

医学论文跨语种抄袭的特征分析和防范措施*

张 和 张海燕 鲁翠涛 丁敏娇[†]

温州医科大学期刊社, 325035, 浙江温州

摘 要 医学论文跨语种抄袭多为英译中文献,目前还没有学术不端检测系统进行有效识别,因此甄别十分困难。本文从语言翻译学和文献内容 2 个角度,总结英译中的跨语种抄袭特征,剖析编辑部出版实践中的典型案例,提出在出版流程中严格把关并发挥主观能动性,开发基于先进计算机技术跨语种抄袭检测软件以及加强学术诚信教育和惩处力度 3 个防范措施,为编辑同行提供借鉴,共同维护学术诚信。

关键词 医学论文;学术不端;跨语种抄袭;防范措施

Feature analysis and preventive measures for cross-lingual plagiarism in medical papers//ZHANG He, ZHANG Haiyan, LU Cuitao, DING Minjiao

Abstract Cross-lingual plagiarism in medical papers is mostly English to Chinese translated literature. Current academic misconduct detection system cannot judge the cross-lingual plagiarism, therefore it is difficult to identify. This paper summarizes the features of English to Chinese lingual plagiarism in aspects of language translation and literature content, and analyzes the classic cases of editorial board. We put three preventive measures to maintain a good academic integrity: close checking and subjective initiative in the publishing process, advanced computer technology-based cross-lingual plagiarism detection software and strengthening academic integrity education and punishment.

Keywords medical papers; academic misconduct; cross-lingual plagiarism; preventive measures

Authors' address Periodical Press of Wenzhou Medical University, 325035, Wenzhou, Zhejiang, China

DOI:10.16811/j.cnki.1001-4314.2023.02.010

21 世纪科学技术迅猛发展,谷歌、百度、有道和腾讯等开发的在线机器翻译系统,能提供实时、免费的翻译,满足绝大多数用户的翻译需求。其中百度翻译采用创新人机共译模式即机器翻译(machine translation, MT) + 译后编辑(post-editing, PE) + 智能增强,让机器翻译内容更准确、更高效。另外随着日趋频繁的国际学术交流,以及互联网上丰富的外文数据库资源,读者获取前沿信息和外文文献方便快捷。因此,在翻译技术和信息文献资源的支持下,开始出现跨语种抄袭,通过翻译外文学术文献进行再次发表,其中以翻译英语文献居多。而目前这类翻译文献还没有学术不端检测系统能进行有效判断,甄别十分困难,且国内学者对这

方面的研究较少。因此,本文通过总结医学论文英译中的跨语种抄袭特征,并理论联系实际,剖析笔者所在《肝胆胰外科杂志》编辑部出版实践中的案例,为编辑防范此类学术不端行为提供借鉴,共同维护学术诚信。

1 定义

跨语种抄袭,是指未经版权许可,未透明原文出处,或未明示二次文献发表的前提下,将已发表的某一语种文献翻译成其他语种,进行再次发表的行为^[1]。根据 CY/T 174—2019《学术出版规范 期刊学术不端行为界定》规定,“未经许可,使用需要获得许可的版权文献和被允许的二次发表不说明首次发表出处”均属于学术不端行为^[2]。

2 国内文献数据库的学术不端检测系统分析

目前国内各大文献数据库开发了学术不端系统,包括万方论文相似性检测系统(WF similarity detection, WFSD)、中国知网科技期刊学术不端文献检测系统(academic misconduct literature check, AMLC)、维普论文检测系统(VIP paper check system, VPCS)和超星大雅论文检测系统(简称超星大雅)等。国内科技期刊在稿件初审时多采用单个或多个系统联合进行检测。超星大雅优势在于图书检测,因此不纳入统计分析。WFSD、AMLC 和 VPCS 三大文字相似性检测系统的异同点总结如下。

1) 比对资源库,3 个检测系统都有丰富的比对资源,包括数据库各自收录的学术期刊、优先出版论文、重要会议论文、硕博学位论文、中国专利文献全文、英文数据库等资源。针对重要会议论文资源,WFSD 包括国内外会议论文,AMLC 和 VPCS 仅包括国内会议论文。AMLC 和 VPCS 有图书资源和港澳台学术文献库,但 WFSD 没有;WFSD 和 VPCS 有中国标准全文数据库,但 AMLC 没有。英文数据库以 AMLC 收录资源较多,涵盖期刊、硕博论文、会议的英文数据以及外文 Springer、Taylor&Francis 期刊数据库等;WFSD 和 VPCS 的英文比对资源库为外文特色文献数据,但未给出具体数据库名称。

2) 英文检测功能,三者均有此项功能,但只能与该数据库收录的英文文献资源进行比对。其中 AMLC

* 浙江省教育厅一般科研项目(Y202249697)

[†] 通信作者

包含多语种检测功能,可以检测中英文互译文献;WFSD和VPCS的英文检测功能只能针对英文稿件与数据库中的英文文献,不支持检测中英互译文献。

3 医学论文英译中跨语种抄袭的特征分析

从语言翻译学和文献内容2个角度,笔者结合编辑出版实践经验,并阅读相关文献和医学英汉互译书籍,总结医学论文英译中跨语种抄袭文献的特征。

3.1 语言翻译学特征

跨语种抄袭稿件多采用机器翻译后作者稍加修饰,虽然部分机器翻译采用创新人机共译模式来提高翻译的准确性,但由于自然语言复杂多变,英汉思维方式的差异,以及医学论文有较多专业术语,因此英译中文献多存在不自然、不连贯,不符合中文语言表达习惯,违背专业领域常识性知识等问题。《医学英汉互译技巧与实践》是一本十分实用的工具书,此书从翻译理论、翻译技巧和翻译实践等多方面进行阐述^[3]。张重毅等^[4]和袁松翔等^[5]分别从语言表达习惯和译文风格方面分析了跨语种抄袭特征。本文从语言翻译学角度出发,总结英译中跨语种抄袭存在符号错误、字词错误、句法错误、逻辑错误4个方面的特征。

3.1.1 符号错误

标点符号如半角、全角不分,顿号和逗号之间的误用,连词“和”之前使用逗号;变量符号前后不一。

3.1.2 字词错误

1) 错别字:文中出现“的、地、得”3字的混用。

2) 人名、地名和外语音译错误:医学论文中有很多以外国科学家名字命名的疾病名称和统计学方法,多采用中外文结合的方式,有自己固定的翻译习惯,如“Parkinson's syndrome”通常被译为“帕金森综合征”,而“Fisher test”多不译为“费歇尔检验”。

3) 通用词语错误:尤其一词多义的英文单词更容易出现错译,“Bile duct failed to develop properly.”中“develop”正确翻译为“发育”,而非“发展”。

4) 术语错误:医学期刊存在较多专业术语,术语具有单义性、科学性、系统性和简明性等特征^[6]。若是不熟悉该领域的人员或者机器翻译,会出现翻译错误或不准确,产生歧义,甚至让人啼笑皆非。如“open-surgery”正确翻译为“开腹手术”,而非“开放手术”;“stone size”被错误翻译为“石头尺寸”,正确翻译为“结石大小”。

3.1.3 句法错误

由于中英思维习惯不同,英译中跨语种抄袭论文会出现较多不符合中文表达习惯的句子。

1) 被动句较多:跨语种抄袭论文含“被、使得、导

致、致使”等被动句出现频率高,而中文多使用主动语态。

2) 句子重心错误:英语往往重心在前,开篇点题,尤其是后置定语从句、后置状语从句较多,而中文却相反。

例:“Patients were converted to open surgery because of dense adhesions, unclear anatomy and large impacted stones.”

中文多按照因果逻辑顺序翻译为“患者因粘连致密、解剖不清、巨大结石嵌顿结石而转为开腹手术”,而不采用因果倒置。

3) 句子内容松散:医学英语有较多长难句,一个句子中通常包含若干个从句,关系复杂,需要仔细分清主干和各分句之间的关系。

例:“Mirizzi syndrome is a rare condition, which caused by the obstruction of the common bile duct or common hepatic duct by external compression from multiple impacted gallstones or a single large impacted gallstone in Hartman's pouch.”

该句中,主语是“Mirizzi syndrome”,谓语动词“is”,宾语“a rare condition”,“which”引导了一个非限制性定语从句。第1个“by”引导一个原因状语从句,第2个“by”引导一个定语从句,修饰“obstruction”。正确翻译为“Mirizzi综合征是一种罕见疾病,是由于单个或多个结石嵌顿于Hartman袋,压迫引起的胆总管或肝总管梗阻。”

3.1.4 逻辑错误

医学论文结构紧凑、逻辑清晰,若不能准确把握语义和语境,尤其是多篇文献翻译拼凑,会出现全文段落和章节语言风格迥异,语义跳跃,逻辑不连贯,导致内容表述混乱,不知所云。

3.2 文献内容特征

3.2.1 主题内容

医学跨语种抄袭文献的主题内容多为医学热点、焦点或冷门研究领域,比如涉及miRNA、lncRNA和circRNA的稿件就是跨语种抄袭重灾区;还有部分抄袭者深谙国内各大数据库的比对资源库,对国外会议论文收录较少,因此抄袭国际会议中最新的研究方法和成果。

3.2.2 图表特征

图表有较多英文表述,但无中文对照说明,祁丽娟等^[7]也认为这是跨语种抄袭的常见特征之一。

3.2.3 写作格式

医学跨语种抄袭文献的写作格式套用外文,比如结构式摘要用“背景”代替“目的”,正文单独列出“结

论”,其后包含“致谢”。

3.2.4 文末参考文献

文末参考文献以英文文献居多,且内容陈旧。正文中引用的观点与文末参考文献并不一致,可能是作者仅做简单翻译,替换参考文献,但并未认真查阅原文。

4 案例分析

笔者总结了本刊6篇明确存在跨语种抄袭稿件,并对其中3例典型案例进行深入剖析。6篇稿件全部重新进行文字相似性检测,详见表1,结果发现:

1)4篇稿件文字重复率均在20%以下;例3为发表后撤稿稿件,文字重复率较高为80%以上,主要集

中在与已发表中文文献的文字重复内容;例6的WFSD和AMLC文字重复率15%以下,但VPCS较高为35.34%,阅读比对报告发现,除中文文字有重复外,中文文献的英文摘要与已发表外文原文有重复。

2)AMLC多语种检测结果均为0,表明AMLC跨语言检测能力有待提高。

3)文末参考文献绝大多数为英文,这与跨语种抄袭文献内容特征的第4点相符。

4)利用文题、主题词、关键词等在学术搜索引擎和外文数据库进行检索,能找出83.3%(5/6)的英文原文,其中1篇未找到的原因可能是多篇外文文献拼凑的结果。

表1 6篇跨语种抄袭文献的检出情况

序号	中文文题	文字重复率/%			第一发现人员	文末英文/总参考文献数量	利用文题、主题词、关键词等是否检索到英文原文
		WFSD	AMLC ¹⁾	VPCS			
1	《选择性胆管造影对复杂性肝内胆管结石治疗的临床可行性与安全性》	6.37	2.5	12.23	外审专家	9/15	否
2	《丙泊酚与异氟醚对原发性肝癌手术后疼痛影响的对比研究》	6.67	0.5	6.46	外审专家	4/18	是
3	《lncRNA在肝细胞癌中的研究现状》	89.04	81.6	91.87	读者	53/54	是
4	《液体活检在胆道癌中的临床应用》	2.47	12.7	7.73	责任编辑	45/45	是
5	《胰腺脂肪浸润与胰腺癌关系的研究进展》	10.51	3.5	15.60	外审专家	47/49	是
6	《LCBDE治疗困难性胆管结石的安全性和有效性单中心1450例LCBDE+LC回顾分析》	13.65	6.7	35.34	外审专家	25/26	是

注:1)AMLC跨语种检测的文字重复率全部为0。

例1 2019年稿件《选择性胆管造影对复杂性肝内胆管结石治疗的临床可行性与安全性》在同行评议过程中,其中1位外审专家指出文中存在多处语言翻译学特征,比如“一个负担过重且由公共资金资助的医疗服务体系当中,很少有机构会拥有提供此类服务的资源”,这种表述只有国外医院才有;被动句较多;术语错误“结石尺寸”。由于并未找到相应的外文原文,只能判定为疑似跨语种抄袭文献。

分析:编辑仔细阅读中文文献,进一步发现,正文讨论中的部分内容与文章主题无关;6篇引用的中文参考文献与正文内容完全不相关。然后利用文题、主题词、关键词等在各大学术搜索引擎和数据库检索也未找到原文,推测有可能此文是多篇中外文文献拼凑的结果。最后与作者电话联系为第三方中介代写代投稿件,因此本文存在伪造、篡改、抄袭、代写代投等多种学术不端,予退稿处理。

例2 2018年来稿《丙泊酚与异氟醚对原发性肝癌手术后疼痛影响的对比研究》也是外审专家发现,指出这是一篇跨语种抄袭文献,并将原文发至编辑部。编辑部查阅发现,英文原文《Effect of intravenous

anesthesia with propofol versus isoflurane inhalation anesthesia in postoperative pain of inguinal herniotomy: a randomized clinical trial》已于2017年发表在《Medical Gas Research》,2篇文献的文题也很雷同,只是将疾病类型由腹股沟斜疝改成原发性肝癌,麻醉用药完全一样。正文资料、方法和结果完全照搬外文,只是涉及腹股沟斜疝的前言和讨论中的内容稍作改动,替换为原发性肝癌,文中3张表格内容完全一样,见图1。文末参考文献替换补充了部分中文文献。

分析:这是比较“高明”的抄袭者,虽然中文写作流畅,无明显语言翻译痕迹,但存在篡改疾病名称,张冠李戴,替换参考文献。如果编辑对来稿保持高度责任感和警惕性,还是能查出这些学术不端行为。编辑用“propofol”“isoflurane”“postoperative pain”3个关键词在Pubmed中检索可迅速找到英文原文。

例3 2019年文献综述《lncRNA在肝细胞癌中的研究现状》见刊后,有读者来信反映该文抄袭2018年发表于《EXCLI Journal》的《lncRNA involvement in hepatocellular carcinoma metastasis and prognosis》,经过人工比对,2篇内容雷同,中文文献只是在介绍各条参与

Table 1: The baseline measurements for comparing the age, heart rate, systolic and diastolic blood pressure, oxygen saturation, ETCO₂ and BIS before induction of anesthesia

Item	Isoflurane group	Propofol group	P
Age (year)	36.33±9.85	32.86±9.35	0.071
Heart rate (beats per minute)	90.1±16.8	87.6±15.4	0.438
Systolic blood pressure (mmHg)	128.4±23.9	125.9±125.9	0.576
Diastolic blood pressure (mmHg)	79.6±14.4	78.4±14.9	0.686
Oxygen saturation (%)	97.8±1.6	98.3±1.4	0.174
ETCO ₂ (%)	36.5±2.5	36.3±2	0.267
BIS (Score)	0.96±6.86	0.88±6.30	0.952

表1 麻醉诱导前两组患者生命体征的基线值[□]

变量 [□]	异氟醚组 [□]	丙泊酚组 [□]	P 值 [□]
年龄(岁) [□]	36.33±9.85 [□]	32.86±9.35 [□]	0.081 [□]
心率(每分钟) [□]	90.1±16.8 [□]	87.6±15.4 [□]	0.438 [□]
收缩压(mmHg) [□]	128.4±23.9 [□]	125.9±25.6 [□]	0.576 [□]
舒张压(mmHg) [□]	79.6±14.4 [□]	78.4±14.9 [□]	0.686 [□]
氧饱和度(%) [□]	97.8±1.6 [□]	98.3±1.4 [□]	0.174 [□]
呼气末二氧化碳浓度(%) [□]	36.5±2.5 [□]	36.3±2.2 [□]	0.267 [□]
双频指数 [□]	0.96±6.86 [□]	0.88±6.30 [□]	0.952 [□]

图1 跨语种抄袭文献的表格抄袭

肝细胞癌 lncRNA 的前后顺序上进行了调整,核实后联系各大数据库作撤稿处理。

分析:这篇文献符合跨语种抄袭文献内容特征的第一点,主题为医学领域的热点。但通篇用词流畅,符合中文语言习惯,并未发现语言翻译学特征,隐蔽性极强。该文为读者来信反映,因此审稿时编辑需要多留个心眼,仔细甄别。编辑用“lncRNA”“hepatocellularcarcinoma”“metastasis”“prognosis”4个关键词在 PubMed 中也能检索到英文原文。

5 措施

5.1 在出版流程中严格把关,发挥主观能动性

1)提高期刊编辑自身素养。为了规避学术不端检测系统,增强隐蔽性,跨语种抄袭稿件往往存在多种语言翻译学和跨语种抄袭文献内容特征,还有多种形式的学术不端行为。一些抄袭者首先会将机器翻译文献进行译后编辑,掩盖翻译痕迹,使语言表达流畅,符合中文习惯;然后对原文进行改头换面,替换前言背景和部分英文参考文献,篡改部分数据、图片、检测指标和疾病类型等。因此甄别跨语种抄袭十分困难,期刊编辑需要不断提高自身素养,拥有良好的英文阅读能力和医学专业知识储备量,时刻关注学科前沿动态,并在稿件处理过程中时刻保持警惕,层层把关,从源头杜绝学术不端行为。①在稿件初审时,通读全文并利用 word 中的“拼写和语法”功能进行检测,发现存在文字拗口,可读性欠佳,不符合中文语言习惯。如符合4个语言翻译特征,可初步判断为跨语种抄袭。②抓住跨语种抄袭文献的内容特征,并借助外部工具,常规使用多个学术不端检测系统进行查重,然后利用文题、主题词、关键词等在 Google 和百度等学术搜索引擎或 PubMed 和 Web of Science 等外文数据库检索,再进行人工比对,深入分析判别。③若高度怀疑存在英译中跨语种抄袭,应及时联系作者核实确认。

2)依靠同行评审专家。由于外审专家具备丰富的专业知识和大量阅读外文文献的基础,能发现部分跨语种抄袭文献,本组就有4篇在同行评议时发现。

3)接受读者监督。群众的眼睛是雪亮的,期刊要接受广大读者的监督,例3就是读者来信举报后,编辑部进行查证发现的。

5.2 开发基于先进计算机技术跨语种抄袭检测软件

依靠编辑、同行评审专家和读者去发现的概率有限,这与个人学术素养有关,存在一定偶然性。但单纯依靠文字相似性检测系统,并不能准确判断跨语种抄袭。虽然中国知网 AMLC 系统增加了多语种检测功能,但是由于英文比对资源有限,实际检出率并不高;VPCS 的检测比对结果对跨语种抄袭有一定的提示作用,但需要人工比对,深入分析判断。国外学者^[8-11]提出采用单词嵌入、知识图谱、语料库和主题分割算法等先进计算机技术检测跨语种抄袭。国内学者刘刚等^[12]开发了一种基于指纹融合的跨语种剽窃检测技术,李光曦^[13]通过构建一种基于多特征的跨语种剽窃检测模型从译文挖掘出的多种特征来解决跨语种剽窃问题。由于医学论文包含大量专业术语,建议将语料库和术语库整合到跨语种抄袭检测系统,扩大外文文献的比对资源,并将人工智能(AI)技术如自然语言处理、机器学习等应用到系统开发中,提高检出率。

5.3 加强学术诚信教育和惩处力度

对从事学术科研的单位和个人开展学术诚信教育,可利用融媒体技术,制作短视频,线上、线下会议,充分解读学术诚信相关的条例、法规,并结合真实案例进行讲解,做到心中有尺,做人有度。另外加大对学术不端行为的曝光和惩处力度,共同营造良好的学术环境。

总之,医学论文跨语种抄袭具有较强的隐蔽性,学术不端文献检测系统检出率较低,需要在出版流程中严格把关,开发基于先进计算机技术跨语种抄袭检测

软件以及加强科研诚信教育和惩处力度3个方面共同努力,促进学术诚信建设。

6 参考文献

- [1] 林琳,姜永茂,李英华.医学期刊编辑出版伦理规范[M].北京:人民卫生出版社,2018
- [2] 学术出版规范期刊学术不端行为界定:CY/T 174—2019[S].北京:中国书籍出版社,2019
- [3] 罗磊.医学英汉互译技巧与实践[M].北京:人民军医出版社,2004
- [4] 张重毅,方梅.科技论文隐性学术不端行为判别特征分析[J].中国科技期刊研究,2019,30(1):25
- [5] 袁松翔,刘功申.基于译文特征的中英文跨语种抄袭识别[J].上海交通大学学报,2012,46(6):991
- [6] 方梦之.从译学术语看翻译研究的走向[J].上海翻译,2008(1):5
- [7] 祁丽娟,耿静漪,方梅.跨语种抄袭和代写代投类学术不端行为的甄别和防范[J].中国科技期刊研究,2021,32(11):1347
- [8] ARIFIN Y, ISA S M, WULANDHARI L A, et al. Develo-

- ping a bilingual model of word embedding for detecting Indonesian English plagiarism[J]. Journal of Theoretical and Applied Information Technology, 2021, 99(17): 4338
- [9] HOURRANE O, BENLAHMAR E H. Graph transformer for cross-lingual plagiarism detection[J]. IAES International Journal of Artificial Intelligence (IJ-AI), 2022, 11(3): 905
- [10] MOHTAJ S, ASGHARI H. A corpus for evaluation of cross language text re-use detection systems[J]. Journal of Information Systems and Telecommunication, 2022, 10(39): 169
- [11] EHSAN N, SHAKERY A. Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information[J]. Information Processing & Management, 2016, 52(6): 1004
- [12] 刘刚,左权,杨倩茹.一种基于指纹融合的跨语言剽窃检测技术[J].计算机应用研究,2019,36(1):168
- [13] 李光曦.基于多特征的跨语言剽窃检测模型构建技术研究[D].哈尔滨:哈尔滨工程大学,2017

(2022-12-02收稿;2023-01-14修回)

对一篇《办刊之道》栏目稿件的退稿意见

本文详细介绍了作者所在期刊为提高学术影响力的具体做法,可以看出该刊编辑同人对于办刊是很认真的。但文章存在以下3个比较严重的问题。

1) 面面俱到地罗列的“举措”,几乎都来自发表的文献,是科技学术期刊办刊的常规做法,并没有显示与该刊办刊宗旨、学科性质、学术定位等相关的突出新意。

2) 通过图表及文字表述总结出的提高学术影响力的“成效”,都落脚到了被引频次、影响因子、下载率等指标的提升,以及进入这个“核心”那个“计划”,未摆脱长期以来形成的“唯指标论”的传统思维的影响,似乎这些就是办刊人追求的目标。在新时代,《编辑学报》不倡导、不鼓励这个导向。

3) 本文引用了《编辑学报》2018年第1期发表的《科技期刊学术影响力保障机制》中的观点“科技期刊的学术影响力可以用‘被引频次’‘影响因子’‘他引率’‘基金论文比’‘论文下载情况’‘国际论文比’等

多种指标进行计量”,并以此作为文章论述的依据。不可否认的是在2017年前后,关注评价指标、想方设法追求指标提升,确实几乎是科技期刊界的共识。但需要提醒的是,《编辑学报》发表的那篇文章的观点是明显有别于“共识”的。原文是这样的:“科技期刊的学术影响力主要体现在被引频次、影响因子、他引率、基金论文比、论文下载情况、国际论文比等多个指标上,以及编委、专家或作者的学术水平与学术地位,高水平的学术带头人作为期刊的编委、审稿专家或作者,可显著提升期刊的被关注度和影响力。”该文在论述“学术影响力保障机制”时,始终未具体提及评价指标!而本文的作者在引用该文时篡改了原文:一是将“主要体现在”改为“可以用”;二是删去了“以及”后的文字,更突显了评价指标。这种篡改文献原意的引用,属于学术不当,《编辑学报》也是不支持的。

审稿结论:鉴于以上,不建议《编辑学报》录用此文。