汉语学习者文本多维标注语料库建设 *

王莹莹[1] 孔存良[1] 杨麟儿[1] 胡韧奋[2] 杨尔弘[1] 孙茂松[3]

（1. 北京语言大学 北京 100083；
2. 北京师范大学 北京 100875；
3. 清华大学 北京 100084）

[摘要] 本研究以中介语多元对比分析的理论和方法为指引，以计算机智能辅助写作为导向，构建了一个大规模、高质量、篇章级别的汉语学习者文本多维标注语料库——YACLC。YACLC 设计了一套多维度富信息标注体系，包括最小改动、流利度提升、句子可接受度、上下文依赖性四个维度，采用众包策略标注了 2,421 篇、32,124 句语言使用场景下的汉语学习者文本，获取到 331,292 个最小改动标注句和 137,708 个流利提升标注句。YACLC 的建设既解决了现有汉语学习者语料库语料来源封闭、标注结果单一和流利维度欠缺的问题，又为学界分析学习者语言与两个参照语变体三者之间的多元互动，揭示二语习得的规律提供了参考及扩展空间。

[关键词] 学习者语料库；流利度；众包；语法自动纠错

[中图分类号] H087 [文献标识码] A [文章编号] 1003–5397（2023）01–0088–13

DOI:10.16499/j.cnki.1003-5397.2023.01.005

The Construction of Chinese Multi-dimensional Learner Corpus: YACLC

WANG Yingying, KONG Cunliang, YANG Liner, HU Renfen, YANG Erhong, SUN Maosong

Abstract: Guided by the theory and the methods of Contrastive Interlanguage Analysis and intelligent computer-assisted writing, this paper constructs a large-scale, high-quality, document-level, multi-dimensional annotated Chinese learner corpus, Yet Another Chinese Learner Corpus (YACLC). YACLC designs a multi-dimensional informative annotation guideline, including minimal edit, fluency

[收稿日期] 2022–03–11

[作者简介] 王莹莹，北京语言大学博士生，主要研究计算机辅助语言学习和语言监测；孔存良，北京语言大学博士生，主要研究计算机辅助语言学习；杨麟儿，北京语言大学副教授，博士，主要研究计算语言学和计算机辅助语言学习；胡韧奋，北京师范大学讲师，博士，主要研究中文信息处理；杨尔弘（通讯作者），北京语言大学教授、博导，博士，主要研究语言信息处理、语言监测和语言资源建设；孙茂松，清华大学长聘教授、博导，博士，主要研究自然语言理解、中文信息处理、Web 智能、社会计算和计算教育学。

* 本研究得到国家语委科研中心重点项目“智能辅助汉语应用文写作研究”（ZDI135-131）、教育部中外语言交流合作中心 2021 年度项目“汉语学习者偏误多维度标注语料库建设”（YHJC21YB-129）、北京语言大学语言资源高精尖创新中心项目“智能辅助汉语写作研究”（TYZ19005）和国家语言资源监测与研究平面媒体中心研究经费资助。本研究得到何姗、陈云、刘正皓老师的帮助与指导，谨此致谢！陆晓蓉和王一君同学参与了数据处理与分析，全体标注员付出了辛勤的劳动，编辑部和审稿专家提出了宝贵的修改意见和建议，在此一并深表谢意！

edit, sentence acceptability, and context dependence. Then YACLC annotates 2,421 Chinese learner texts of language usage scenarios with 32,124 sentences using a crowdsourcing strategy, to obtain 331,292 minimal edit annotations and 137,708 fluency edit annotations. The construction of YACLC not only solves the problems of closed data resources, single annotation and lacking of fluency dimension of the Chinese learner corpus, but also supports and extends the comparative analysis between the learner language and the two reference language variants to reveal the laws of second language acquisition.

Keywords: learner corpus; fluency; crowdsourcing; grammatical error correction; contrastive interlanguage analysis

一　引言

学习者语料库（Learner Corpus，也称中介语语料库[①]）是由第二语言学习者在说或者写的过程中产生的一系列文本的集合（Granger，2004）。一方面，研究学习者语料库提供的大规模语料有助于深入揭示二语发展规律（Nicholls，2003；曹贤文，2020）。另一方面，学习者语料库作为语法自动纠错（Grammatical Error Correction，简称 GEC）任务的数据集，被广泛运用到了自然语言处理领域中（Htut & Tetreault，2019）。研究者们通过构建模型自动获取语料库中学习者语言的特征，应用于计算机智能辅助写作研究（王辰成等，2020）。

其中，“偏误”指学习者掌握一定的目的语语法规则后出现的系统性错误（周小兵等，2007）。目前，研究者们对英语学习者文本的偏误标注研究已经颇有建树，已建成多个大规模、高质量的英语学习者语料库，如国际英语学习者语料库（International Corpus of Learner English，简称 ICLE）（Granger，2003）、朗文学习者语料库（Longman Learners Corpus，简称 LLC）、香港科技大学学习者语料库（Hong Kong University of Science and Technology Learner Corpus，简称 HKUST）（Milton，2001）、剑桥英语学习者料库（Cambridge Learner Corpus，简称 CLC）（Nicholls，2003）、新加坡国立大学英语学习者语料库（NUS Corpus of Learner English，简称 NUCLE）（Dahlmeier et al.，2013）等。此外，还有一些规模较小但标注更为精准和丰富的学习者语料库，主要用作语法自动纠错任务的数据集，如 HOO（Helping Our Own）2011（Dale & Kilgarriff，2011）、HOO2012（Dale，Anisimoff & Narroway，2012）、CoNLL（Conference on Computational Natural Language Learning）2013（Ng et al.，2013）、CoNLL2014（Ng et al.，2014）、BEA（Building Education Application）2019（Bryant et al.，2019）等多个 GEC 评测比赛中发布的数据集，用于评测参赛队伍所提交系统的改错性能。因训练 GEC 模型需要大量数据，所以研究者们往往使用多个学习者语料库做训练，同时也使用规模较大、学习者自发互相纠正错误的 Lang-8 数据集（Mizumoto et al.，2011）。

汉语学习者语料库的资源建设尚处于发展阶段，建设较早的有中山大学留学生中介语语料库（张舸，2008）、HSK 动态作文语料库（张宝林，2009）、南京师范大学中介语偏误信息语料库（周文华、肖奚强，2009）、TOCFL 学习者语料库（Chang，2013）、全球汉语中介语语料库（张宝林、崔希亮，2013）等。这些语料库旨在为汉语教学及相关研究提供一个基础平台，并为汉语本体研究提供参考（张宝林，2010），因此往往从字、词、句、篇、标点符号等多种角度，人工地全面标注偏误信息。近年来，研究者们逐步聚焦于中文语法错误自动检测（Chinese Grammatical Error Diagnosis，简称 CGED）和纠错任务上，相继开展了一系列评测比赛，并从上述语料库中挑选学习者文本形成评测数据集。例如：CGED 评测比赛于 2014~2020 年发布了来自 TOCFL 和 HSK 的多个训练集和测试集（Yu et al.，2014；Lee et al.，

2015；Lee et al.，2016；Rao et al.，2017；Rao et al.，2018；Rao et al.，2020），第七届国际自然语言处理及中文计算会议（Natural Language Processing and Chinese Computing 2018，简称 NLPCC 2018）的中文 GEC 评测比赛发布的数据集，训练集选自 Lang-8 数据集的训练集，测试集选自北京大学汉语学习者语料库（PKU Chinese Learner Corpus）（Zhao et al.，2018）。

从二语习得研究来看，现有的汉语学习者语料库大多数局限于建库时的设计目标，仅仅适用于单一维度的学习者语言分析理论和方法，不足以全面描述学习者语言的特征和规律。因此，有必要以二语习得研究前沿理论和方法为指引，“建设能有效服务于动态、多元、多维的二语习得研究范式的高质量汉语学习者语料库”（曹贤文，2020）。

另一方面，从服务于智能辅助学习的技术研发来看，现有的汉语学习者语料库也还存在如下问题：第一，语料来源主要是语言能力测试的课堂、作业、考试场景，题材和文体相对固定，话题相对封闭，无法全面反映开放场景下学习者的真实语言使用情况。与之相比，英语学习者语料库已经囊括多个话题下的作文文本，并且开始关注话题较为自由的辅助写作平台的文本。例如，BEA2019 发布的 W&I 数据集（Bryant et al.，2019）采集自一个供英语非母语者使用的辅助写作平台。第二，大多数汉语学习者语料库仅由一位标注员提供一种标注结果。这种单一的修改结果，极易出现 GEC 模型修改正确但与答案不匹配的现象，从而导致模型学习困难，评测结果不够准确。无论是在真实的语言教学中还是在学习者文本的偏误标注时，一个偏误句子都有可能存在多种不同的修改方案。尤其在 GEC 任务上，Tetreault & Chodorow（2008）在研究中指出，提供同一句子的多种偏误修改方案，可以显著提升纠错模型的效果。目前，英语的面向纠错任务的学习者语料库或数据集大多由多位标注员同时标注，如学习者语料库 NUCLE（Dahlmeier et al.，2013）有两位标注员，JFLEG（JHU Fluency-Extended GUG corpus）（Napoles et al.，2017）和 W&I 的测试集（Bryant et al.，2019）均有五位标注员。第三，现有的汉语学习者语料库和数据集基本都是采用最小改动的标注方式，缺乏针对流利性的偏误纠正结果。Napoles 等（2017）在构建英文 GEC 评测数据集 JFLEG 时提到，已有的最小改动方式标注仅仅修改了原句中的语法错误，而忽略了将其改为更流利、更符合母语者习惯的表达。因此，他们提出了流利提升的偏误标注方式，在忠于原意的基础上允许较大幅度的修改，以得到更为地道的标注结果。

从上述三个问题出发，本研究选取话题更为丰富的语言使用场景中的汉语学习者文本，设计一套多维度富信息标注体系，招募 183 名国际中文教育相关专业的标注员，采用众包标注策略，构建了一个包括最小改动和流利提升两种纠错维度的大规模、高质量、篇章级别的汉语学习者文本多维标注语料库 YACLC（Yet Another Chinese Learner Corpus）。

二　建库目标、理论方法和语料采集

语料库建设皆是目标驱动的，不同的目标决定不同的建库原则、方法和技术。YACLC 在设计之初立意为：既可以为汉语教学和二语习得研究提供数据支持和检索服务，也可作为语法自动纠错算法的训练与评测数据，服务于智能辅助写作技术研究。因此，在语料选取上，要考虑语料覆盖的全面性，既包括学习者的自身属性，如母语背景、学习等级等，又包括文本产出的场景、领域、话题等等，便于从不同维度进行学习者语言的对比分析。在语料标注上，要考虑如何设计合理的标注体系，保证语料标注的高质量；在满足二语习得研究多元、多维的研究范式的同时，解决应用于智能辅助写作技术研究中前述的三个现存问题。基于此，YACLC 的建库目标是一个大规模、学习者背景全面、语言场景开放、话题丰富、标注维度多样、结果丰富、质量优异的汉语学习者语料库。

YACLC 以二语习得理论中的中介语对比分析（Contrastive Interlanguage Analysis，简称 CIA）为建库的理论和方法指导。CIA 是使用学习者语料库进行研究的常见方法，该方法主张从多个角度在学习者语言和本族语之间展开比较分析（Grange，1996）。中介语多元对比分析（Grange，2015）是该方法的进一步发展，突出强调中介语和参照语（Reference Language）的变异性，以及中介语和参照语的各变体之间的多元互动对比特征（曹贤文，2020）。YACLC 的建设借鉴中介语多元对比分析，为汉语学习者语言和标注后的参照语设计语言功能、地域、学习者等多种角度的变体，进而在学习者语言和标注参照语以及各变体之间，从词汇、语法、语篇等语言层面进行多元对比分析研究，全面提供学习者语言发展的信息和规律。YACLC 支持采用综合测量框架对学习者语言进行多维表现分析，反映不同群体的汉语学习者的共同特点和变异特征。这些规律和特征一方面可“反馈到教学大纲的制定、教材的编写以及课堂教学实践等环节中”（曹贤文，2013），加强汉语教学的针对性和有效性；另一方面可以应用到自动判错、纠错、校对、润色等辅助写作技术的模型建构和训练，增强模型性能。

在收集汉语学习者文本环节，我们聚焦于开放场景下学习者的语言使用，采集了 Lang-8 平台[②]的汉语学习者作文语料。Lang-8 平台汇集了各个背景的二语学习者用目的语进行写作的文本，由精通该语言的母语者进行修改和批注。平台通过这样的“语言交换”鼓励机制来促进多语言交流（Mizumoto et al.，2011），从而汇集了一大批话题开放、内容丰富的学习者语料。Lang-8 拥有约 50,000 名汉语学习者用户，产出的汉语学习者语料库约为 29,595 篇文章，441,670 个句子（Zhao et al.，2014）。然而，由于这些语料是由学习者自发地互相修改，而没有相应的规范要求，修改后的数据质量参差不齐，甚至存在修改后的句子仍存在错误的情况。因此，我们仅提取了汉语学习者的作文原文为生语料，经语料清洗和过滤，最终形成 2,421 篇、32,124 个句子的待标注语料。

三　多维标注体系和实践

针对汉语学习者语料库的标注结果单一、流利度维度欠缺的问题，我们聚焦于汉语学习者语言的独特性，设计了一套以中介语多元对比分析理论和方法为指引、以计算机智能辅助写作技术为导向的多维度富信息标注体系。该体系包括以下四个维度：最小改动偏误纠正；流利度提升偏误纠正，即修正错误的同时使语句更符合汉语表达习惯；句子可接受度评分；句子修正与上下文的依赖关系。多维信息的标注互相依赖、互相限制，有助于控制标注质量，较全面地反映汉语学习者的语言使用特征。最小改动和流利提升维度可视为与学习者语言对照的不同功能的参照语变体，通过多元对比分析揭示二语习得的规律。

（一）标注原则

标注原则是制定标注规范的前提，与标注目的密切相关，对标注的内容与方法有重要制约作用（张宝林，2013）。我们确立了以下两条标注原则，对标注任务进行总体性指导。

第一，粒度为词。词是最小的能独立运用的语言单位（黄伯荣、廖旭东，1991），也是从汉语信息处理需要出发，适用于汉语信息处理使用的、具有确定的语义或语法功能的基本单位。因此，我们将词作为标注的基本单位，要求标注员在词层面上进行标注的相关操作。

第二，忠于原意。周小兵等（2007）提到标注学习者语料的首要原则即是“准确表达原句作者的意思；不能出于最小改动，改变了作者的意思”。

在标注原则的指导下，标注员从最小改动和流利提升两个维度对每个句子标注词级别的偏误，对句子可接受度维度标注评分，从上下文依赖性维度标注强弱。

（二）偏误类型

早期汉语学习者语料库的建设主要面向汉语教学及相关研究，越精细的分类，越有助于二语习得研究或学习者语言的偏误分析。因此，这些语料库大多采用包含字、词、句、篇等多个层面的复杂的偏误分类方法，以及基于这些分类的标注规范和标记集，如 HSK 从字、词、句、篇、标点符号等角度分类，约 50 种偏误标记（张宝林，2009）。然而，过于精细的标记系统不利于标注员做出统一判断，从而增加标注难度，降低标注效率。TOCFL 在对汉语学习者文本进行标注时率先使用更为简单的偏误八大分类：词汇 [L]、语法 [G]、形式 [F]、语序 [W]、语义 [S]、冗词 [R]、缺词 [M]、话题 [T]（Chang，2013）。后续的语法错误自动检测和纠正的评测比赛数据集就仅将偏误分为词语赘余（redundant words，R）、词语遗漏（missing words，M）、词语误用（word selection，S）、词序错序（word ordering errors，W）四类，大大简化了偏误标注的难度，更有助于训练 GEC 模型。

同样地，英语学习者语料库对偏误类型的标注也呈现出由繁到简的趋势。FCE[③]（Cambridge Learner Corpus First Certificate in English）（Yannakoudakis et al.，2011）有大约 80 种偏误类型，NUCLE（Dahlmeier et al.，2013）有 27 种。JFLEG（Napoles et al.，2017）舍弃了复杂的偏误分类和标记，只分为拼写、语法和不流利三种类型。BEA2019 发布的 W&I 数据集（Bryant et al.，2019）更是未曾标注偏误类型，仅有最小改动的结果，并使用工具 ERRANT（Bryant et al.，2017）自动标注错误类型。

鲁健骥（1994）认为，对外汉语教学的语法项目都会有“遗漏、误加、误代、错序”的偏误。以该观点为参照，考虑到英语和汉语已有学习者语料库的偏误分类，我们将偏误类型确定为成分缺失、成分冗余、词汇误用、语序错误四类。针对这四种偏误类型，在标注实践中，标注员分别通过“添加”“删除”“修改”“调序”这四种操作进行修正。

表 1　句子偏误标注示例

原句	偏误类型	对应修改
我喜欢坐茶馆看书。	成分缺失	我喜欢坐在茶馆里看书。（添加）
我的最好的朋友是安妮。	成分冗余	我最好的朋友是安妮。（删除）
学习的时间很小。	词汇误用	学习的时间很少。（修改）
武汉大学以好风景著称。	语序错误	武汉大学以风景好著称。（调序）

（三）最小改动维度

最小改动（Minimal Edit，M）偏误纠正的标注维度，要求标注员尽可能少地改动原句中的成分，使句子符合汉语语法规则。Napoles 等（2017）通过总结众多英语学习者语料库的标注，概括出“最小改动”原则，即在最小的范围（一般为 1~2 个词）内对句子进行改动，以达到直接改正语法错误的目的。在汉语学习者语料库标注中，周小兵等（2007）同样提出“最简化”原则，要求尽可能好地维持原句的结构，尽可能少地增删、替换句中的词语，使句子符合汉语语法规则。张宝林（2013）也强调过“在修改标注环节需最大限度地保持其语料的原汁原味”。最小改动的偏误纠正结果可看作是符合语法要求下的一种参照语变体。

（四）流利提升维度

流利提升（Fluency Edit，F）偏误纠正的标注维度，进一步要求将句子修改得更地道，符合汉语母语者的表达习惯。最小改动和流利提升这两个维度的标注具有逐步递进的关系。对同一个带有偏误的句子而言，一位标注员至少应先提供一条最小改动标注，再进行流利提升的标注。若标注员认为所提供的最小改动标注已经符合汉语的表达习惯，则可不再提供流利提升标注。若标注员认为原句无语法错误，则可不提供最小改动标注，只提供流利标注。

表 2 是最小改动与流利提升标注的示例。相较于最小改动，流利提升的偏误纠正结果可看作是符合汉语表达习惯的要求下的另一种参照语变体。这两种变体与学习者语言三者之间可从多个语言层面进行对比分析，揭示不同需求下学习者语言的共同和变异特征。

表 2　最小改动标注和流利提升标注示例

原句	他的名字是王。
最小改动标注	他的姓是王。
流利提升标注	他姓王。

（五）可接受度评分维度

语言学上，一个句子的可接受度（Acceptability）由母语者的内省判断决定（Warstadt et al.，2019）。Leech（1993）探讨过标注的主观性，他认为任何一种标注方案都可能产生标注的分歧，因为标注本质上是对语言特征的解释，不同的人可能会产生不同的解释。这种主观性在学习者语料的标注中普遍存在，不同的标注员会给出不同的修改方案（张宝林，2013）。因此，我们可以把句子可接受度作为标注的主观性体现，用可接受度评分量化标注主观性，保留标注员的行为信息。句子可接受度评分标注一方面与最小改动和流利提升标注呼应，当最小改动与流利提升标注不一致时，对应不同的可接受度；另一方面，后续可对标注员本身进行建模，消除标注偏差（Geva et al.，2019），更好地服务于提升 GEC 算法性能。

综合考虑，我们从句意、语法和流利度三个方面把句子分为 4 个等级，评分规则如表 3 所示。如果一个句子句意明确无歧义，语法无误，表达流利、自然，则可标注为 4 分；如果该句句意明确，语法无误，但表达不够流利、自然，则也可标注为 3 分；如果该句句意明确，但有语法偏误存在，则标注为 2 分；如果该句连句意都不明确，标注为 1 分。标注员根据上述原则对每个句子打分，若标为 4 和 3 分，则无需提供最小改动标注而只需提供流利提升标注；若为 2 和 1 分，则至少提供一个最小改动标注，再提供一个或多个流利提升标注。

表 3　可接受度评分标注示例

句意	语法	流利	可接受度评分	示例
√	√	√	4	明年我想去中国学汉语。
√	√	×	3	九月我宝贝出生了。
√	×	×	2	前天我尝一尝酸辣汤了。
×	×	×	1	在图书馆借书以后，总是靠近着照片的店。

（六）上下文依赖性维度

有些句子的偏误信息需要依靠上下文确定，据此我们设计了上下文依赖性的标注维度，要求标注员偏误纠正时考虑篇章信息。上下文依赖性分为强弱两种情况。

1. 强依赖。指原句的偏误信息不明确，需要依赖上下文才能确认和纠正偏误。如“他不能说话”。脱离原文来看符合汉语语法和语义表达。但在原文中该句位于“他的儿子现在 1 岁”之后。显然，句子中“他”的指代有歧义，且根据上下文信息，1 岁的孩子应该是“不会”说话，而不是“不能”说话。因此该句的偏误信息属于上下文强依赖情况。

2. 弱依赖。原句偏误信息明确，上下文信息对纠正的帮助不大，如“她写汉语写地非常流利”。标注员无需查看上下文即可进行偏误纠正。

（七）基于众包的多人标注方案

在英文的数据集中，已有最多 5 位标注员同时标注的方案（Bryant et al.，2019）。而汉语仅有 NLPCC 2018 的测试集启用了 2 位标注员（Zhao et al.，2018）。但标注员之间并非独立进行标注的，而是在一位标注员完成标注后，由另一位标注员审核其标注，若有不同意见

再给出第二种修改结果。在 2000 条句子的测试集中，仅有 261 个句子有两个修改结果（Zhao et al.，2018）。

YACLC 的建设采用众包策略。我们搭建了可供多人同时在线标注和逐句审核的众包标注平台④，招募了 183 名国际中文教育相关专业的标注员，分组、分阶段地进行偏误标注和审核工作。同一组内人员标注同一批学习者文本，可以解决标注结果单一的问题。这种多人标注方案符合教学中同一偏误可能会有多种解释和修改方案的实际情况。多样化的修改结果既能为增强 GEC 模型容错能力的研究与更准确的评测提供数据基础，同时可建模标注员行为信息，引入到纠错或质量评估（Quality Estimation）（Chollampatt & Ng，2018；Liu et al.，2021）模型，对多个纠错结果进行重排序，消除标注偏差（Geva et al.，2019），进一步提升纠错效果。

图 1 和图 2 分别是标注平台的句子标注和审核界面，标注员首先需要点击原句后跟的星级标识来评判句子可接受度，一星到四星分别对应 1 至 4 分；然后进行最小改动和流利提升维度的标注，分别以 G 和 F 标识，以词为粒度，通过添加、删除、修改、调序四种操作将原句修改为正确句子；点击每条标注后的“眼睛标识”标注上下文依赖性。当鼠标悬停在原句文本上时，标注员即可查看前后两句的上下文信息。审核员通过点击“通过”或者“驳回”按钮审核标注后的句子，亦可提供审核意见供标注员参考。

图 1　数据标注界面

图 2　逐句审核界面

四　数据集分析

目前，YACLC 包含 2421 篇学习者文本共 32,124 个句子的多维标注，每个句子的标注员数量维持在 9~11 人。根据标注维度，又可分为最小改动标注子库、流利提升标注子库、

句子可接受度子库。图 3 示例了一条原句的最小改动和流利提升标注结果。表 4 给出了标注统计数据。我们使用斯坦福 CoreNLP（Manning et al.，2014）进行了分词操作。

```
ntence_id": 4308, // 句子id
ntence_text": "我只可以是坐飞机去的，因为巴西离英国到远极了。", // 学习者f
ticle_id": 7267, // 该句所属的作文id
ticle_name": "我放假的打算", // 作文标题
tal_annotators": 10, // 共多少个标注者参与了该句的标注
ntence_annos": [ // 多标注信息
{ // 最小改动偏误纠正
    "sentence_score": 2, // 句子可接受度评分
    "is_grammatical": 1, // 标注类型：1表示最小改动，0表示流利提升
    "depend_context": 1, // 上下文依赖性：1表示弱依赖，0表示强依赖
    "correction": "我只能坐飞机去，因为巴西离英国远极了。", // 标注结果
    "edits_count": 3, // 共有几处修改
    "annotator_count": 6 // 共有几个标注者标注为这一结果
},
{ // 最小改动偏误纠正
    "sentence_score": 2,
    "is_grammatical": 1,
    "depend_context": 1,
    "correction": "我只能是坐飞机去的，因为巴西离英国远极了。",
    "edits_count": 2,
    "annotator_count": 1
},
{ // 流利提升偏误纠正
    "sentence_score": 2,
    "is_grammatical": 0,
    "depend_context": 1,
    "correction": "我只能坐飞机去，因为巴西离英国太远了。",
    "edits_count": 6,
    "annotator_count": 2
}
```

图 3　标注数据示例

表 4　YACLC 数据统计

	句子数量	词次	字次	词种	句均词次	句均字次	句均修改数
原句	32,124	373,749	577,679	23,374	11.63	17.99	
最小改动	331,292	3,931,509	6,077,742	33,042	11.87	18.35	2.68
流利提升	137,708	1,405,090	2,172,490	24,240	10.20	15.78	3.07

根据表 4 的统计结果，最小改动子库共包括 32,124 个原句的 331,292 个标注，平均每个句子有 10.31 个最小改动标注；流利提升子库共 137,708 个，平均每个句子有 4.29 个流利提升标注。这符合 YACLC 的建设预期：多个标注维度下提供丰富多样的标注结果。从标注数量来看，每个原句的平均流利提升标注的数量较平均最小改动标注要少。从词种数来看，流利提升标注较原句有部分提升，而最小改动标注则有大幅提升，说明标注员在最小改动标注时引入了更多词语。但从句长来看，最小改动标注较原句差别不大，但流利提升标注则较原句减少了 1.4 个词。从标注的句均修改数来看，流利提升标注的修改较最小改动标注多了约 0.5 处。这一统计结果符合“流利提升标注较最小改动标注的修改幅度更大”的设计原则（Napoles et al.，2017），有利于后续多元对比分析学习者语言和这两种参照语变体。

基于众包的多人标注方案以及最小改动和流利提升两种维度的偏误纠正维度，使得 YACLC 解决了以往汉语学习者语料库标注结果单一和流利度欠缺的问题。YACLC 标注结果的丰富性可从原句的平均标注数量体现，多样性则可进一步结合标注员的行为信息进行分析。我们将每个句子的标注结果进行去重，两个维度下的句均标注结果数量就分别减少为 5.74 和 1.79。我们统计了每个不重复标注结果的标注员人数，如图 4 所示，其中横坐标

为标注员人数的不同取值，纵坐标为各取值所对应的不重复标注结果的总数量。可以看到即使 YACLC 设计由 11 位标注员同时标注一个原句，无论是最小改动还是流利提升维度，绝大多数的标注结果都是仅由一名标注员提供，不同的标注员偏向于给出不同的修改方案。这是从标注结果的整个句子是否相同的统计，即句子级别的一致性分析。

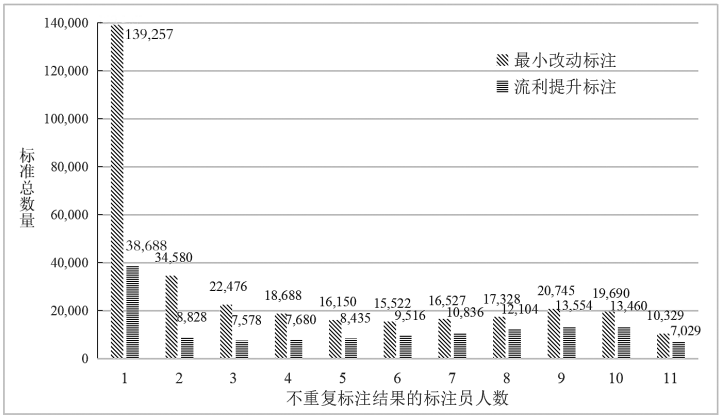

图 4　标注结果的标注人员分布图

鉴于偏误纠正的特殊性，可对各标注结果的每处修改进行统计，即进行词级别的一致性分析。我们使用 Krippendorff（2013）提出的 Alpha 系数计算最小改动和流利提升两个维度下每个原句的所有标注结果之间词级别的一致性。结果如图 5 所示，其中横坐标为每个原句的标注结果一致性分数的取值区间，纵坐标为各区间所对应的原句的数量。一般认为，$\alpha \geq 0.8$ 时，数据具有较强的一致性。最小改动标注的平均一致性为 0.45，流利标注的平均一致性为 0.59，皆低于阈值。也就是说，无论是最小改动标注还是流利提升标注，其标注一致性都偏低。这说明对同一句子进行标注的大部分标注员在词级别的修改上也未能达成一致，标注的差异较大，进一步证实了 YACLC 标注结果的多样性。

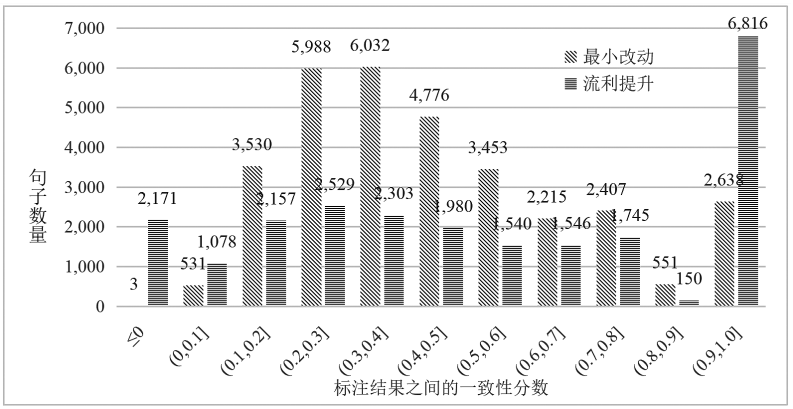

图 5　两个维度标注结果的一致性分布直方图

同时，图 5 的数据表明，流利提升标注的词级别平均一致性高于最小改动标注。在最小改动标注中，有 2,443 个句子的 Alpha 系数高于阈值，具有较强的标注一致性，仅占比约 7.60%。而在流利提升标注中，有 6,966 个句子的标注一致性较强，占比达 29.01%。就峰值而言，最小改动标注在一致性分数为（0.3, 0.4] 时达到峰值，而流利提升维度则是在（0.9, 1.0] 时。一致性分数为 1.0 意味着标注员们提供同一种修改结果。但在流利提升维度，6,553 个

得分为 1.0 的句子是标注员们都未对原句进行修改。这也反映出标注员在两个维度上的行为差异，我们统计每个原句里每位标注员提供的标注数量，结果如图 6 所示。其中横坐标为每个标注员在两个维度下对每个原句提供的有修改的标注数量的取值，“0”表示标注员对原句无修改，纵坐标为各取值所对应的所有原句的标注总数量。可以看到绝大多数标注员都会只提供一个最小改动标注，也只提供一个流利提升标注。同时，标注员们流利标注的数量较少，且大多是无修改标注。这主要是由于我们并未强制要求标注员提供流利标注。因而在任务模式下，标注员们更倾向于标注更多原句而非提供更多样的标注结果。如何鼓励标注员积极提供改写式的流利标注，需要进一步思考和探索。

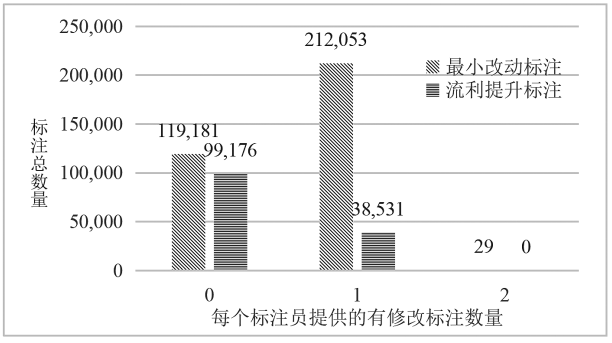

图 6　标注员提供的有修改标注数量分布图

五　结语

本研究聚焦于汉语学习者语言的独特性，针对目前汉语学习者语料库建构中存在的语料来源封闭、标注结果单一和流利度维度欠缺的问题，设计了一套以二语习得研究和智能辅助写作理论和方法为指导的多维度标注体系，构建了一个大规模、高质量、篇章级别的汉语学习者文本多维标注语料库 YACLC。具体工作可总结如下：

（1）针对语料来源集中于考试场景的问题，本研究选取了辅助写作平台 Lang-8 的语言使用场景下的学习者作文文本。

（2）针对偏误标注结果单一的问题，本研究采用基于众包的多人标注方案，获得句均 14.60 个偏误标注结果，标注结果更加丰富、多样，为偏误自动发现与修改技术、评测提供了较为新颖的数据集。

（3）针对欠缺流利提升维度标注的问题，本研究设计了两个维度的标注：既能提供多种仅纠正不符合汉语语法规则的最小改动标注，也能从流利提升的维度提供更符合母语表达习惯的修改结果，适用于多种需求下的辅助写作技术的研发。

（4）本研究设计了句子可接受度评分、上下文依赖性等标注维度，既有助于提升标注质量，又保留标注员的主观性信息，可进一步用于标注行为建模，提升数据的应用效益。

（5）本研究使用自建的众包标注平台，构建完成一个包括 2,421 篇、32,124 句及 331,292 个最小改动标注和 137,708 个流利标注的篇章级别的多维标注语料库。

YACLC 既可直接作为数据资源供自动判错、纠错、校对、润色等辅助写作技术使用，也可为汉语教学和二语习得研究提供例证数据支持，同时为进一步开展中介语多元对比分析研究提供参考。具体而言，对照中介语多元对比分析模型，YACLC 的最小改动维度和流利提升维度可视为与学习者语言对应的不同功能变量下的参照语变体。因此，基于 YACLC 的研究可将多元对比的角度从学习者语言和目标语之间或之内的两两对比，拓展

到学习者语言与所对应的不同修改维度的两个变体这三者之间的多元互动对比，从理论上拓宽了二语习得理论的研究视野，从实践上丰富了利用学习者语料库的多元对比分析方法的研究路径。我们在语料采集时也获取到了学习者元信息，因此可使用 YACLC 从学习者变体的角度对比分析不同群体的学习者语言的共同特点和变异特征，包括不同母语背景、学习水平、写作题材等多个学习者个体因素变量。同时可结合中介语多维表现分析方法（曹贤文，2020），从准确性、流利性、复杂性和多样性多个维度，从汉字、词汇、语法和语篇等多个层面，综合测量、对比分析不同群体的学习者语言，揭示二语习得的发展特征和规律。

［附　注］

① 学习者语料库即中介语语料库，学习者语言即中介语，本研究对两种表述不作区分。

② Lang-8 平台网址为：https://lang-8.com/。

③ FCE 数据集是 CLC：（Nicholls，2003）的一部分。

④ YACLC 标注平台网址为：https://yaclc.wenmind.net/。

［参考文献］

［1］曹贤文 . 应用语言学实证研究方法与量化数据分析——对外汉语教学研究视角［M］. 北京：世界图书北京出版公司，2013.

［2］曹贤文 . 二语习得研究“需求侧”视角下的汉语学习者语料库建设［J］. 华文教学与研究，2020，（1）.

［3］黄伯荣，廖序东 . 现代汉语（增订版）［M］. 北京：高等教育出版社，1991.

［4］鲁健骥 . 外国人学汉语的语法偏误分析［J］. 语言教学与研究，1994，（1）.

［5］王辰成，杨麟儿，王莹莹，杜永萍，杨尔弘 . 基于 Transformer 增强架构的中文语法纠错方法［J］. 中文信息学报，2020，（6）.

［6］张宝林 . “HSK 动态作文语料库”的特色与功能［J］. 汉语国际教育，2009，（4）.

［7］张宝林 . 汉语中介语语料库建设的现状与对策［J］. 语言文字应用，2010，（3）.

［8］张宝林 . 关于通用型汉语中介语语料库标注模式的再认识［J］. 世界汉语教学，2013，（1）.

［9］张宝林，崔希亮 .“全球汉语中介语语料库建设和研究”的设计理念［J］. 语言教学与研究，2013，（5）.

［10］张　舸 . 程度副词结构作状语、谓语和补语的语义及句法差异［A］.“第二届中青年学者汉语教学国际学术研讨会”资料汇编［C］，2008.

［11］周文华，肖奚强 . 基于语料库的外国学生兼语句习得研究［J］. 暨南大学华文学院学报，2009，（3）.

［12］周小兵，朱其智，邓小宁等 . 外国人学汉语语法偏误研究［M］. 北京：北京语言大学出版社，2007.

［13］Bryant, C., Felice, M., Briscoe, E. Automatic annotation and evaluation of error types for grammatical error correction［A］. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.

［14］Bryant, C., Felice, M., Andersen, Ø. E., et al. The BEA-2019 shared task on grammatical error correction［A］. Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications, 2019.

［15］Chang, L. P. TOCFL 作文语料库的建置与应用［A］. 第二届汉语中介语语料库建设与应用国际学术讨论会论文选集［C］，2013.

[16] Chollampatt, S. & Ng, H. T. Neural Quality Estimation of Grammatical Error Correction[A]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[17] Dahlmeier, D., Ng, H. T., & Wu, S. M. Building a large annotated corpus of learner English: The NUS corpus of learner English[A]. Proceedings of the eighth workshop on innovative use of NLP for building educational applications, 2013.

[18] Dale, R., & Kilgarriff, A. Helping our own: The HOO 2011 pilot shared task[A]. Proceedings of the 13th European Workshop on Natural Language Generation, 2011.

[19] Dale, R., Anisimoff, I., & Narroway, G. HOO 2012: A report on the preposition and determiner error correction shared task[A]. Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, 2012.

[20] Geva, M., Goldberg, Y., & Berant, J. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets[A]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019.

[21] Granger, S. From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora[A]. *Languages in Contrast. Text-based cross-linguistic studies*[C]. Lund: Lund University Press, 1996.

[22] Granger, S. The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research[J]. *Tesol Quarterly*, 2003, （3）.

[23] Granger, S. Computer learner corpus research: Current status and future prospects[J]. *Applied Corpus Linguistics*, 2004, （2）.

[24] Granger, S. Contrastive interlanguage analysis: A reappraisal[J]. *International Journal of Learner Corpus Research*, 2015, （1）.

[25] Htut, P. M., & Tetreault, J. The Unbearable Weight of Generating Artificial Errors for Grammatical Error Correction[A]. Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2019.

[26] Krippendorff, K. *Content analysis: An introduction to its methodology*[M]. Thousand Oaks, CA: Sage, 2013.

[27] Lee, L. H., Yu, L. C., & Chang, L. P. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis[A]. Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications. 2015.

[28] Lee, L. H., Rao, G., Yu, L. C., Xun, E., Zhang, B., & Chang, L. P. Overview of the NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis[A]. Proceedings of the Workshop on Natural Language Processing Techniques for Educational Applications （NLP-TEA’16）, 2016.

[29] Leech, G. Corpus annotation schemes[J]. *Literary and linguistic computing*, 1993, （4）.

[30] Liu, Z., Yi, X., Sun, M., Yang, L., & Chua, T. S. Neural Quality Estimation with Multiple Hypotheses for Grammatical Error Correction[A]. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

[31] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. The Stanford Core NLP Natural Language Processing Toolkit[A]. Proceedings of 52nd Annual Meeting of the

Association for Computational Linguistics: System Demonstrations, 2014.

[32] Milton, J. Elements of a written interlanguage: A computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students[R], 2001.

[33] Mizumoto, T., Komachi, M., Nagata, M., & Matsumoto, Y. Mining revision log of language learning SNS for automated Japanese error correction of second language learners[A]. Proceedings of 5th International Joint Conference on Natural Language Processing, 2011.

[34] Napoles, C., Sakaguchi, K., & Tetreault, J. JFLEG: A fluency corpus and benchmark for grammatical error correction[A]. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017.

[35] Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., Tetreault J. The CoNLL-2013 shared task on grammatical error correction[A]. Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 2013.

[36] Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. The CoNLL-2014 shared task on grammatical error correction[A]. Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 2014.

[37] Nicholls, D. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT[A]. Proceedings of the Corpus Linguistics 2003 conference, 2003.

[38] Rao, G., Zhang, B., Xun, E., & Lee, L. H. IJCNLP-2017 task 1: Chinese grammatical error diagnosis[A]. Proceedings of the IJCNLP 2017 Shared Tasks, 2017.

[39] Rao, G., Gong, Q., Zhang, B., & Xun, E. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis[A]. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, 2018.

[40] Rao, G., Yang, E., & Zhang, B. Overview of NLPTEA-2020 Shared Task for Chinese Grammatical Error Diagnosis[A]. Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, 2020.

[41] Tetreault, J., & Chodorow, M. Native judgments of non-native usage: Experiments in preposition error detection[A]. Proceedings of the workshop on human judgements in computational linguistics, 2008.

[42] Warstadt, A., Singh, A., & Bowman, S. R. Neural Network Acceptability Judgments[A]. Transactions of the Association for Computational Linguistics, 2019.

[43] Yannakoudakis, H., Briscoe, T., & Medlock, B. A new dataset and method for automatically grading ESOL texts[A]. Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011.

[44] Yu, L. C., Lee, L. H., & Chang, L. P. Overview of grammatical error diagnosis for learning Chinese as a foreign language[A]. Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications, 2014.

[45] Zhao, Y., Jiang, N., Sun, W., & Wan, X. Overview of the NLPCC 2018 shared task: Grammatical error correction[A]. CCF International Conference on Natural Language Processing and Chinese Computing, 2018.

（责任编辑　陈丽湘）