Modern Linguistics 现代语言学, 2024, 12(1), 423-428
Published Online January 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.121061

Hans汉斯

基于词汇丰富性的汉语二语作文质量研究

彭盼盼

鲁东大学文学院，山东 烟台

收稿日期：2023年12月12日；录用日期：2024年1月16日；发布日期：2024年1月25日

摘 要

本文通过分析HSK动态作文语料库的近15万字作文语料，以词汇丰富性为基点，从词汇复杂度、词汇密度、词汇多样性以及词汇等级四个角度出发，通过SPSS统计分析，发现词汇复杂度与作文质量呈现显著性相关，词汇密度与作文质量无显著性相关关系，词汇多样性则因其计算方式的不同而呈现出较为复杂的态势，词汇等级也对作文质量具有相当的影响。

关键词

词汇丰富性，作文质量，汉语二语作文

Study on the Quality of Chinese Second Language Composition Based on Vocabulary Richness

Panpan Peng

College of Liberal Arts, Ludong University, Yantai Shandong

Received: Dec. 12th, 2023; accepted: Jan. 16th, 2024; published: Jan. 25th, 2024

Abstract

In this paper, through the analysis of HSK dynamic composition corpus nearly 150000 words composition corpus, based on vocabulary richness, from the vocabulary complexity, vocabulary density, vocabulary diversity and four levels, through the SPSS statistical analysis, it is found that there is a significant correlation between vocabulary complexity and composition quality, but no significant correlation between vocabulary density and composition quality. Vocabulary diversity shows a more complex trend due to its different calculation methods, and vocabulary

文章引用: 彭盼盼. 基于词汇丰富性的汉语二语作文质量研究[J]. 现代语言学, 2024, 12(1): 423-428.
DOI: 10.12677/ml.2024.121061

level also has considerable influence on composition quality.

Keywords

Vocabulary Richness, Composition Quality, Chinese Second Language Composition

1. 引言

随着我国经济的快速发展和综合国力的不断提升，汉语的快速传播以及对外汉语教学事业的蓬勃发展促进了汉语作为第二语言习得研究的发展，许多国内研究者在引入和吸收国外先进成果的基础上，积极探索适用于汉语作为二语的习得理论体系。而如何有效衡量二语学习者的语言水平和发展过程一直是国内外研究者们所关注的关键问题。从语言习得的语言要素角度出发，词汇作为语言的建筑材料，在二语习得过程中起着至关重要的作用。

词汇丰富性指“文本中词汇知识的质量”(Nation & Webb, 2011)。Laufer & Nation (1995)提出，词汇丰富性的测量维度包括词汇变化性(Lexical variation)、词汇密度(Lexical density)、词汇复杂性(Lexical sophistication)和词汇独特性(Lexical orginality)；Engber (1995)发现词汇密度不适于衡量学习者的词汇水平和发展情况，他建议将词汇错误(Lexical error)作为词汇丰富性的内部维度。Read (2000)提出，词汇独特性无法反映学习者的词汇能力发展，认为应该从词汇多样性、词汇复杂性、词汇密度和词汇错误这四个维度来衡量学习者的词汇丰富性，而后大部分学者都采用了 Read 的观点。近年来，复杂度 、准确度、流利度(CAF)已经得到了学界的广泛关注，词汇丰富性是二语学习者语言表达过程中词汇运用能力的体现，也是衡量二语学习者整体语言运用水平的重要尺度。

综上所述，本文主要从二语作文中词汇的丰富性着手，对其词汇多样性、词汇密度、词汇复杂性进行统计分析，探究词汇丰富性与汉语二语作文质量之间的关系，进而为提高汉语二语学习者作文质量针对性地提出建议。

2. 研究过程

2.1. 语料来源

本文所用语料来源于北京语言大学 HSK 动态作文语料库 2.0，该语料库为汉语二语学习者的中介语语料库。本文选用语料从“如何看待安乐死”、“吸烟对个人健康和公众利益的影响”、“如何解决代沟问题”、“我对男女分班的看法”、“我看流行歌曲”、“最理想的结交方式”、“由三个和尚没水喝想到的”、“绿色食品与饥饿”这八个题目入手，每个题目各选取作文若干，并对其进行人工矫正检查。语料共分为三个等级，一级语料得分 95~90，二级语料得分 75~70，三级语料得分 55 以下，三级语料之间分值差异较大且平均，每个级别 5 万字，共计 15 万字，语料在数量和差异性上具有相当的参考价值。

2.2. 参考标准

本文基于词汇丰富性对汉语二语作文质量进行研究，其参考要素为以下三个方面。

词汇密度：指文本中实词与整体词汇的数量比，用于考察文本的信息含量，词汇密度的计算公式为：

词汇密度 = 实词总数/总词数 × 100% [1]。词汇复杂性：指文本中低频词汇和全部词汇的比例，本文以2021 年最新发布的《国际中文教育中文水平等级标准》为参考，该等级标准中将词汇分为“三等九级”[2]。本文研究对各个文本中的词汇进行统计分析，汇总出其各级词汇数量，根据该等级标准我们将初等一到三级词设立为高频词，中等四级到六级词设立为中频词，进而计算出词汇复杂度。具体计算公式有两种：1、(一级词 + 二级词 + 三级词)/词汇总数 × 100%。2、(一级词 + 二级词 + 三级词 + 四级词 + 五级词 + 六级词)/词汇总数 × 100%。

词汇多样性：指学习者使用词汇的范围。该指数计算方式多种，本文采取以下三种词汇多样性测量手段：多样性 1：总词数。多样性 2：总词种数(未重复出现的词汇总数)。多样性 3：Log 2 总词数/Log 总词数-Log 总词种数(优博指数)，国内英语和汉语二语词汇多样性较为常用的词汇多样性测量手段，可简单理解为文本中总词数与总词种数之比的变体。

本研究的参考标准并不唯一，主要体现在词汇多样性的计算方式上，在以往关于词汇丰富性的研究中关于各个标准的定义及计算各有不同，为了确保统计结果的准确性及其多重参考价值，因此选用了多样计算方式。

2.3. 分析统计

2.3.1. 词汇多样性统计与质量分析

Table 1. Statistics of vocabulary diversity of different grades
表 1. 不同等级二语作文词汇多样性情况统计

	高分组	中分组	低分组
平均词汇数量	259	233	143
平均词种数量	159	131	87

根据表 1 可知，多样性 1：文本词汇数量统计，本文所统计的各等级文本共约 15 万字，低中高得分文本字数均等，各约 5 万字左右。对其各文本所包含词数进行统计，并最终相加得出总词数。高分组共 99 篇作文，总词数 25665 个，平均每篇文章约 259 个词；中分组共 115 篇作文，总词数 26852 个，平均每篇文章约 233 个词语；低分组共 191 篇文章，总词数 27469 个，平均每篇文章约 143 个词语。从以上对不同分组作文所含词语数量的统计可得出，词汇数量对作文质量具有相当的影响，高中分组作文均为及格作文，分数均在 70 分以上，两组各篇词汇数量相当，并无明显差异，但第三组低分作文的各篇词汇数量却与上面两组相差巨大，有近 100 词的差异，且低分组均得分 55 分及以下，为不合格作文，由此可知词汇数量是否足够是影响作文质量的重要因素，并且在相当程度上决定着一篇作文合格与否。

多样性 2：词种数在一定程度上是与词汇数量有正相关的关系的，所以一定的词种数是构建合格作文的必要因素。综合以上两点可以得出，高分作文往往具备论据详实、观点论证充分等特征，而这些都需要文本有一定篇幅才可实现的，进而一篇文章的词汇数量及词种数量对其质量具有重要影响。

多样性 3：即优博指数，通过其计算公式我们可以发现，该指数为文本词汇数量与词种数量之比的相关变体，在使用 SPSS 进行分析的过程中我们发现该数据只有高分、中分组符合正态分布，而低分组不符合，因而其与作文质量的相关性关系并不能进行分析[3]，究其原因，一方面，在文本创作过程中[4]；另一方面，由于在文本统计中，HSK 作文语料库收集文本资源有限，本研究所选用话题所需语料篇数在中高分组中采集各话题各篇数大致相同，但低分组由于多数作文字数少，所以在搜集语料过程中一来会产生同一话题下收集相较于中高分组更多篇数的作文，二来各个话题收集语料篇数相差过大的问题。以上问题则会对低分组作文的词汇数词种数产生较大影响，使得该组词汇中产生相较于中高分组的较多重

复，进而导致总词种数和总词数之比没有明显变化。

2.3.2. 词汇复杂度统计与质量分析

本文所探索文本分为三个组别，故而在数据分析中采用 SPSS 的多因素方差分析，首先我们检验出其数据符合正态分布，具有分析价值，针对两组不同计算方法的词汇复杂度数据进行统计分析，结果表 2 所示：

Table 2. Correlation analysis between lexical complexity and composition quality
表 2. 词汇复杂度与作文质量的相关性分析

	平方和	df	平均值平方	F	显著性
群组之间	0.815	2	0.407	88.893	0.000
在群组内	1.843	402	0.005		
总计	2.657	404			

通过表 2 我们可以发现，词汇复杂度 1、2 检验的显著性值，即 p 值为 0.000，小于 0.05，差异具有统计学意义，表明词汇复杂度与文本质量具有显著相关性关系。

2.3.3. 词汇密度统计与质量分析

词汇密度的数据在正态性分布检验上也与词汇多样性 3 优博指数产生了相同的问题，高分组的数据不符合正态分布，因而词汇密度这一项检验因素与作文质量之间难以产生相关性关系[5]。在实际写作中，大量重复使用实词虽然能够提高词汇密度，但这会使文章用词反复，从而降低词汇变化性。

2.3.4. 词汇等级统计与质量分析

Table 3. Statistics of vocabulary levels are included in the composition of each level
表 3. 各等级二语作文包含词汇等级情况统计

	一级词	二级词	三级词	四级词	五级词	六级词	七—九级
高分组	33.18%	14.11%	12.16%	6.20%	3.71%	2.87%	5.70%
中分组	43.16%	16.90%	11.06%	5.30%	2.73%	1.68%	3.38%
低分组	44.26%	17.31%	10.58%	5.23%	2.40%	1.47%	2.68%

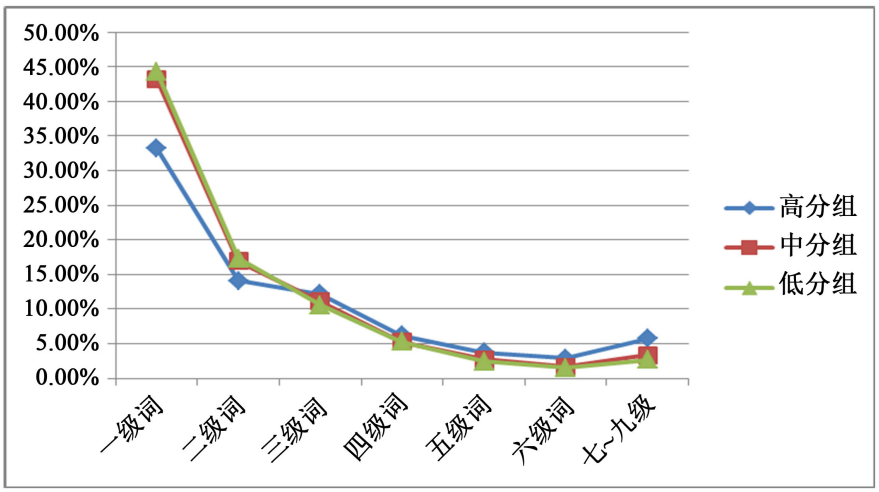

Figure 1. The composition of each grade contains the statistics of the vocabulary grade line chart
图 1. 各等级二语作文包含词汇等级折线图统计

表 3 是根据《国际中文教育中文水平等级标准》中的一至九级词相应地对各分组作文文本中的词汇进行数量统计，并计算出各级词汇所占百分比。为更直观反映数据变化趋势我们将其绘制为相应的折线图，如图 1 所示。

通过统计和观察我们可以得出以下结论：首先从总体看来，低、中、高三组文本使用不同等级词汇的数量变化整体趋势大致相同，遵从一级词至九级词使用数量愈来愈少的原则，符合学习者习得认知规律。其次通过观察数据可以发现，三级词为一个重要节点，在此之前的一级、二级词，使用数量是随着文本等级的降低而逐渐增多的，但从三级词开始往上，文本的各等级词汇使用情况则随着文本等级的降低也有所减少。最后从数量上看，一级、二级、三级词在总词汇数量中占据了大部分比例。从以上数据分析我们可以得知：在汉语作为第二语言习得者的作文产出中，一二三级这些初等词汇是构建文本的主体，不论在高分作文还是低分文本中都是如此，三级词汇虽然有随着文本难度升级使用数量逐渐减少的趋势，但由于其使用比例均在百分之十往上，亦可当作构建基础性文本的一部分。中等及高等词的使用比例均在百分之十以下，且根据其数量趋势变化我们可以得知，词汇等级对作文质量是具有相当的影响的，文本使用高级词汇数量越多，质量越好。

3. 分析结果与建议

在上述我们对影响词汇丰富性的相关要素进行了简要阐述和说明，并进行了基于词汇丰富度对作文质量的相关性统计分析，现将结论大致归纳如下：经由 python 设计程序对文本进行统计分析，归纳出词汇数量、词种数量、词汇密度、词汇复杂度等多项数据，并利用 spss 对其进行分析总结，最终得出，词汇复杂度与作文质量具有显著相关性关系，词汇密度与作文质量无相关性关系，词汇多样性由于计算方式多样因而其测量方式也不同，其与作文质量的相关性关系也呈现出较为复杂的态势，其中词汇数量与词种数量对作文质量是具有较大影响的，且存在决定作文质量是否合格的重要词汇数量节点。除此之外，本文还对不同分组作文中的词汇等级进行了分级统计，并计算出各级别所占百分比，可以发现，词汇等级与作文质量之间存在显著相关性。针对以上结论，在汉语作为第二语言的作文学习及产出中我们可以相应地提出一些建议，使其对二语教学中的作文质量提高有所裨益。首先，本文所采集语料为 HSK 动态作文语料库中的相关作文，虽然不同人所创作的作文质量不尽相同，相应地作文分数也具有显著差异，但作文创作者参加同等水平考试，可以默认其都接受了大致相同的汉语水平教学。在此前提下，对作文质量影响最为显著的便是词汇数量及词种数量，学习者应注意对词汇的积累与应用，对同一话题使用尽可能多、变化丰富的词汇，延长文本所需长度，这是一篇文章要达到合格首要的要求[6]。敦促学习者在多种语境下尝试使用不同的词汇，并且写作可以帮助学习者对语言形式的关注，这对于提高学习者词汇的多样性和准确性都有帮助。

其次词汇等级也对文章质量具有一定的影响，根据上述数据我们可以很明显地看出等级词在不同分数作文中的应用情况及变化趋势，尽量选取三级以上词汇，增加其在总体词汇中的比例，对提升作文质量具有一定的帮助[7]。

本研究中借鉴以往学者对二语学习者产出性词汇即作文产出的质量影响因素研究，从词汇角度出发，其中的构成词汇丰富度的相关要素的界定及计算方式也仍有待统一，但不可否认其仍具有一定的研究价值。加之本文的语料数量较大，采用的 python 语言进行分词、汇总等工作相较于人工难免存在误差，在后续的研究中可以对其进行进一步精确。其次，作文质量的影响因素众多，词汇丰富度只是其中之一，本文对词汇丰富度与作文质量之间的相关性关系做了统计分析，并不否认其他诸如语法、句长等因素对作文质量的影响。

参考文献

[1] 苏丹洁, 陆俭明. “构式-语块”句法分析法和教学法[J]. 世界汉语教学, 2010(4): 557-567.

[2] 张江丽. 汉语第二语言学习者产出性词汇复杂性研究[J]. 云南师范大学学报, 2020, 18(5): 48-57.

[3] 周敏, 俞芳芳, 仇心浩. 英语专业大学生产出性词汇研究[J]. 英语广场, 2019(10): 126-127.

[4] 李春琳. 汉语二语学习者产出型词汇水平和写作质量相关关系分析[J]. 华文教学与研究, 2017(3): 54-61.

[5] 左葳岳. 语料库视域下英语学习者写作词汇丰富性研究[J]. 大众文艺, 2022(20): 102-104.

[6] 赵玮. 汉语作为第二语言词汇教学“语素法”适用性研究[J]. 世界汉语教学, 2016, 30(2): 276-288.

[7] 张家强, 郭丽. 外语写作焦虑与写作质量: 词汇丰富性差异研究[J]. 二语写作, 2021(2): 26-37.