学术不端医学论文中重复文字的分布

刘 清 海
《中山大学学报(医学科学版)》编辑部,510080,广州

摘 要 采用 CNKI 的学术不端文献检测(AMLC)系统,对《中山大学学报(医学科学版)》230 篇来稿进行了重复文字比例的检测,发现约有 50% 的文章存在不同程度的重复文字现象,不同重复比例的文章数呈偏态分布,罕见重复文字比极高的文章。对其中重复比例较高的 64 篇文章进行了其各部分重复文字标记数的比较,发现医学论文中方法、结果和讨论部分重复文字标记数较多,而摘要、引言和材料等部分较少。对其中的临床研究和基础研究文章进行了比较,发现基础研究论文在引言、方法和结果部分重复文字的标记数高于临床研究论文,但总复制比却没有发现有统计学意义的差别。认为反对和抵制学术不端行为,是多个系统和部门共同协作的系统工程,期刊编辑也应尽其学术把关的职责。

关键词 学术不端文献;医学论文;重复文字比例;分布特征;AMLC 系统

Distribution of repeated words in medical manuscripts checked out by AMLC system//LIU Qinghai

Abstract The author of this article checked repeated words in 230 manuscripts contributed to the *Journal of SUN Yat-sen University* (*Medical Sciences*) by AMLC system provided by CNKI, found that almost 50% of the manuscripts had repeated words to some extend, and it was un-normal distributed of different word-repetition ratio articles, high word-repetition ratio articles were rarely seen. Among them, 64 articles were found relatively high word-repetition ratio, and were analyzed deeply from all parts of the articles. Word-repetition marks were more frequently found in the methods, results, and discussion parts than in the abstract, introduction, and materials parts. Compared with clinical articles, basic research articles showed more word-repetition marks in the introduction, methods, and results parts, but total repetition ratios were not significant. To argue against and reject academic misconducts is a complicated task that needs cooperation from various systems and departments, but journal editors should better bear their responsibilities as gate keepers.

Key words academic misconduct literature; medical article; words-repetition ratio; distribution characteristic; AMLC system

Author's address Editorial Department of Journal of SUN Yat-sen University, 510080, Guangzhou, China

近几年来,社会上的浮躁之风蔓延到了学术界,极大地损伤了科学共同体的诚信声誉。学术失范、学术不端乃至腐败的愈演愈烈之势已经引起了我国政府和学术界的高度重视,它们相继采取了果断措施,制定了《高等学校哲学社会科学研究学术规范(试行)》《科技工作者科学道德规范》,签署了《关于坚决抵制学术不端行为的联合声明》,出台了《关于严肃处理高等学校学术不端行为的通知》,编制了《高校人文社会科学学术规范指南》等[1-4]。而 2008 年底由 CNKI 科研诚信管理系统研究中心开发推出的检测学术文献当中不端行为的 AMLC 系统,则无疑为科研管理界和期刊编辑界提供了反学术不端论文的锐利武器。

《中山大学学报(医学科学版)》编辑部自 2009 年初起使用该 AMLC 系统,检测出部分学术不端论文,并提出了一些使用和完善该系统的建议[5]。本文中,我们对查出的可疑文献进行了进一步的比对和分析,希望发现学术不端的医学论文中重复文字的规律,为医学编辑和有关人员进一步研究提供依据。

1 研究对象选择与计数方法

我们检测了自使用 AMLC 系统 4 个月内《中山大学学报(医学科学版)》所有来稿和未及检测而已发表的最近 2 期稿件,共 230 篇。系统按照重复文字的比例,只提供相似文字比例≥5% 的文字复制比,并以 10%、30%、50% 为界划分为 5 种情况,即轻度句子抄袭、句子抄袭、轻度段落抄袭、段落抄袭和整体抄袭,并有相应的绿、黄、橙、红等颜色醒目提示[6]。进一步点击可疑论文发现,重复文字来源通常为多源,即与多篇文献的部分文字重复;因此,与某一篇文字的重复比例要小于总计的文字复制比,且与每一篇重复来源文献的复制文字比皆不尽相同。

为了有效地分析重复文字的分布情况,我们选定总复制比≥20% 的论文进一步点击复制比详情,若与其中某篇论文的复制比 >10%,则选定其中复制比最大的一篇作为重复来源论文进行比对。比对时,记录在中英文摘要、引言、材料与方法、结果与讨论各部分的重复文字标记数,标记的一段包括多个部分时,各部分分别计 1 次,重复文字若大部分是层次标题,则不予记录,材料与方法部分进一步细分为材料、方法、标准、统计学方法等 4 个小部分,图表的编排可能混在不同部分,皆还原为方法或结果部分。

为了便于比较不同类型论文中重复文字是否具有不同分布,我们区分了基础研究、临床研究、技术研究、信息研究、护理论文、综述等内容或类型的文章。

2 被检测论文的整体分布情况

在230篇被检论文中,重复文字比例≥5%的共116篇,<5%的(未标记,可以认为没有重复文字)为114篇。若按AMLC提供的分段标准计,则230篇论文的分布如表1;如细分之,按5%的区间进行分析,则其分布呈明显偏态分布,约一半论文没有重复文字,约12%的论文重复文字在5%~10%之间,重复文字比例较高者论文比例较小,最高者有1篇文献重复文字比例达81%,被系统诊断为整体抄袭。

表1 不同文字复制比按段统计的结果

文字复制比	区段篇数	百分比/%	累计篇数	累计百分比/%
<5%	114	49.56	114	49.56
≥5% ~ <30%	80	34.78	194	84.35
≥30% ~ <50%	22	9.56	216	93.91
≥50%	14	6.09	230	100.00

3 重复文字的分布情况

按照论文中文字总复制比≥20%且与某篇的最大复制比≥10%的原则,在116篇有重复文字的论文中筛选出64篇论文,并与其中具有最大复制比的重复来源文献进行了比对。结果发现,在该64篇论文中,总复制比的中位数为35.5%,最大复制比的中位数为24.5%,属于自我抄袭者24篇(37.5%),与重复文字来源文献有引证关系者仅4篇(约6%)。而每一篇论文的重复文字比对中发现,方法、结果与讨论中的标记重复数最多,而中英文摘要、标准与统计学方法重复文字的标记数较少,结果如表2。

表2 64篇较大复制比文章各部分重复文字标记数

论文各部分	最小值	最大值	中位数	4分位数
中文摘要	0	5	0	0~2.00
英文摘要	0	5	0	0~0
引言	0	5	1.5	1.00~3.00
材料	0	6	1.0	1.00~2.00
方法	0	11	3.0	1.50~6.00
标准	0	6	0	0~1.00
统计学	0	2	0	0~1.00
结果	0	17	4.0	0.25~5.00
讨论	0	18	5.0	3.00~8.00
总复制比	20	81	35.5	25.00~53.25
最大复制比	12	77	24.5	19.00~41.75

为了解不同内容或类型论文在重复文字方面是否有分布差异,我们记录了论文的类型。不过,最后发现,技术研究、信息研究、护理论文、综述文献极少,未达到统计分析的样本量,于是仅比较了基础研究(22篇)和临床研究(32篇)的分布,结果如表3所示。发现引言、方法、结果部分这2类论文重复文字标记数不同,基础研究的重复文字标记数多些,而其他部分尚未发现有统计学意义。

表3 基础研究与临床研究论文重复文字标记数比较(中位数(4分位数))

比较项目	临床研究(32篇)	基础研究(22篇)	*P*值
中文摘要	0(0~1)	0(0~2)	0.922
英文摘要	0(0~0)	0(0~0.25)	0.524
引言	1(1~2)	2.5(1~3)	0.026
材料	1(0~2)	1.5(1~3)	0.085
方法	2(1.00~4.75)	5(2.00~8.25)	0.001
标准	0(0~1)	0(0~0.25)	0.195
统计学	0(0~1)	0(0~1)	0.667
结果	3.5(0~4.75)	4(2~9)	0.027
讨论	6(3.25~8.25)	5(2.75~8.00)	0.571
总复制比	38(29.25~53.75)	34(25~54)	0.622
最大复制比	27.5(19.25~43.5)	26(21.75~43.75)	0.972

4 分析与讨论

我们前期检测了178篇论文,发现约有40%的论文或多或少存在重复文字现象,方法和讨论部分重复文字较多,标记的段落也较长,而结果部分重复较少,标记的段落也较短,多为句子或图表标题与注释等[5]。

本次被检测的230篇文章中,约有50%的文章存在文字重复现象,比例有所增加,可能与近段时间来稿中多了一些专门投本刊增刊的稿件有关。从检测的结果中,发现方法、结果、讨论的重复文字标记数较多,中位数分别为3、4、5,与前期的研究是一致的。因为方法与讨论部分确实存在较多重复文字,而结果部分虽然重复文字较少,但是由于其重复文字较短,标记数却不少;因此,在本次研究中显示其标记数与方法、讨论处在同一层次。

在筛选出的较大重复文字比例的不同类型论文比较中,发现基础研究的总复制比中位数为34%,略低于临床研究的38%,但没有统计学意义,而在不同部分的比较中,基础研究的引言、方法和结果重复文字标记数多于临床研究论文。可能是因为基础研究的引言部分通常较长、介绍科学背景知识较多而易于导致文字重复,方法部分则可能是由于一方面基础研究的实验通常需要较多种方法互相补充或佐证而导致方法部分内容较多,另一方面基础研究的方法多为通用性方法而难于创新,而临床上则相对较易于有一定的小改进。对于结果部分,则可能因基础研究的图表较多而导致图表标题和注释重复较多。

研究发现,医学论文重复文字现象比较严重,不同类型的论文中,不同部分重复文字的分布略有不同,但总体文字复制比并无统计学差异,说明不同身份的作者学术不端行为的比例并无不同。

学术不端行为造成的结果很多,而学术不端文献仅是其中之一。马勇进[7]列出了学术期刊中的不端行为,分为4 类;韩丽峰等[8]列举了学术成果发表中的不端行为,从作者、审者和编者 3 个方面共列举了 30 种不端行为;常亚平等[1]则主要依据《科技工作者科学道德规范》综述了 7 类学术不端行为共 25 小项。实际上,学术不端行为与学术腐败是不同的概念[9],依文献[4],不正当的学风分为 3 个层次,一是学术失范,二是学术不端,三是学术腐败,而学术腐败是学术不端的极端情况。众多学者对学术不端行为进行分类列举,有些分类却未能明确定义和区分这 3 种情况。陶范还对学术期刊中的失范行为进行了探析[10]。

无论如何,学术不端行为表现形式多样,结果复杂,影响因素也很多,需要政府、科研管理部门、学术界、社会评价环境、期刊编辑、社会舆论等多个系统与部门的通力合作才能比较有效地进行打击和遏制。学术期刊的编辑,作为反对和抵制学术不端行为的重要一个环节,必须担当起学术把关者的神圣职责。

5　参考文献

[1]　常亚平,蒋音播,阎俊. 基于组织因素的高校学术不端行为影响因素的敏感性分析[J]. 管理学学报,2009,6(2):264-270

[2]　教育部. 教社科[2009]3 号关于严肃处理高等学校学术不端行为的通知[S]

[3]　关于坚决抵制学术不端行为的联合声明[J]. 学术论坛,2009(4):封 2

[4]　教育部社会科学委员会学风建设委员会. 高校人文社会科学学术规范指南[M]. 北京:高等教育出版社,2009:1-6

[5]　刘清海,王晓鹰,孙慧兰,等. AMLC 检测医学论文的特点与期刊编辑部的应用对策[J]. 编辑学报,2009,21(6):526-529

[6]　中国学术期刊(光盘版)电子杂志社,同方知网(北京)技术有限公司. 科技期刊学术不端文献检测系统用户使用手册[S]. 2008:2

[7]　马勇进. 抵制学术不端行为:学术期刊的神圣职责[J]. 青海社会科学,2008(6):195-199

[8]　韩丽峰,徐飞. 学术成果发表中不端行为的形式、成因和防范[J]. 科学学研究,2005,23(5):623-628

[9]　何跃,袁楠. 学术腐败与学术不端的区别及其区分意义[J]. 科技进步与对策,2008,25(3):124-127

[10]　陶范. 学术期刊失范行为探析[J]. 编辑学报,2009,21(6):480-481

(2010-03-04 收稿;2010-03-15 修回)

编　辑　感　悟　6　则

赵　大　良

1)稿件处理得越慢,退稿的阻力越大,到最后作者没有其他地方投稿、发表,一定会与编辑据理力争;所以,应该加快稿件处理速度,避免干扰选稿的原则,特别是退稿时“没有”确切理由的稿件,不宜拖延。建议:不要可惜,不妨提出来大家一起讨论,及早决定。如果承诺作者 4 个月给出处理结果的话,至少应该在 3 个月前处理掉,避免被动。

2)引文的关联性,应提倡科学、适度、合理,反对人为操作、崇洋媚外、歧视中文文献等。“不引用投稿期刊的文章是不正常的”,我不完全赞同但觉得有一定道理:不是作者没有阅读或不了解所投期刊,就是期刊发文偏离了主学科方向或水平太低。合理引用既是对他人著作权的尊重,也利于与读者共享相关信息。

3)审稿人是同行专家,但不是期刊出版专家,所以审稿人的意见应该服务于办刊宗旨,而不是一味地遵从专家评价。如何采信专家意见,需要研究,既要尊重专家,又不唯专家是从。对单纯说好话的评审意见,需要慎重采信;否定性意见,需要评估侧重点。学术期刊应鼓励创新,特别是涉及本学科或跨学科方面的“上层模式”构建,新的理论、思想和方法的提出等等,具有较普遍指导意义的研究,哪怕在某方面不完整、不充分,但也有价值,甚至更有价值。

4)注重与作者的沟通,沟通需要站在“学术”“学科”的高度,而不是本刊规定的角度简单化处理。学术期刊应该具有学术精神,担当学术责任,但“每个刊物有各自的分工和能力”,这样落脚到“本刊要求”上来比简单地回绝要更好。通过沟通,传达期刊的办刊理念,树立期刊的品牌,是学术期刊出版的重要工作,不亚于发表几篇高水平论文。

5)学术出版的核心是创新,发表创新性的科研成果是一方面,同时也要求编辑出版工作本身创新,至少编辑出版人员要有创新的思维。否则,编辑不仅跟不上学科的发展,没有与科研人员对话的底气,也体会不到编辑工作的无穷乐趣。

6)研究和写作并不是科学家的专利,实践性工作更需要总结。在总结中提高是做好期刊编辑出版工作的基础,也是一条成功的捷径。没有时间对编辑实践进行总结和写作,那只是懒得思考、不求上进的借口。