

日语学习者产出作文词汇丰富性的研究

——以日语学习者语料库（I-JAS）为对象

杨纯 王星薄

天津外国语大学 天津市 300204

摘 要：词汇丰富性是衡量语言能力的重要指标之一，能够反映学习者在产出过程中对词汇的掌握和运用情况。在二语习得研究中，词汇的使用情况不仅影响语言表达的准确性和流畅性，也与学习者的整体语言发展水平密切相关。因此，对学习者的产出文本中的词汇丰富性进行系统分析，有助于深入理解不同阶段学习者的词汇发展特点。本文选取 I-JAS 语料库中 N5-N1 各等级学习者的 150 篇作文作为分析对象，从词汇多样性、词汇密度及词汇复杂度三个维度考察了不同日语能力等级学习者在产出作文中的词汇使用特征。研究发现，随着日语能力等级的提升，学习者产出作文的词汇丰富性整体呈现出上升趋势，具体而言，日语学习者在 N5 到 N4 以及 N3 到 N2 这两个阶段产出作文的词汇多样性显著提升；在 N4 到 N2 阶段中词汇密度经历了量变到质变的过程；在 N3 到 N2 的阶段，其词汇复杂度显著提升。

关键词：词汇丰富性；二语习得；日语学习者

1. 引言

词汇习得一般可分为接受性词汇和产出性词汇。接受性词汇指的是学习者能够理解但未必能主动使用的词汇，通常出现在阅读和听力输入中；而产出性词汇则指学习者能够在口头或书面表达中自主运用的词汇，直接关系到语言输出的质量。研究表明，学习者的接受性词汇通常大于产出性词汇，即使在高级阶段，许多学习者仍然难以将被动掌握的词汇转化为主动使用（Nation, 2001）^[2]。因此，通过考察学习者作文中的产出性词汇，不仅能揭示其实际的词汇运用能力，还能反映词汇习得过程中从输入到输出的转化情况，而词汇丰富性则是用于评估学习者词汇使用水平的关键概念。

2. 文献综述

学习者产出性词汇丰富性相关的过往研究中，对于词汇丰富性的考察角度都略有偏差。Laufer（1991）将词汇丰富性定义为词汇独特性、词汇密度、词汇复杂性和词汇多样性^[1]。Read（2000）则认为，词汇丰富性可以通过词汇多样性、词汇复杂度、词汇密度、词频概貌以及词汇错误这几个维度进行评估^[3]。随后的研究多延续 Laufer（1991）和 Read（2000）的框架，利用不同指标对学习者的作文中的词汇进行了量化分析。

以日语学习者作为对象的先行研究在数量上相形见绌，

但基本也是基于上述指标进行分析。张建华（2015）通过自建作文语料库，从词频分布、词汇多样性、词汇复杂性以及词汇密度这四个维度考察了 2-4 年级的中国日语学习者在词汇广度知识发展过程^[7]。杨晓敏（2021）则自建学习者纵向语料库，从词汇多样性、词汇复杂性、词频分布、词汇密度四个维度研究了学习者作文中的产出词汇从初级到中级阶段性的变化^[6]。总体而言，以日语学习者作为对象的词汇丰富性相关研究数量较少，且对学习者的日语能力阶段的划分不够细致，无法细致把握日语学习者从初级到中级再到上级过程中产出词汇丰富性的整体倾向。

基于此，本文尝试通过对各等级学习者产出作文的词汇丰富性各个指标进行对比，旨在揭示不同日语能力等级学习者在词汇使用上的发展轨迹。

3. 研究设计

3.1 研究问题

基于上述先行研究，本文主要回答以下问题：N5-N1 不同等级日语学习者随着等级的提升产出作文的词汇丰富性整体呈现出怎样的发展路径？

3.2 语料来源和处理

本文使用的学习者作文语料来自日语学习者横断语料库（I-JAS）中海外环境学习者的作文数据①^[4]。

I-JAS 语料库包含了 20 个不同国家学习者的口语和书面语数据,且通过日语能力测试 J-CAT ②精准测定其日语水平。本文以此为依据划分学习者日语能力水平。

I-JAS 中收录了来自海外教室环境、日本国内教室环境以及日本国内自然环境的日语学

习者和日语母语者撰写,主题为「私たちの食生活:ファーストフードと家庭料理」(我的饮食生活:快餐和家庭料理)的作文素材。本文选取了其中海外教室环境下学习者的作文样本。根据 J-CAT 官网中对各等级分数的划分标准,本文抽取了 N5-N1 各等级学习者的作文各 30 篇,共计 150 篇。

语料处理方面,本文沿用了张建华(2015)对文本的处理方式。首先,删除了文本中出现的地名、食物名以及店铺名。其次,对作文中能辨识的拼写错误和不规范写法进行修正。此外,将系统无法识别的复合词分解为单一词汇,以防止软件在词频统计时出现误判。

3.3 研究方法和工具

本文参考 Read(2000)对词汇复杂度的分类,综合考虑了词汇密度、词汇多样性和词汇复杂性三个测量指标对语料中的词汇进行不同维度的分析。

词汇多样性是衡量学习者词汇使用范围和种类的关键指标。本文参考郑咏滢(2015)^[8]的做法,采用 Uber 指标来反映作文的词汇多样性。公式如下:

$$\text{Uber} = (\text{Tokens})^2 / (\text{Tokens} - \log \text{Types})$$

在上述公式中,形符数(Type)和类符数(Token)是计算词汇多样性的基础数据。

词汇密度主要测量一篇文本中实词所占的比率,体现文本的信息含量。本文参考杨晓敏(2021)的做法,将名词、代词、动词、形容词、副词、连体词和接续词定义为实词,计算它们在文本中的比率。公式如下:

$$\text{词汇密度} = (\text{名词} + \text{代词} + \text{动词} + \text{形容词} + \text{副词} + \text{连体词} + \text{接续词}) / \text{Tokens} \text{ ③}$$

词汇复杂度是衡量文本中低频词占比的重要指标,能在一定程度上反映学习者词汇产出的发展水平(鲍贵,2008)^[5]。本文参考张建华(2015)的做法,利用在线分析工具「リーディングチュウ太」④来获取每篇语料中各等级词汇的数量。「リーディングチュウ太」以新日语能力考试词汇表为基准,将词汇分为级外、N1、N2N3、N4、N5 及其他这 6 个等级,并自动统计各个级别词汇的类符数及其占比。本文

将 N1 等级以及级外词汇视为低频词,并计算其在整篇文本中的比例。公式如下:

$$\text{词汇复杂度} = (\text{N1 词数} + \text{级外词数}) / \text{Tokens}$$

然而,在分析过程中发现题目中的「ファーストフード」(快餐)和「食生活」(饮食生活)被错误归类至级外词汇,因其并非学习者自主产出,本文在计算级外词数时进行了人工剔除。

本文使用的数据分析方法为多元方差分析,使用的分析软件为 IBM SPSS Statistics 27.0.1。

4. 研究结果和讨论

在进行数据分析之前,本文首先对数据进行了正态性和方差齐性检验。Shapiro-Wilk 正态性检验结果表明,词汇密度($p=0.254$)和词汇多样性($p=0.387$)均符合正态分布,而词汇复杂度($p < .001$)不满足正态性假设。此外,Levene 基于平均数的方差齐性检验显示,词汇密度($p=0.45$)与词汇复杂度($p=0.844$)满足方差齐性,而词汇多样性($p=0.018 < 0.05$)不符合方差齐性假设。

由于方差分析要求数据近似正态分布,本文采用 Box-Cox 变换对词汇复杂度数据进行转换。Box-Cox 变换是一种常用于处理非正态分布数据的变换方法,通过对数据应用一个幂变换,使其分布接近正态分布。转换后的数据在 Shapiro-Wilk 检验($p=0.198$)和 Levene 检验($p=0.294$)中均满足正态性与方差齐性假设。因此,本文最终采用转换后的数据进行方差分析,以确保分析结果的可靠性和稳健性。

4.1 词汇多样性的考察结果

为了考察不同日语能力等级学习者产出作文中词汇多样性的差异,本文对 Uber 的结果进行了方差分析。从表 2 数据可知,不同等级学习者在词汇多样性上存在显著差异 [$F(4, 145) = 23.540, p < .001$],说明日语能力等级对词汇多样性有显著影响。

另外为了进一步探究不同等级间词汇多样性的具体差异,本文对各等级的数据进行了多重比较 Tamhane's T2 检验,结果显示, N1 和 N2 等级间词汇多样性无显著差异,但与其他等级差异显著。N3 和 N4 等级间无显著差异,与其他等级则有显著差异。N5 等级与其他所有等级差异显著。

这些结果表明,初级日语学习者在词汇使用方面与中高级学习者存在较大差距。同时强调了随着学习者能力的提升,其词汇多样性也随之增加,特别是在 N5-N4 以及

N3-N2 这两个阶段,学习者在词汇的掌握和运用能力上取得了显著进步,能更加灵活地使用词汇来丰富其语言表达。

4.2 词汇密度的考察结果

为了考察不同日语能力等级学习者产出作文中词汇密度的差异,本文对词汇密度的结果进行了方差分析。方差分析结果表明,不同日语能力等级学习者产出作文中词汇密度存在显著差异 [$F(4, 145) = 6.549, p < .001$],说明日语能力等级对词汇密度存在显著影响。

另外为了进一步探究不同等级间词汇多样性的具体差异,我们对各等级的数据进行了多重比较 Scheffé 检验。结果显示, N1 与 N2、N2 与 N3 等级间在词汇密度上无显著差异,而 N3、N4 和 N5 等级学习者之间的词汇密度差异也不显著。然而, N4 与 N2 等级之间存在显著差异,这表明尽管 N4 至 N3 再到 N2 两阶段单独比较时未显示明显变化,但在 N4-N2 的整体进程中,学习者的词汇密度能力产生了显著提升,呈现出量变到质变的过程。

4.3 词汇复杂度的考察结果

为了探究不同日语能力等级学习者产出作文中词汇复杂度的差异,本文对词汇复杂度 Box-cox 转换后的结果进行了方差分析。分析结果显示,不同等级学习者在词汇复杂度上存在显著差异 [$F(4, 145) = 17.109, p < .001$],这表明日语能力等级对词汇复杂度有显著影响。

另外, N5 等级学习者的词汇复杂度平均值 (-1.47) 略高于 N3 (-1.48) 和 N4 (-1.52) 等级,这反映了初级学习者在日语学习初期更倾向于大胆使用多样化词汇。随着日语能力的提升,学习者可能更加注重用词的准确性,从而导致词汇复杂度有所降低。

此外,通过 Scheffé 多重比较检验进一步分析不同等级间的词汇多样性差异,结果显示 N1 和 N2 等级学习者在词汇复杂度上无显著差异,且均显著高于 N3、N4、N5 等级,而 N3、N4、N5 等级间的差异并不显著。这些结果表明,日语学习者在从 N3 到 N2 的阶段,其词汇复杂度有显著提升,反映了他们在语言运用上的更高熟练度和词汇知识的丰富性。

结语

本文基于 I-JAS 语料库中海外学习者的作文数据,分析了 N5 至 N1 等级日语学习者作文中的词汇多样性、密度和复杂度,以探讨其随等级提升的词汇丰富性发展路径。结果显示,随着日语能力提高,学习者作文的词汇丰富性整体呈上升趋势。

具体来说,词汇多样性在 N5-N4 以及 N3-N2 这两个阶段显著提高;词汇密度在 N4-N2 阶段经历了量变到质变的过程;词汇复杂度在 N3-N2 阶段存在显著提升的趋势。

然而,研究仍存在局限性,如样本量有限且仅涵盖海外学习者,可能无法全面代表所有日语学习者。未来研究应扩大样本范围,纳入更多背景的学习者,以增强研究结果的广泛性和适用性。

参考文献:

- [1] Laufer B. The development of L2 lexis in the expression of the advanced learner[J]. The Modern Language Journal, 1991, 75(4): 440-448.
- [2] Nation I S P. Learning Vocabulary in Another Language[M]. Cambridge: Cambridge University Press, 2001: 26-30.
- [3] Read J, Read J A S. Assessing vocabulary[M]. Cambridge University Press, 2000.
- [4] 迫田久美子 (2020) 「I-JAS 誕生の経緯」迫田久美子・石川慎一郎・李在鎬 (編著) 『日本語学習者コーパス I-JAS 入門: 研究・教育にどう使うか』 (pp.2-13) くろしお出版.
- [5] 鲍贵. 二语学习者作文词汇丰富性发展多纬度研究[J]. 外语电化教学, 2008, (05): 38-44.
- [6] 杨晓敏, 钟勇, 赵寅秋. 中国学习者日语词汇丰富性发展实证研究[J]. 第二语言学习研究, 2021, (02): 59-72+133-134.
- [7] 张建华. 日语专业学生产出性词汇广度知识发展的研究[J]. 当代外语研究, 2015, (03): 31-35.
- [8] 郑咏滢. 基于动态系统理论的自由产出词汇历时发展研究[J]. 外语教学与研究, 2015, 47(02): 276-288+321.

注释:

- ①『多言語母語の日本語学習者横断コーパス: I-JAS』の官方网址为: <https://www2.ninjal.ac.jp/jll/lsaj/ijas-search-info.html> (2025/3/14)
- ② J-CAT (Japanese Computerized Adaptive Test) 的官方网址为: <https://www.j-cat2.org/> (2025/3/14)
- ③本文词汇多样性以及词汇密度的计算均通过 Python 编程语言实现。
- ④「リーディングチュウ太」是由东京国际大学言语コミュニケーション学部の川村よし子教授主导开发的日语阅读教学辅助系统, 官方网址为: <https://chuta.cegloc.tsukuba.ac.jp/> (2025/3/14)