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言区别性特征与写作成绩的关系如何？哪些特征能有效预测写作质量？预测程度如何？

(2)内容质量和写作质量的关系如何？6个语言区别性特征与内容质量的关系如何？哪些特征能有效预测内容质量？预测程度如何？

(3)具体来看，语言水平(初、中、高)是否会影响以上6个语言测量指标对写作成绩的预测作用？即语言水平与各测量指标对写作成绩的预测是否存在交互作用？

(4)评分员对写作质量和内容质量评分时，主要考虑哪些因素？存在哪些差异？质性访谈的结果和量化分析的结果是否一致？

2.2 语料来源和学生分级标准

语料来自于首都师范大学2015～2017三年的入学分班考试作文《我的爱好》，要求不少于100字，闭卷考试，分班考试总时间为1小时，作文满分9分。参加考试的韩语母语者共210人，初、中、高汉语水平各70人。分级标准为：根据入学分班考试的笔试试卷成绩，总分20～45分定为初级，55～69分定为中级，80分及以上定为高级。

2.3 写作质量和内容质量评估及评分员访谈

写作质量评估：第一步，每次笔试分班考试结束后，笔者逐一鉴别每篇作文的字数是否

达到要求，然后扫描字数达到要求的试卷。第二步，招募 3 名参与过 HSK 或 AP 中文考试作文阅卷的老师，依据《新汉语水平考试》的作文评分标准（见附录 1）对阅卷教师进行统一培训，主要涉及总体内容和结构、语法、汉字和词汇三大块，采用整体性评分的方式，满分 9 分。第三步，为减少阅卷时间对评分员的影响，3 名评分员统一在同一教室和同一时间段对扫描的电子版作文进行评分。为使作文评分更加公正，我们的评分分为三步——两轮试评和一轮正式评分。第一轮，评分员首先根据评分标准对 10 份作文进行试评，评分完毕后，计算三位评分员评分的肯德尔和谐系数（Kendall coefficient of concordance），当系数低于 0.8 时，三名评分员对差异较大的作文评分进行讨论并重新进行评分，直到相关系数达到 0.8 或以上。第二轮，再对另外 10 份作文进行试评，我们对评分者信度进行计算，相关系数达到 0.87，达到较高的一致性。第三轮，对 210 篇韩国学生作文进行整体性打分，当两名教师的评分相差 3 分或 3 分以上时由三位老师商定，给出最后的成绩。阅卷完成后，我们对评分者信度进行计算，肯德尔和谐系数为 0.85，该结果表明，评分者一致性较高。最后，每篇作文的最终成绩我们取三位评分员评分的平均分。

内容质量评估：对于写作内容质量的评估，我们借鉴 Kuiken & Vedder（2017）的功能充分性评分量表，并对其进行简化（见附录 2），分成 6 个等级，满分 6 分，最低分 1 分，具体考察的维度包括：第一，是否很好地完成了任务要求，即写作内容是否紧扣题目、没有跑题，举例和阐释是否细致恰当；第二，文本是否易于理解；第三，语句之间和段落之间的衔接是否自然。内容质量的评估是在写作质量评分完成一周之后进行的，评分员仍是写作质量评分的三位老师，评分程序也与写作质量评分程序相同——两轮试评和一轮正式评分，唯一不同之处是当评分员之间的打分相差 2 分或 2 分以上时，共同商定给出最后的分数。阅卷完成后，我们对内容质量的评分者信度进行计算，肯德尔和谐系数为 0.87，该结果表明，评分者一致性较高。最后，每篇作文的内容质量打分我们取三位评分员的平均分。

评分员访谈：在对写作质量和内容质量评分完毕以后，我们对三位评分员进行采访，问题如下：在对写作质量和内容质量打分时，主要依据哪些文本特征？评分标准是否存在差别？原因是什么？

语料的处理：打分完成后，将学生作文逐一转写为电子文本，建立本研究所需要的语料库。然后利用国家语委的相关软件对文本逐一进行分词、词性标注以及字词频率统计。

2.4 本文所用测量指标及工具

（1）词汇多样性测量工具。词汇多样性是指在写作中使用多种不同的词如同义词、上位词和其他关系的词，而避免重复使用某些词（Read，2000：200）。本文采用 Uber index 来测量词汇多样性，具体计算方法如下：

$$\text{Uber index: } U=\frac{(\log \text{Tokens})^2}{(\log \text{Tokens}-\log \text{Types})} \qquad \text{(Jarvis, 2002)}$$

（2）词汇复杂性测量工具。词汇复杂性是指在写作中选择使用更适合主题的低频词，而不是选择日常的普通词汇，使用技术名词、术语和其他各种有特点的词来精确地表达作者的意思（Read，2000：200）。本文词汇复杂性的操作定义是每篇作文中使用《汉语水平词汇与汉字等级大纲》（国家汉语水平考试委员会办公室考试中心，2001。以下简称《大纲》）中的乙级词、丙级词、丁级词及《大纲》未收录词的词汇种类（type）总数占每篇作文词汇种类总数的

比例。

(3)词汇正确性测量工具。在计算词汇正确性时,我们主要统计的是语义错误和搭配错误,其中语义错误包括易混淆词中的理性意义基本相同的词、有相同语素的词、母语一词多义对应的汉语词、单双音节近义词等。搭配错误我们主要考察的是动宾、动补等结构语义上是否搭配,但如果出现词汇错序,如"房间整理",我们则把它作为语法错误。另外,词形错误我们放在汉字层面进行统计。考虑到文本长度对计算结果的影响,我们计算每个学习者的词汇错误率,即每个学习者词汇错误的总数除以每篇作文的总词数。最后用1减去词汇错误率,即为词汇正确率。

(4)汉字正确性测量工具。我们将汉字错误分成两类:一类是字形错误,即错别字;第二类是用拼音代替汉字。首先,我们利用国家语委的字词统计软件计算每篇作文的总字数(不包含标点在内),然后在每篇作文中,用两类汉字错误的总个数除以总字数,即为汉字错误率。最后用1减去错误率即为汉字正确率。

(5)语法正确性测量工具。我们借鉴井茁(2013)的方法计算语法正确性,即每篇作文中无语法错误小句总数与小句总数的比值。其中,小句包括简单整句和复句中的子句。

(6)句法复杂性测量工具。吴继峰(2018b)发现T单位长度指标既可以有效区分韩国学生的汉语水平,也能有效预测其汉语写作成绩,故本研究中句法复杂性采用T单位长度作为测量指标。T单位本文采用Jiang(2013)的定义,即"汉语的T单位是指包含一个独立谓语和其他附属小句或嵌入小句的独立主句"。具体来说,可包括以下几条标准:(1)简单句作为一个T单位。(2)有两个及两个以上分句组成的复合句(compound sentence),根据每个分句中是否含有谓语划分为不同的T单位。分句中含有谓语才能算作一个T单位。(3)复杂句(complex sentence)中的嵌入分句不作为独立的T单位。

2.5 语言水平与6个语言测量指标的交互效应

因为本文研究目的之三是考察语言水平与6个语言测量指标对写作成绩的预测是否存在交互作用,所以在进行回归分析之前,应先计算出6个交互效应。以词汇多样性为例,语言水平和词汇多样性的交互效应=(原始卷面分-原始卷面分的总体平均分)*(词汇多样性U值的原始数据-U值的总体平均数)。

三 结果

3.1 描述统计

写作成绩、内容质量评分和6项语言测量指标的描述性数据见表1。

表1 写作成绩、内容质量评分和各项测量指标的描述性数据表

	平均数	标准差		平均数	标准差
写作成绩	6.31	1.27	汉字正确性	0.97	0.03
内容质量评分	4.23	1.06	词汇正确性	0.96	0.03
词汇多样性	16.44	5.15	语法正确性	0.88	0.12
词汇复杂性	0.19	0.07	句法复杂性	10.36	1.69

3.2 相关及回归

为考察210名韩语母语者6个语言区别性特征、内容质量与写作质量的关系,以及汉语

水平对语言特征预测写作成绩的影响程度，我们使用 SPSS16.0 进行相关及多重回归分析。此回归模型包括 6 个语言区别性特征自变量及语言水平与 6 个语言测量指标的交互效应，因变量为写作成绩。各变量之间的相关关系如表 2 所示。

表 2　各变量之间的相关矩阵(N=210)

	写作成绩	内容质量	词汇多样性	词汇复杂性	汉字正确性	词汇正确性	语法正确性	句法复杂性
写作成绩	——							
内容质量	0.874***	——						
词汇多样性	0.404***	0.349***	——					
词汇复杂性	0.431***	0.432***	0.340***	——				
汉字正确性	0.444***	0.350***	0.184**	0.101	——			
词汇正确性	0.403***	0.386***	0.154*	0.140*	−0.012	——		
语法正确性	0.439***	0.418***	0.178**	0.174**	0.001	0.208**	——	
句法复杂性	0.415***	0.424***	0.262***	0.409***	0.180**	0.217**	0.138*	——

注：* $p<0.05$，** $p<0.01$，*** $p<0.001$。

由表 2 可知，6 个语言区别性特征均与写作成绩显著相关，r 值范围处于 0.403～0.444 之间(相应的 p 值均小于 0.001)。该研究的被试量为 210 人，达到被试数量与预测指标的比率不低于 5∶1 的标准(Plonsky & Ghanbar，2018)。同时，我们结合数据并利用直方图、P−P 图、散点图进行正态性检验、线性检验、方差齐性检验(Tabachnick & Fidell，2013：161)，结果显示残差呈正态分布，与预测值之间符合直线关系，而且方差齐性，满足进行多重线性回归分析的前提条件。

为考察每个自变量对写作成绩的贡献程度，我们进行逐步回归分析(stepwise regression analysis)，结果见表 3。

表 3　写作成绩的多重逐步回归结果摘要表(N=210)

Entry	Variable added	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change
1	汉字正确性	0.444	0.198	0.194	1.144	0.198
2	语法正确性	0.625	0.390	0.384	1.000	0.192
3	词汇正确性	0.703	0.495	0.488	0.912	0.105
4	词汇复杂性	0.758	0.575	0.567	0.839	0.080
5	词汇多样性	0.770	0.593	0.583	0.823	0.018
6	语言水平 * 词汇多样性	0.790	0.625	0.613	0.792	0.032
7	语言水平 * 汉字正确性	0.800	0.640	0.627	0.778	0.015
8	句法复杂性	0.806	0.649	0.635	0.769	0.009

对于成功的回归模型来说，估计标准误(Std. Error of the Estimate)应大大小于因变量的标准差(Martin & Bridgmon，2012，转引自秦晓晴、毕劲，2015：459)，本例中 8 个模型的估计标准误均小于写作成绩的标准差 1.27，所以该回归模型是成功的。

通过表 3 可知，在逐步回归分析中，汉字正确性、语法正确性、词汇正确性、词汇复杂性、

词汇多样性、句法复杂性以及显著的交互效应能解释写作成绩总变异的64.9%。具体来看，各测量指标解释写作成绩总变异的比例分别为：汉字正确性19.8%，语法正确性19.2%，词汇正确性10.5%，词汇复杂性8%，词汇多样性1.8%，句法复杂性0.9%。按照Cohen(1988:413－414)的效应量参照体系，R^2的小、中、大效应量标准分别是0.02、0.13、0.26，那么汉字正确性和语法正确性达到中等效应量。如果按照语言维度来统计的话，词汇三个测量指标能解释写作成绩总变异的20.3%，语法两个测量指标能解释20.1%，汉字一个测量指标能解释19.8%。语言水平与汉字正确性的交互效应、语言水平与词汇多样性的交互效应显著，这说明在不同语言水平上汉字正确性、词汇多样性对写作成绩的影响有显著差异。为进一步考察交互效应，我们对三个水平分别进行回归分析，结果发现，汉字正确性在初、中、高水平上的贡献率分别是16.4%、7.9%、7.4%，初级水平汉字正确性的贡献率显著高于中、高级水平；但是未发现词汇多样性在三个水平上的贡献。另外，各自变量在进入回归模型之后，容忍度(tolerance)均大于0.5，方差膨胀因子(VIF)均小于2。以上数据说明该多重回归分析中不存在多重共线性。

另外，为考察各自变量对内容质量评分的贡献程度，我们也进行一次逐步回归分析，结果见表4。

表4　内容质量评分的多重逐步回归结果摘要表(N＝210)

Entry	Variable added	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change
1	词汇复杂性	0.432	0.186	0.182	0.961	0.186
2	语法正确性	0.554	0.307	0.301	0.889	0.121
3	汉字正确性	0.637	0.406	0.397	0.825	0.099
4	词汇正确性	0.694	0.482	0.472	0.772	0.076
5	句法复杂性	0.711	0.506	0.494	0.756	0.024
6	语言水平＊句法复杂性	0.722	0.521	0.507	0.746	0.015

首先，通过表2可知，6个语言特征均与写作质量有显著的相关关系(r值范围在0.349～0.432，p值均小于0.001)，而且被试量、正态性检验、线性检验、方差齐性检验的结果均满足进行多重回归分析的前提条件，所以可以进行多重回归分析。

其次，通过表4可知，本例中6个模型的估计标准误均小于内容质量评分的标准差1.06，说明该回归模型是成功的。在此次逐步回归分析中，词汇复杂性、语法正确性、汉字正确性、词汇正确性、句法复杂性以及显著的交互效应能解释内容质量评分总变异的52.1%。具体来看，各测量指标解释内容质量评分总变异的比例分别为：词汇复杂性18.6%，语法正确性12.1%，汉字正确性9.9%，词汇正确性7.6%，句法复杂性2.4%。按照Cohen(1988:413)的标准，词汇复杂性达到中等效应量。如果按照语言维度来统计的话，词汇两个测量指标能解释内容质量评分总变异的26.2%，语法两个测量指标能解释14.5%，汉字一个测量指标能解释9.9%。语言水平与句法复杂性的交互效应显著，这说明在不同语言水平上句法复杂性对写作成绩的影响有显著差异。为进一步考察交互效应，我们对三个水平分别进行回归分析，但结果未发现句法复杂性在三个水平上的贡献。另外，每个自变量在进入回归模型之后，容忍度均大于0.5，方差膨胀因子均小于2。以上数据说明该多重回归分析中不存在

多重共线性。

四　讨论与分析

4.1 语言区别性特征与写作质量的关系

本文首先考察了词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言区别性特征与写作成绩的关系，并检验了这6个特征对写作成绩的预测作用。研究发现，从总体上看，这6个语言特征均与写作质量有显著的相关关系。另外，6个语言特征均能显著预测写作成绩，这些语言特征的组合以及显著的交互效应能解释写作成绩总变异的64.9%。其中，汉字正确性、语法正确性、词汇正确性的贡献最大，三者构成的正确性能解释总变异的49.5%；其次是词汇复杂性；最后是词汇多样性和句法复杂性。以上数据说明，语言正确性是评分员最看重的评分维度，书写是否正确、是否正确使用词汇和语法直接影响最后的得分，这与之前的部分研究结果是一致的。吴继峰(2018a)基于中级汉语水平的英语母语者写作的数据发现，语法正确性和词汇正确性二者可以解释写作成绩总变异的51.5%，是贡献力最大的维度，但是该研究是将汉字和词汇作为一个维度考察的，没有单独考察汉字正确性的贡献。而本研究发现从总体上看，作为单个维度，汉字正确性的贡献最大，能解释总变异的19.8%，且达到中等效应量。这符合人们的一般认知，因为汉字使用正确与否直接影响到评分员的识读和理解，汉字错误过多或者直接用拼音代替，会妨碍评分员的阅读、增大阅读难度、降低阅读速度，尤其是初级水平学生的作文。另外，语法正确性和词汇正确性的贡献也很大，其中，语法正确性也达到中等效应量，语法和词汇是否使用正确和地道也会直接影响评分员的阅读和理解，吴继峰(2018a)发现在语言特征维度，语法正确性是影响中级汉语水平英语母语者写作成绩的最大因素，达到大效应量；Crossley et al.(2015)考察词汇内部各维度对英语二语写作成绩的预测作用时，发现词汇搭配正确性对写作成绩的预测作用最大。可见，虽然上述研究中的被试群体母语背景、汉语水平、目的语不同，但语法和词汇正确性在写作质量评估中都具有重要作用。

除了语言正确性以外，词汇复杂性也是重要的预测指标，能解释写作成绩总变异的8%。低频词的使用情况能帮助评分者区分汉语水平。下面结合学生的作文具体说明：

例1：高级水平学生作文

每个人都有自己的课余爱好，我也不是例外。你知道我的爱好是什么？告诉你吧，就是看书。

我每天都要看书，可想而知我是一个见到书便爱不释手的小书迷。好像看书已经成为我生命当中的一部分，可能是因为我所有家人都喜欢看书。

我喜欢看书也是因为看书让我在知识的海洋里遨游，让我增长了许多知识，更让我知道了生活的五彩斑斓。

该作文虽有几个语法错误(首段第二和第三小句)，但是正确使用了几个低频词，用得恰如其分，直接提升了作文质量，例如“爱不释手”“遨游”“五彩斑斓”。此外，该文还正确使用了地道的修辞手法，如“看书让我在知识的海洋里遨游”。三位评分员一致认为该文的写作质量较高，给出8或8.5的高分。可见，恰当使用低频词，并配合使用汉语修辞手法，能大大提升作文质量。Crossley & McNamara(2012)基于英语二语写作的研究也发现低频词是有

效区分二语写作质量的标准之一。但本文与王艺璇(2017)的研究结果不同,王文发现词汇复杂性不是预测写作质量的有效参项。造成两项研究结果不同的原因可能是,两项研究中词汇复杂性的操作定义不同,王艺璇(2017)词汇复杂性的范围划定为《汉语水平词汇与汉字等级大纲》丙级、丁级词及超纲词,而本文词汇复杂性的划定范围是乙、丙、丁级词及《大纲》未收录词,比王文多了乙级词。之所以把乙级词也放到词汇复杂性的范围内,是因为前人研究(孙晓明,2009)发现留学生即使到了高级阶段,仍倾向于使用甲级词,使用乙级词的数量并不多。吴继峰(2016)也证明在汉语二语写作研究中,把乙级词放到词汇复杂性的范围内是可行的,因为其不仅可以区分中、高级汉语水平,而且也能有效预测写作质量。

最后,词汇多样性和句法复杂性对写作成绩也有显著贡献,分别能解释总变异的1.8%、0.9%,这与之前的相关研究结果是不一致的。吴继峰(2016、2018a)基于英语母语者汉语写作的两项研究发现,虽然词汇多样性与写作成绩显著相关,但是并未能有效预测写作成绩,而本文发现词汇多样性可以有效预测写作成绩。可能是因为词汇多样性测量指标易受语言样本大小的影响,本文样本量大,共210个,远远高于前两项研究的46个和50个,词汇多样性指标在计算预测作用时不易显著,在210人的大样本中只能解释写作成绩总变异的1.8%也能说明这一点。另外,句法复杂性对写作成绩的预测作用也不大,仅为0.9%。访谈发现,三位评分员在打分时往往是将句法复杂性和语法正确性结合起来对作文质量进行判断。Banerjee et al. (2015) 基于英语二语的写作研究发现,采用T单位长度作为测量指标未能显著预测写作成绩,认为可能的原因之一是评分员在执行语法标准时,往往将句法复杂性和语法正确性合并在一起,因为语法正确性能更好地帮助他们区分写作水平。本文采用T单位长度作为句法复杂性的测量指标能够有效预测写作成绩,可能是因为初级汉语水平的韩国学生作文中T单位长度和平均句长都很短,而且大都是简单的单句,随着汉语水平的提高,T单位长度和平均句长明显变长,评分员也容易观察到。我们的结果与施家炜(2002)关于韩国学生个案研究的结果是一致的,即韩国学生汉语平均句长随着水平提高而增长。

4.2 内容质量与写作质量、语言区别性特征的关系

内容质量指的是学习者是否达到成功交际。在《我的爱好》这篇作文中,内容质量具体指作文是否紧扣题目进行阐述、内容是否充实、举例是否恰当、全文是否易于理解、语句之间和段落之间的衔接连贯是否顺畅。从表2可知,内容质量和写作质量的评分达到高相关(r=0.874),这说明作文质量的高低在很大程度上取决于内容质量的好坏。Révész et al. (2016)发现,英语二语学习者口语产出的内容质量和整体口语评分的相关系数达到0.66。可见,内容质量对写作和口语评分的影响很大。

本文之所以没有将内容质量和6个语言特征一起放入回归模型考察这7个指标对写作成绩的预测作用,是因为内容质量和6个语言测量指标之间有交叠,会影响回归分析的结果。例如,内容质量的评估中包含一个可理解性维度,可理解性是指全文是否易于理解,而影响理解的因素包括汉字、词汇、语法、逻辑等因素,这与汉字、词汇、语法维度的测量在很大程度上是重合的。

我们也考察了词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言区别性特征与内容质量评分的关系。研究发现,这6个语言特征均与内容质量评分有显著的相关关系。另外,除了词汇多样性以外,词汇复杂性、词汇正确性、汉字正确

性、语法正确性、句法复杂性5个语言特征均能显著预测内容质量评分，这些语言特征的组合以及显著的交互效应能解释内容质量分数总变异的52.1%。如果按照单个指标来看，词汇复杂性的贡献最大，达到中等效应量；其次是语法正确性、汉字正确性、词汇正确性；最后是句法复杂性。以上数据说明，影响评分员对内容质量打分的因素依次是语言正确性、词汇复杂性和句法复杂性。可见，语言正确性也是影响内容质量打分最重要的因素，汉字、词汇、语法使用是否正确恰当直接影响评分员的理解。但是与写作质量回归分析的结果相比，在内容质量的回归分析中，词汇复杂性解释总变异的比例明显增大（从解释写作成绩总变异的8%到解释内容质量成绩总变异的18.6%），解释力更强，可能是因为作文成绩和内容质量评分的标准有些许不同，内容质量更看重作者是否围绕题目进行详细举例说明，如学生能恰当使用更多的低频词，围绕题目进行细致阐述，就能大大提升内容质量，如例1的学生作文。另外，句法复杂性的解释力也有所增强（从解释写作成绩总变异的0.9%到解释内容质量总变异的2.4%），但增加幅度不大。

另外，针对评分员对写作质量和内容质量的打分差异，我们就其采用的主要评分依据进行访谈。三位评分员表示，在给写作质量和内容质量打分时是有少许差异的，在给写作质量评分时，他们表示在较短时间内评阅大量试卷，他们更看重学习者语言的正确性，即汉字、词汇和语法有无明显错误，也会看重低频词的使用，高级词汇的使用会让他们眼前一亮；另外，他们会注意学生行文的逻辑性，句子之间和段落之间是否在语义上衔接自然。在给内容质量评分时，他们更看重内容是否切题、举例是否恰当，如学生使用低频词同时配合使用排比、比喻等修辞手法，将句与句之间自然衔接，恰当阐释主题，会更加吸引评分者；同时，他们也会注意汉字、词汇和语法的使用是否会造成理解困难。可见，评分员在给作文质量评分时最为重视的是语言正确性，而在给内容质量评分时更为重视的是高级词汇的使用以及语言正确性。质性访谈的结果和我们量化分析的结果是一致的。

4.3 语言水平对各语言测量指标预测写作成绩的影响

本文的第三个研究问题聚焦在汉语水平是否影响语言特征测量和写作成绩的关系，也就是说，语言特征测量和写作成绩关系的改变是否依赖于汉语水平。本研究发现，语言水平和汉字正确性、词汇多样性都存在交互效应，这说明在不同语言水平上汉字正确性、词汇多样性对写作成绩的影响有显著差异。对三个水平分别进行回归分析时发现，在汉字正确性方面，初级水平汉字正确性的贡献率显著高于中、高级水平。这可以从两个角度进行分析，从语言发展的角度看，在语言发展过程中各水平阶段有各自显著的特征，在初级阶段，正确书写汉字的能力还在发展中，所以这个阶段正确书写汉字的能力可以区分不同能力的学生。从语言测试的角度看，在评分时，评分员实际关注的层面或对不同语言水平学生作文的评分标准有所区别，这也得到很多实证研究的证实（Humphry & Heldsinger，2014）。在本研究中，由于初级阶段汉字错误较为突出，评分员更能根据汉字错误情况区分作文质量。

对三个水平分别进行回归分析时，三个水平上都没有发现词汇多样性的显著贡献，但是对总体（210人）进行回归分析时词汇多样性有显著贡献。这可能是因为词汇多样性这个指标的统计检验力不够大，在样本数量不够大时，不易显著。我们的总体回归中样本量是210个，结果显著；三个水平分开进行回归分析时，每个水平的样本量是70个，结果都不显著。

五 结论与启示

本文基于不同汉语水平的210篇韩国学生分班考试的作文，考察了词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言区别性特征与写作成绩、内容质量评分的关系及其预测作用，还考察了写作成绩和内容质量评分之间的关系，以及语言水平对6个语言测量指标和写作成绩关系的影响。研究发现，词汇多样性、词汇复杂性、词汇正确性、汉字正确性、语法正确性、句法复杂性6个语言区别性特征均与写作成绩和内容质量评分显著相关，均能显著预测写作成绩，其中汉字正确性、语法正确性、词汇正确性贡献程度最大，其次依次是词汇复杂性、词汇多样性和句法复杂性；除了词汇多样性以外，其余5个语言特征均能显著预测内容质量评分，其中，词汇复杂性的贡献程度最大，其次依次是语法正确性、汉字正确性、词汇正确性和句法复杂性。如果将语法正确性、汉字正确性、词汇正确性看作语言正确性一个维度，语言正确性在写作成绩和内容质量评分中的预测作用都是最大的。另外，内容质量评分和写作成绩达到高相关，说明写作成绩的高低很大程度上取决于内容质量的好坏，即取决于内容是否切题、举例阐释是否细致恰当、语言是否可理解、句子之间和段落之间是否衔接连贯自然。最后，研究还发现语言水平与汉字正确性存在交互效应，与中高级水平相比，初级水平汉字正确性对写作成绩的预测作用更大。

以上研究结果给予我们的启示是，在语言特征方面，作文评分主要依靠的是汉字正确性、语法正确性、词汇正确性和词汇复杂性，语言使用的正确性是重中之重。所以，为提升学生的汉语写作水平，必须打好语言基本功，确保语言的正确使用。同时，要探索行之有效的方法扩大学生乙级词和低频词的产出量。另外，写作成绩的高低很大程度上取决于内容质量的好坏，所以教师应该训练学生解题的能力，使其写作内容切题、举例阐释细致恰当；同时，也要训练学生的逻辑思维方式，教给学生句子之间和段落之间常用的衔接和连贯方式以及比喻、排比等修辞手法，并通过相关的训练使其逐渐掌握，使其连词成句、连句成篇。

本研究虽然较为全面地考察了语言特征和内容质量对写作质量评估的影响，但是在学习者方面并未考察，如写作过程中的认知因素和修改过程对于作文质量评估的影响。另外，本文所考察的是任务比较简单的记叙文，对于任务更难的议论文和说明文结果如何，值得进一步考察。本文所考察的对象是汉字圈的韩国学生，对于非汉字圈的学习者结果是否一致，也值得进行对比研究。

参考文献

国家汉语水平考试委员会办公室考试中心（2001）《汉语水平词汇与汉字等级大纲》，北京：经济科学出版社。

金 檀、刘 力、郭 凯（2016）口语测试评分标准研究与实践三十年，《现代外语》第6期。

井 茁（2013）从中介语发展分析到高级汉语课程设置——内容依托型教学研究的启示，《世界汉语教学》第1期。

秦晓晴、毕 劲（2015）《外语教学定量研究方法及数据分析》，北京：外语教学与研究出版社。

施家炜（2002）韩国留学生汉语句式习得的个案研究，《世界汉语教学》第4期。

孙晓明（2009）留学生产出性词汇的发展模式研究，《民族教育研究》第4期。

王佶旻（2002）三类口语考试题型的评分研究，《世界汉语教学》第4期。

王艺璇（2017）汉语二语者词汇丰富性与写作成绩的相关性——兼论测量写作质量的多元线性回归模型及方程，《语言文字应用》第 2 期。

吴继峰（2016）英语母语者汉语写作中的词汇丰富性发展研究，《世界汉语教学》第 1 期。

吴继峰（2018a）语言区别性特征对英语母语者汉语二语写作质量评估的影响，《语言教学与研究》第 2 期。

吴继峰（2018b）韩语母语者汉语书面语句法复杂性测量指标及与写作质量关系研究，《语言科学》第 5 期。

辛　平（2007）基于语言能力构想的作文评分标准及其可操作性研究，《暨南大学华文学院学报》第 3 期。

朱世芳（2009）韩国汉语第二语言学习者口语语篇发展研究，北京语言大学硕士学位论文。

Banerjee, Jayanti, Xun Yan, Mark Chapman & Heather Elliott (2015) Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing* 26(4): 5－19.

Cohen, Jacob (1988) *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Crossley, Scott A. & Danielle S. McNamara (2012) Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2):115－135.

Crossley, Scott A.,Tom Salsbury & Danielle S. Mcnamara (2015) Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics* 36(5): 570－590.

De Jong, Nivja H., Margarita P. Steinel, Arjen F. Florijn, Rob Schoonen & Jan H. Hulstijn (2012) Facets of speaking proficiency. *Studies in Second Language Acquisition* 34(1): 5－34.

Humphry, Stephen Mark & Sandra Allison Heldsinger (2014) Common structural design features of rubrics may represent a threat to validity. *Educational Researcher* 43(5): 253－263.

Jarvis, Scott (2002) Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* 19(1): 57－84.

Jiang, Wenying (2013) Measurements of development in L2 written production: The case of L2 Chinese. *Applied Linguistic* 34: 1－24.

Jin, Tan & Barley Mak (2013) Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing* 30(1): 23－47.

Kuiken, Folkert & Ineke Vedder (2014) Rating written performance: What do raters do and why? *Language Testing* 31(3): 329－348.

Kuiken, Folkert & Ineke Vedder (2017) Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing* 34(3): 321－336.

Kuiken, Folkert, Ineke Vedder & Roger Gilabert (2010) Communicative adequacy and linguistic complexity in L2 writing. In Inge Bartning, Maisa Martin and Ineke Vedder (eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 81－100. Eurosla Monographs Series 1. European Second Language Association.

Martin, William E. & Krista D. Bridgmon (2012) *Quantitative and statistical research methods: From hypothesis to results*. San Francisco: Jossey-Bass.

Pallotti, Gabriele (2009) CAF: Defining, refining and differentiating constructs. *Applied Linguistics* 30(4): 590－601.

Plakans, Lia, Atta Gebril & Zeynep Bilki (2016) Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, https://doi.org/10.1177/0265532216669537. (Pre-published 26 Sep 2016)

Plonsky, Luke & Hessameddin Ghanbar (2018) Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal* 102(4): 713－731.

Read, John (2000) *Assessing vocabulary*. Cambridge: Cambridge University Press.

Révész, Andrea, Monika Ekiert & Eivind Nessa Torgersen (2016) The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics* 37(6): 828–848.

Tabachnick, Barbara G. & Linda S. Fidell (2013) *Using multivariate statistics*, 6th edn. Boston: Pearson Education.

附录1:写作质量评分量表

该写作质量的评分标准主要改编自新HSK的评分标准,主要涉及总体内容和结构、语法、汉字和词汇三大块(国家汉办《汉语考试阅卷员培训手册》:19–20),改编时对评分标准进行相应的分数段设计,如下:

等级、分数段	总体内容和结构	语法	汉字和词汇
5档 9分	短文内容较充实,结构完整,内容逻辑性强,条理清楚,衔接顺畅,表达得体。总体上表达自然,能顺畅阅读完全文。	1. 语句通顺,表达形式丰富,能使用复杂的句式结构。 2. 语法准确,无简单语法错误,特别是虚词的使用和词序。	1. 词汇使用丰富、准确,恰当得体,有个别错误,但不影响交际。 2. 允许有一两个错别字。
4档 7–8分	短文内容比较充实,结构较完整,有一定逻辑性,条理清楚,短文衔接比较顺畅连贯,存在个别语句连接上的不恰当。总体上能顺利阅读完全文,理解作者所要表达的主要内容,有一些语法词汇错误影响意思的表达和理解。	1. 语句通顺,表达形式不单一,在使用复杂的句式结构时有错误出现,但不影响交际。 2. 有个别的语法错误,主要是词序和虚词的错误,但不影响交际表达的准确性。	1. 词汇使用比较丰富、准确、得体,有个别错误,但不影响交际。 2. 有个别错别字。
3档 5–6分	短文内容结构基本完整,有一定逻辑性和条理性,前后基本连贯。总体上能理解作者所要表达的内容,有语法词汇方面的错误影响部分意思的表达和理解,但能读懂。	1. 语句基本通顺,表达形式简单。 2. 有一定语法错误,主要表现在虚词、词序使用方面。	1. 有一定的词汇量,能使用简单常用词语满足任务表达的需要,有些词语使用不恰当。 2. 有少量错别字。
2档 3–4分	内容简单,条理性和连贯性较差,有连接上的错误和不恰当。总体上基本能理解作者要表达的思想内容,有需要通过猜测理解的部分。	1. 句式简单,句子欠通顺。 2. 语法错误较多,影响意义理解和表达,有些语句影响交际,使用简单的连接词,但有错误或不恰当。	1. 能使用有限的词汇,词不达意,影响语句意义的表达。 2. 有一些错别字。
1档 1–2分	短文基本上看不出有条理性,连贯性差,总体上难以读懂。	语法错误太多,导致无法交际。	1. 词汇使用错误多,使用不恰当。 2. 错别字很多。

附录 2：内容质量评分量表

该内容质量的评分标准主要改编自 Kuiken & Vedder(2017：335－336)的功能充分性(functional adequacy)评分量表，主要包括内容、任务要求、可理解性、衔接和连贯四部分，改编时对评分标准进行相应的分数设计，如下：

分数	内容	任务要求	可理解性	衔接和连贯
6	观点的数量非常充足，而且彼此之间非常一致。	回答了所有的问题和任务要求。	文本非常容易懂，可读性很高。观点和目的都说明得很清楚。	文本非常连贯。通过使用连接词或者连接短语整合文本中新的观点。有规律地使用前后照应手段。很少有不相关的推进，没有连贯的中断。文本结构衔接非常自然，能熟练使用连接词，经常使用它们来描述观点之间的关系。
5	观点的数量很充足，而且彼此之间很一致。	回答了绝大部分问题和任务要求。	文本很容易理解，阅读起来很顺畅，可理解性不是一个问题。	文本很连贯。当作者介绍一个新话题时，一般使用连接词或者连接短语。不经常使用重复。有很多前后连接手段。没有连贯的中断。文本衔接自然，能很好地连接观点。
4	观点的数量充足，而且足够一致。	回答了大部分(超过一半)的问题和任务要求。	文本是可以理解的。只有几个句子不清楚，但是可以懂，在阅读第二次之后，不用付出太多努力就可理解。	文本是连贯的。不相关的推进少，但是写作者有时依赖于重复来达到连贯。使用足够数量的前后照应，但有一些连贯的中断。文本衔接较自然。能较好地使用连接词，有时不仅仅局限于连词。
3	观点的数量比较充足，虽然它们不是非常一致。	回答了大概一半的问题和任务要求。	文本稍微可以理解。有的句子第一次读很难读懂，第二次读可以帮助澄清文本的目的和传达的观点，但是还留有疑问。	文本比较连贯。经常出现不相关的推进和重复。使用一些前后照应的手段。有几个连贯的中断。使用一些连接词，但是它们大部分都是连词。
2	观点的数量几乎不充足，观点缺乏一致性。	回答了一些(不到一半)的问题和任务要求。	文本几乎不能理解。目的描述得不清楚，读者很难理解写作者的观点。读者不得不猜测大部分观点和目的。	文本几乎是不连贯的。写作者经常使用不相关的推进方式，经常使用重复的方法来实现连贯。只是用几个前后照应的技巧。有一些连贯的中断。文本不是非常衔接。观点没有被连接词连接好，也很少用连接词。
1	观点的数量一点儿也不充足，观点之间没有关系。	没有回答任何问题和任务要求。	文本完全不能理解。观点和目的说明得不清楚，读者付出的努力来理解文本是无效的。	文本完全不连贯。没有关系的推进，连贯的中断非常常见。作者没有使用任何前后照应的手段。文本完全不衔接。几乎不使用连接词，观点之间没有关系。

Assessing Chinese L2 Writing Quality on Basis of Language Features and Content Quality

Wu Jifeng　Zhou Wei　Lu Dawei

Abstract This research involves 210 South Korean speakers who are of different Chinese levels. A correlation and a linear regression analysis are conducted to explore the relationship between six language features (lexical diversity, lexical sophistication, lexical accuracy, Chinese character accuracy, grammar accuracy, and syntactic sophistication), writing score and content score. Then the relationship between content score and writing score is explored, which is in turn followed by an analysis of the effect of Chinese level on the relationship between language features and writing score. The results show that all six language features are significantly correlated with writing score and content score and can predict writing score effectively. Of the six features, Chinese character accuracy, grammar accuracy and lexical accuracy contribute the most, with the former two achieving medium effect size. Except lexical diversity, the other five language features can all predict content score, lexical sophistication contributing the most to achieve medium effect size. We also find high correlation between content score and writing score, as well as interaction effect between language level and character accuracy, in that character accuracy best predicts writing score on the basic level. In the mean time, by combining students' writing performance, interview with scorers, and an examination of the internal construct of measure indices, we make a quality analysis of the reasons for the different contributions of language features to writing score and content score, and the relationship between writing quality and content quality.

Keywords Chinese as a second language, writing quality, content quality, distinguishing feature, assessment

作者简介

吴继峰(通讯作者),男,博士,首都师范大学国际文化学院副教授,主要研究领域为汉语二语教学、习得与测试。[Email: ubaid@163.com]

周蔚(通讯作者),男,博士,首都师范大学心理学院副教授,主要研究领域为阅读认知过程。[Email: zhouwei@cnu.edu.cn]

卢达威,男,博士,中国人民大学文学院讲师,主要研究领域为中文信息处理。[Email: wedalu@163.com]