人工智能在日语自动写作评价中的应用

王嘉欣

西安外国语大学，陕西西安，710128；

摘要：本文探讨了 ChatGPT 在日语自动写作评价（AWE）中的应用潜力，并与现有的日语 AWE 系统 jWriter 和 GoodWriting Rater 进行了比较。研究分析了 ChatGPT、AWE 系统 jWriter 和 GoodWriting Rater 对 I-JAS 语料库中的 100 篇作文的评价结果，与日语能力测试 J-CAT 的词汇、语法、阅读成绩及名词、动词、助词等语言统计量之间的相关性。结果显示，与 AWE 系统 jWriter 和 GoodWriting Rater 相比，ChatGPT 的评分结果与 J-CAT 各项分数呈现更强的正相关性，在语言特征量分析中，ChatGPT 虽与 AWE 系统在部分语言特征量上存在相关性，但其评价方式更加灵活，不依赖固定规则。研究表明，以 ChatGPT 为代表的人工智能在日语自动写作评价中具有较大的应用潜力，为未来的发展提供了新的方向。

关键词：人工智能；ChatGPT；自动写作评价（AWE）；日语写作

DOI：10.69979/3041-0673.25.09.073

引言

人工智能时代的到来，推动了教育的智慧化发展，为现有教学模式、教育评估以及学习方式等带来了革新与挑战。OpenAI 公司于 2022 年推出的基于大语言模型的生成式人工智能 ChatGPT，在语言理解、智能对话、文本生成等方面展现出优秀表现，具有广阔的应用前景。外语写作教育作为其中一个领域，ChatGPT 的应用贯穿教学的多个环节。在教学准备阶段中，它能为学习者制定个性化学习计划，为提供教师教学资源和课程设计支持（龚邵华，2024；毛文伟，2023）；在教学过程中，可以辅助教师指导，给予学习者个性化学习体验与即时反馈（郭茜等，2023；焦建利 2023；陈茉和吕明臣，2024）；在教学评估中，能够协助教师纠错并提出评价反馈，减轻教师负担、提升教学效率。（魏爽，2023；Guoet al.，2024；Mizumoto & Eguchi，2024）。作为一种强大的自然语言处理工具，ChatGPT 的应用涵盖多领域多层面，为外语教育开辟了新的发展途径，带来了深远的影响。

自动写作评价（Automated Writing Evaluation，AWE）研究是当前智慧教育方兴未艾的领域。早在 20 世纪 60 年代，Page 教授团队开发了首个自动写作评价系统（以下简称 AWE 系统）PEG(Project Essay Grade)。20 世纪 90 年代以后，随着自然语言处理技术不断提升，自动写作评价研究也得到了进一步发展，IEA(Intelligent Essay Assessor)、e-rater（Electronic Essay Rater)相继出现。其中 e-rater 被广泛用于研究生入学考试（GraduateRecord Examination,GRE）和面向母语非英语者英语能力考试（Test of English as a Foreign Language,TOEFL）等大规模考试。相较于人工评价，AWE 系统不仅在日常写作教学中有效缓解了教师的工作负担，还在大规模考试中发挥了节省经济和时间成本的优势，提高了评价的效率和一致性。因此 AWE 逐渐成为语言学习和教学领域的重要工具。然而，自动写作评价研究主要集中于英语教育中，在日语教育方面尚处于发展初期阶段，目前有 jWriter（2017）和 Goodwriting Rater（2018）两种 AWE 系统，它们在写作能力评价方面展现出了一定的有效性。但因为其主要依赖统计语言特征量的评价方式，未能充分考虑语言知识的正确使用及内文章的内容结构，评价结果的准确性和全面性有待提高，所以仍需结合教师人工评价。近年来，基于大语言模型的生成式人工智能的兴起，自然语言处理技术得到了显著性提升，以 ChatGPT 为代表的生成式人工智能具有出色的语言理解和文本生成能力，为日语自动写作评价的发展带来了新的突破口。

本研究旨在探讨以ChatGPT为代表的人工智能在日语自动写作评价中应用的可能性，为日语自动写作评价进一步研究提供参考。具体而言，首先介绍当前日语写作教育中现存的 AWE 系统并梳理相关先行研究。其次，具体分析 ChatGPT 在自动写作评价中的表现，与现存 AWE 系统进行比较来评估其优势与局限性。最后，总结研究成果并展望未来自动写作评价的发展方向。

1 日语自动写作评价系统概要

jWriter 是面向日语学习者的 AWE 系统。jWriter 以“多母语日语学习者横向语料库（International Corpus of Japanese as a Second Language: I-JAS）”中 12 个国家初、中、高级别的 373 名母语者作文为标准数据，提取了总字数、句数、段数、平均字数、词汇多样性、连词等 44 种语言特征，其中重点统计平均字数、统计总字数、动词、不同等级词汇数、多样性、和语词、汉语词七大特征量，使用多重回归分析法构建评价公式对作文进行级别入门、初级、中级、高级、超级五个级别的综合判定。待评价文本最低字数要求 300 字以上，字数越多准确性越高，系统建议 600 字左右为宜。超过 1000 字以上的文本还可以进行逻辑性评价，逻辑性评价以期刊及中小学教科书等为标准数据来构建评价公式，将文章划分为优、高、中、低四等级。从评价结果来看逻辑性较弱的文本有两个特点：一是句与句之间的连词使用较少，二是体现文章信息量的名词、动词及形容词等使用较少。jWriter 建议学习者可以通过改善这两点来提高文章的逻辑性。为了给学习者提供更好的反馈以提高其写作能力，jWriter 会显示字数、句数、段数等文本数量信息，按照级别或词典的形式等显示词汇，用颜色区分不同难易度的词汇和不同功能的连词，此外还能以词云及词汇网络等可视化的形式进行专业的文本分析。

GoodWriting Rater 是面向日语教师专门针对议论文的 AWE 系统。选取 I-JAS 语料库之中选取的 611 篇及日本、韩国、中国台湾大学生日语议论文语料库中的 134 篇等共计 1056 篇日语学习者议论文作为标准数据，将日语教师的人工评分数据作为因变量，除总字数、段落数、按词类区分词汇数、按等级区分词汇数外，包括关键词数、元语言表达数、首段字数/总字数、尾段字数/总字数等不同角度评价观点在内的 62 种文本特征作为自变量，使用逻辑回归法构建出评价模型，最后通过评价模型得出 1-2（中级前半接近中级）、3（中级后半接近中极）、4（中级后半，N2）、5-6（高级前半，N1）四种评分。与 jWriter 不同，该系统由整体评价（Holistic scoring）和多元评价（Multiple-trait scoring）两部分构成。整体评价指的是评价时不过于拘泥细节，更关注整体印象，像英语考试雅思和托福一样，对文章进行综合性评价。多元评价分为目的和内容、结构和连贯性、日语表达三个角度。目的和内容中，目的指是目标达成，即是否比较并陈述了观点，内容方面评价论点是否明确统一，论据是否可靠有力；结构和连贯性中，结构指的是写作中是否有结构意识和段落意识以及引言、正文、结论整体结构，连贯性指段落之间的连接是否连贯；最后日语表达的部分指评价语言知识的正确性、多样性和适当性。除了判定分数，GoodWriting Rater 也可以像 jWriter 一样显示总字数、总句数、汉字使用率、每句平均字数、首段占全文百分比、尾段占全文百分比等基础文本信息并使用元语言按顺承、转折、对比、举例等功能对连词进行重点标注。

目前，日语教育中 AWE 系统停留在研究层面，并未在大规模考试及师生之间被广泛使用。小森（2018）在分班测试中发现，jWriter 评价结果与教师整体性评价与学生的日语能力之间存在相关性，这一结果展现出了 jWriter 在一定程度上具有代替教师进行写作评价的可能性。小森（2020）进一步发现 jWriter 在分班测试中的评价等级不仅体现了学习者语言知识量的差异，还体现了其正确运用语言能力的差异。随后小森（2022）研究发现 jWriter 评价结果与日语客观性测试中的文字词汇、语法、阅读部分均呈现相关性，这意味着可以用 jWriter 来推测学习者的语言能力。此外还发现，jWriter 评价结果与总字数、平均句长、汉字词等 20 项语言特征量呈现相关性，说明 jWriter 对于衡量学习者语言知识量和质两方面的能力有一定的参考价值，然而，由于其只是对语言信息量进行了统计，对语言准确性和适当性、内容结构等质方面的分析仍需要依赖教师的人工判断。从不同 AWE 系统的角度出发，小森（2024）对 jWriter 和 GoodWriting 的评价结果进行了比较，并将两种系统的评价结果与教师评价结果进行对比，发现两种 AWE 系统评价结果之间存在相关性，教师评价与 jWriter 之间存在相关性，与 GoodWriting 之间不存在相关性。也就是说，在作文整体性评价方面，jWriter 的结果更具参考性，对作文的内容、结果方面的评价上，GoodWriting 的结果更具有效性。因此，在实际运用中，应根据不同的评价方式合理选择适当的 AWE 系统，以提高评价结果的有效性和可靠性。此外，影山（2019）则使用 GoodWriting 对日本学生的作文进行了评价，并调查了对评价结果的接受度探讨 AWE 系统是否适用于日本学生写作能力的评估。从结果来看，在整体性评价方面，70%的学生写作能力被评为最高等级；在多样性评价中，80%的学生在日语能力方面达到最高水平。然而，在内容、结构与连贯性两方面，评价结果呈现较大离散性，反映出了日本学生在写作能力上的差异。有关日本学生对于评分的接受度，发现评价的接受度与评分高低呈正

相关，即评价结果越高，学生的接受度越高。另外，对于 GoodWriting 提供的反馈，影山认为即使提供了文本信息和元语言信息，日本学生仍难以有效运用，为了提高 GoodWriting 的应用效果，教师的介入与指导至关重要。

2 人工智能 ChatGPT 在日语自动写作评价中的应用

2.1 研究问题

基于大规模语言模型的 ChatGPT 经过海量数据集（包括新闻、书籍、报纸等）的训练下，掌握了丰富的日语表达方式，它不仅能准确理解词语的含义，还能深入分析上下文，为日语写作自动评价提供了技术方面的支持。包括 jWriter 和 Goodwriting 在内的 AWE 系统通常都是在提取了语言特征量后，经过预先选定的作文数据集进行训练，从而来构建评价公式对作文进行判定，基于这一特点，我们研究 ChatGPT 在日语自动写作评价的应用时首先有必要讨论评价结果与语言特征量的关系。其次写作能力作为日语能力的重要体现，我们还需要探究 ChatGPT 的评价结果是否能真实反应学习者的日语能力。在此期间，我们会将 ChatGPT 这两个方面的结果分别和 AWE 系统 jWriter 和 Goodwriting 进行对比，讨论 ChatGPT 是否具有更好的写作评价效果，为人工智能引入日语自动写作评价提供参考。因此本研究分为两个研究问题：

（1）ChatGPT 对日语作文的评价结果是否与语言特征量呈现相关性

（2）ChatGPT 对日语作文的评价结果是否与日语能力测试 J-CAT 分数呈现相关性

2.2 研究方法

随机选取 I-JAS 语料库中题为「食生活」的 100 篇作文。分别使用 jWriter 和 Goodwriting 进行评价，记录评价结果并统计每篇作文的语言特征量，然后对 ChatGPT4o 发出日语指令：假如你是经验丰富的日语老师，请根据题目要求从内容和结构两方面分别对文章进行评价（1-5 分）并写出清晰具体的评语。题目要求：私たちは日常生活で、ファースト･フードと家庭でゆっくり味わう手作りの料理を食べています。ファースト･フードと家庭料理を比較し、それぞれの良い点や悪い点などを説明して、「食生活」についての意見を 600 字程度で書いてください。记录评价结果并与语言特征量和日语能力测试 J-CAT 分数进行相关性分析。J-CAT 是在线评估非日语母语者日语能力的测试，主要以客观题来考察听力、词汇、语法、阅读这四项核心语言技能，采用计算机自适应测试技术，随时根据考生的答题情况动态调整题目难度，具有灵活、全面和准确的特点。

2.3 ChatGPT、jWriter 和 Goodwriting 的评价结果与日语能力测试 J-CAT 之间的相关性

使用皮尔逊相关系数分别对 ChatGPT、jWriter 和 Goodwriting 三者的评价结果与日语能力测试 J-CAT 的词汇、语法、阅读之间进行了相关性分析。从表格中可以看出 ChatGPT 的评价结果与总分、词汇、语法、阅读四项均呈现正相关性，该结果与李（2023b）研究结果 ChatGPT 的评价结果与客观测试总分之间存在相关性（r=.41）相符，说明 ChatGPT 能够较好地反映学习者的综合日语水平。此外，与词汇呈现正相关性说明其能够反应文本中词汇丰富性，与语法呈现正相关性说明其能够在一定程度上体现文本的语法，与阅读呈正相关性说明其在文本的整体可读性和理解性方面具有较高的评价能力。Goodwriting 的评价结果除语法外，与测试总分、词汇、阅读三项呈现正相关性。值得一提的是比起 AWE 系统 jWriter 和 Goodwriting，ChatGPT 与 J-CAT 测试总分、词汇、语法、阅读四项的相关性更强，意味着 ChatGPT 具有一定的可靠性和有效性，在评价文本质量时更具优势。

表 1 评价结果与日语能力测试 J-CAT 之间的相关性

	ChatGPT	jWriter	GoodWriting Rater
J-CAT 語彙/100	r=0.458**	r=0.366**	r=0.251*
J-CAT 文法/100	r=0.409**	r=0.286**	r=0.118
J-CAT 読解/100	r=0.462**	r=0.392**	r=0.281**
J-CAT 合計/400	r=0.578**	r=0.434**	r=0.321**

2.4 ChatGPT、jWriter 和 Goodwriting 的评价结果与语言特征量之间的相关性

对 ChatGPT、jWriter 和 Goodwriting 三者的评价结果与语言特征量之间的相关性进行分析后，我们可以发现：ChatGPT 的评分结果与名词，助词，汉语词，和语词，总字数，每句平均字数共 6 项之间有着正相关关系。jWriter 的评分结果与名词，动词，助词，汉语词，和语词，外来语，连接词，总字数，每句平均字数共 9 项之间有着正相关关系。GoodWriting Rater 的评分结果与名词，动词，助词，汉语词，和语词，外来语，总字数，每句平均字数共 8 项之间有着正相关关系。可以看出 ChatGPT、jWriter 和 Goodwriting 三者均与名词、助词、汉语词、和语词、总字数这五项相关，这表明三种评分标准中都包含了基本的语言特征量。然而，由于 ChatGPT 的训练方式与 AWE 系统 jWriter 和 Goodwriting 存在本质区别，其评价方式并不依赖固定的语言特征量或基于类似语法规则的标准，而是通过学习大量来自书籍、网站等文本数据的语言模式，并利用对语言的理解和生成能力进行评价。ChatGPT 是一个不断更新扩展的动态语料库，包含了正式与非正式、标准与非标准等多种语言信息，因此它的评价标准更加灵活，相较于 AWE 系统给予预设规则和固定的语言特征量进行评分，ChatGPT 更倾向基于统计和概率来判断哪些语言表达，结构更为常见、更符合优质文本的特点。ChatGPT 之所以在评价结果与名词，助词，汉语词等语言特征量呈现相关性，是因为其在广泛的训练数据中学会了与优质文本相关的语言特征。此外，它的评分不仅依赖传统的语法特征，还可能受到语境、上下文逻辑、表达方式等多方面因素的影响。综上，ChatGPT 与 AWE 系统 jWriter 和 Goodwriting 不同的评分方式，使其在评价时更加灵活，而 AWE 系统 jWriter 和 Goodwriting 的评价结果相对稳定，主要侧重于文本的语言特征量分析。

表 2 评价结果与语言统计量之间的相关性

	ChatGPT	jWriter	GoodWriting Rater
名词	r=0.264**	r=0.737**	r=0.482**
动词	r=0.145	r=0.406**	r=0.237*
形容词	r=0.111	r=0.077	r=0.13
助词	r=0.202*	r=0.638**	r=0.362**
副词	r=0.005	r=0.062	r=0.061
汉语词	r=0.311**	r=0.779**	r=0.523**
和语词	r=0.197*	r=0.575**	r=0.315**
外来语	r=0.184	r=0.237*	r=0.229*
连接词	r=0.068	r=0.207*	r=0.01
总句数	r=-0.121	r=-0.186	r=-0.182
总字数	r=0.217*	r=0.632**	r=0.370**
总段落数	r=-0.157	r=-0.177	r=-0.035
每句平均字数	r=0.296**	r=0.734**	r=0.556**
首段占全文比例	r=0.01	r=0.06	r=-0.19
尾段占全文比例	r=0.078	r=0.059	r=-0.041

3 总结与展望

本文探讨了以 ChatGPT 为代表的人工智能在日语自动写作评价中的应用潜力，结果表明。与现有的 AWE 系统 jWriter 和 GoodWriting Rater 相比，ChatGPT 的评分结果与 J-CAT 测试成绩具有更强的正相关性，能够较好地反映学习者的综合日语水平。ChatGPT 与 AWE 系统在部分语言特征量上存在一定相关性，但其评价方式更加灵活，能够基于广泛的语言模式和上下文进行综合判断，而非仅依赖预设规则。

李（2021）指出尽管现存自动评分系统虽在评分精度和评分结构方面仍有不足，面临许多亟需解决的课题，但仍是语言教育领域的一大创新，使计算机可以执行一直以来被认为是只有经验丰富的教师才能胜任的形成性评价任务，今后随着可利用信息的不断增加及多样的

计算模型的涌现，这一领域会得到长足的发展。推动自动写作评价发展的三大途径，一是构建大规模数据库，二是培养具有数据分析素养的日语教育者，三是挖掘多样的自动评价分析指标。从第一点来看，目前大规模日语学习者产出的语料库虽然可以进行字符串检索等，但是作为研究使用，规模有限。而自动写作评价系统要利用自然语言处理技术，在人类提前准备好的的训练数据基础上构建模型对文章进行评价，所以为了尽可能提高评价的准确度，必须构建大规模日语学习者语料库，不断调整评价模型，使其评分结果无限趋近于专业评价者的判断标准。未来，可以结合人工智能与现有 AWE 技术，构建更加精准、高效的日语写作评价系统。

参考文献

[1]龚韶华.ChatGPT 类技术之于大学英语教学的启迪与思考[J].集宁师范学院学报,2024,46(01):70-73.

[2]毛文伟,谢冬,郎寒晓.ChatGPT 赋能新时代日语教学：场景、问题与对策[J].外语学刊,2023,(06):25-33.

[3]郭茜,冯瑞玲,华远方.ChatGPT 在英语学术论文写作与教学中的应用及潜在问题[J].外语电化教学,2023,(02):18-23+107.

[4]焦建利,陈婷.大型语言模型赋能英语教学：四个场景[J].外语电化教学,2023,(02):12-17+106.

[5]陈茉,吕明臣.ChatGPT 环境下的大学英语写作教学[J].当代外语研究,2024,(01):161-168.

[6]魏爽,李璐遥.人工智能辅助二语写作反馈研究——以 ChatGPT 为例[J].中国外语,2023,20(03):33-40.

[7]Kai Guo, Deliang Wang. To resist it or to embrace it? Examining ChatGPT’s potential to support teacher feedback in EFL writing[J]. Educ Inf Technol ,2024(29):8435－8463 .

[8]Atsushi Mizumoto,Natsuko Shintani,Miyuki Sasaki,Mark Feng Teng.Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment[J].Research Methods in Applied Linguistics,2024,3(2):100116-100130.

[9]小森和子・李在鎬・長谷部陽一郎・鈴木泰山・伊集院郁子・柳澤絵美.教師による評価とコンピュータによる自動評価はどの程度一致するのか―中上級日本語学習者の意見文の評価を対象に―[J].2018 年度日本語教育学会秋季大会集,2018:278-283.

[10]伊集院郁子・李在鎬・小森和子・野口裕之.評価コメントに見られる意見文評価の様相－共起ネットワーク及びコレスポンデンス分析に基づく考察－[J].第二言語としての日本語の習得研究.2020(23):26-43.

[11]小森和子・伊集院郁子・李在鎬.日本語学習者の作文における自動評価と教師評価の比較[J].明治大学国際日本学研究,2022(14):41-67.

[12]小森和子.ライティング評価の新潮流 自動評価ツール「jWriter」による作文評価と言語習熟度[J].早稲田日本語教育学,2022(33):35-49.

[13]小森和子.自動評価システムによる作文評価は教師評価の代用になるのか[J].明治
大学国際日本学研究卷,2024(16):117-132.

[14]影山陽子.作文の自動評価システムの日本人学部大学生への活用可能性―評価への納得度と推敲への動機に着目して―[J].アカデミック・ジャパニーズ・ジャーナル,2019(11):28-36.

[15]李在鎬.「書くことを支援する自動評価システム『jWriter』（特集 AI や ICT が変える言語教育）」,2021,40(4):42-51.

[16]jWriter 学習者作文評価システム.https://jreadability.net/jWriter/.

[17]GoodWriting Rater 読み手と構成を意識した日本語ライティング. https://goodwriting.jp/wp/.