

循证医学专业英语语料库建设研究*

周旭 窦川川 彭咏梅 刘海舟 王艳萍 吴勇奇 朱卫丰

(江西中医药大学 南昌 330004)

〔摘要〕 分析我国循证医学论文写作现状,结合项目实践从语料库设计、语料采集、语料加工及语料库应用等方面详细阐述循证医学专业英语语料库建设理论、原则和方法,指出该语料库对循证医学论文写作教学、翻译、教材编写和语法研究具有应用价值。

〔关键词〕 语料库;循证医学;专业英语

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.10.007

Study on the Building of Professional English Corpus for Evidence-based Medicine ZHOU Xu, DOU Chuanchuan, PENG Yongmei, LIU Haizhou, WANG Yanping, WU Yongqi, ZHU Weifeng, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China

〔Abstract〕 The paper analyzes the current situation of evidence-based medicine paper writing in China, elaborates the theories, principles and methods of the building of professional English corpus for evidence-based medicine in detail from the aspects of corpus design, corpus collection and processing, corpus application and so on combined with the project practice, and points out that the corpus has application value in the teaching of evidence-based medicine paper writing, translation, textbook compilation and grammar study.

〔Keywords〕 corpus; evidence-based medicine; professional English

〔收稿日期〕 2020-03-28

〔作者简介〕 周旭,博士,副教授,发表论文40余篇,参编论著1部;通讯作者:朱卫丰,博士,教授,发表论文200余篇。

〔基金项目〕 江西省高校人文社科基金项目“循证医学专业英语语料库开发与应用研究”(项目编号:YY18212);江西省教育科学“十三五”规划课题“大数据语境下的研究生学术英语写作教学研究——以《循证医学》课程为例”(项目编号:20YB148);江西省高等学校教学改革研究课题“基于大数据语料库的本科生英文论文写作能力培养——以《循证医学》课程为例”(项目编号:JXJG-19-12-14)。

1 引言

语料库(Corpus)是指收集海量自然语言进行加工整理、词性标注后转换形成的文本库。由于语料库中的语料数据来自报纸、杂志、演讲和论文等真实语言文本,有利于用户语言输出更加灵活、地道,广泛用于语言教学、翻译、写作和语言学研究等领域。目前COBUILD(Collins Birmingham University International Language Database)和COCA(Corpus of Contemporary American English)等代表性通用英语语料库达到数亿词规模,内容覆盖新闻、小说、杂志、口语、学术等领域^[1]。通用英语语料库规模庞大导致所收词汇全而不精,难以满足医学等专业科研领域使用。建立医学专业英语语料库是医学语言学交叉学科发展的必然趋势。

2 我国循证医学论文写作现状

2.1 现状概述

循证医学诞生于 1990 年,其核心理念是“最优医疗决策”需基于“当前可得最佳证据,结合临床医生掌握的技能和环境并充分考虑患者意愿和价值观”^[2]。经过近 30 年的发展,循证医学在临床医学、护理学、药学和公共卫生等各医学领域被广泛传播和接受,循证临床实践成为临床医生做出最佳医疗决策的基本要求^[3]。循证医学主要研究内容是生产证据,其最终载体是临床科研论文。目前我国循证医学研究产出证据质量普遍不高^[4]。本研究预调查随机抽取和统计 1 000 份 2013–2017 年我国科研人员发表的循证医学 SCI 论文样本,结果显示影响因子 5 分以上的论文仅占 5.5%。在循证医学论文写作方面英语能力不足是导致我国作者论文发表质量难以提升的根本原因。写作高质量循证医学论文对非英语母语作者英语能力有较高要求^[5]。然而我国循证医学研究者在专业学习阶段仅涉及少量英文术语,不足以在 SCI 论文写作中熟练运用词汇、句式和逻辑结构等要素表达论点,降低论文可读性,在较大程度上影响论文科学性,导致难以在高影响因子期刊发表。

2.2 医学专业英语语料库开发和应用

医学专业英语语料库的开发和应用为提高非英语母语作者的写作水平、增强专业英语表达能力提供有效解决方案。目前国内外医学专业语料库尚处于发展初期,建立了部分小型医学专业语料库(如临床医学和药学等),探讨了建库理论和应用方法^[6–7]。循证医学是交叉学科关系,更多涉及医学科研方法学(随机对照试验和系统评价的实施方案等)应用和研究,对应的语料库需突出循证医学特色。目前国内外尚未建立循证医学专业英语语料库。2017 年起笔者所在课题组建成首个 500 万词级循证医学专业英语语料库,在建库过程中逐渐完善建库理论和方法。

3 语料资源和建设方法

3.1 语料来源

发展至今循证医学相关论文在临床医学 SCI 期刊占据重要位置,各类循证医学网站和书籍层出不穷,形成英文语料大数据。语料库建设初期需优先筛选最具代表性、有利于指导循证医学论文写作的语料。经随机分层抽样,本语料库确定语料来源为:一是 2016–2018 年 5 大权威临床医学期刊:《新英格兰医学杂志》(*N Engl J Med*)、《柳叶刀》(*Lancet*)、《美国医学会杂志》(*JAMA*)、《英国医学杂志》(*BMJ*)和《内科学年鉴》(*Ann Intern Med*)发表的循证医学论文 1 277 篇,论文类型包括循证医学的 3 种主要证据类型:系统评价、随机对照试验和观察性研究;二是《如何实施循证医疗》(*How to Implement Evidence-Based Healthcare*)、《循证医学:如何实践和教育循证医学》(*Evidence-Based Medicine: How to Practice and Teach EBM*)和《护理和医疗中的循证实践:最佳实践指南》(*Evidence-Based Practice in Nursing & Healthcare: A Guide to Best Practice*)等 10 本英文循证医学教材或工具书;三是 3 大循证医学证据网站:Cochrane 协作网、UptoDate 和 Best Practice 的高质量证据正文 300 篇。语料库建成后最终容量共 503 万词。

3.2 文本降噪

获取语料原始文档后需经软件自动转换或扫描录入为 docx 文档,会存在大量单词拼写、符号和格式错误或不统一情况,需进行降噪才能实现后续标注和检索功能,确保语料正确性和可用性。降噪处理包括:(1)表格。提取标题、长文本和表注作为文本语料;表中短文本和数据弃用。(2)图片。提取标题和图注作为文本语料;图片内容经截图或转换文件格式保存。(3)格式。调整空格、手动换行符、回车符等错误,修正全半角字符、字母大小写、标点不统一等问题。(4)文本清理。删除作者单位、参考文献和各类附加信息声明等无用内容。

3.3 语料标注

要实现精确检索和匹配,必须对所有语料进行

词性标注。本研究主要使用 TreeTagger 软件进行词性标注,其标注精确度达到 95% 以上。TreeTagger 软件标注结果可被在线语料分析网站 Sketch Engine 等直接识别,有利于语料库推广使用。使用 TreeTagger 灵活应用命令行进行批处理可大幅提高效率。批处理命令为:

```
> set PATH = C: /TreeTagger/bin;% PATH%
> cd C: /TreeTagger
> for /f % a IN ( dir /b D: /Corpus/0/ *. txt) do tag
- english D: /Corpus/0/ % a > D: /Corpus/1/ % a. tag
```

将上述命令写入文本文档中,另存为 Loop. bat。命令实际含义为:将所有生语料“.txt”存放于“D: /Corpus/0/”文件夹,运行 loop. bat 文件即可批量将文件夹内生语料通过 tag - english 标注集转换为熟语料,以相同名称重命名为“.tag”文档,存放于“D: /Corpus/1/”文件夹。为减少遗漏或错误本次建库还使用了 CLAWS4、BFSU Stanford POS Tagger 1.1.2 和 UAM CorpusTool 3.3 等软件辅助标注,人工核对标注后结果,确保语料库检索和匹配等功能的精确使用。

3.4 语料库功能

语料库用户终端开发了检索、统计和匹配 3 项功能,主要使用场景为论文写作及辅助医学专业英语写作教学、教材编写、翻译和语法研究。(1) 检索。语料库开发了单词和词组检索两个功能模块,主要基于 AntConc 软件和 Wordsmith 软件实现。检索结果中提供词块、句子、段落和语篇展示,以及 pdf 原文和图片文件夹链接。(2) 统计。输入关键词后统计词频总数及分类统计,包括按期刊论文、书籍和网站来源分类;按临床医学的二级学科分类;按标题、摘要、背景、方法、结果和讨论等语篇段落分类。(3) 匹配。指定关键词前后名词、动词、形容词、副词和介词等词性匹配是实现论文写作表达准确、地道和多样化的重要功能,可用 AntConc 或 Wordsmith 软件实现,也可将语料库上传至 Sketch Engine 网站中实现匹配检索功能^[8]。

4 经验总结

4.1 建库思路

本研究主要出发点是提高我国循证医学科研从

业者英文论文发表质量,提升我国循证医学证据生产质量和加强本土证据的国际传播。专业语料库的发展及相关建库软件的成熟为解决这项难题开辟新方向。通过开发循证医学专业英语语料库,从循证医学期刊、书籍和网站中采集权威、真实的语料,使用计算机软件对语料进行清理、分类、统计和标注,使语料库实现检索、链接和对比等多功能应用,为循证医学论文写作、翻译、教学和语法研究等应用场景提供可信语料资源和高效工具。循证医学专业英语语料库为论文写作涉及的专业表达提供权威语料作为依据,大幅缩短积累积极语料时间,有助于提高 SCI 论文写作水平。

4.2 语料选择原则

4.2.1 分层随机抽样 目的是让语料充分覆盖各类写作场景。若使用单纯随机抽样则可能出现语料分配不均、某些临床学科或研究类型等语料覆盖率低或缺乏语料的情况。分层数量可按实际需求和语料库规模确定,以每个有实质异质性的小类包含 10 万词以上语料为佳。如本研究考虑到循证医学实际内容包含临床医学所有专业,第 1 级按疾病大类分层,含心脑血管、呼吸、消化、神经、内分泌和代谢、血液、泌尿生殖、恶性肿瘤和其他系统疾病等 9 个类别;第 2 级分层按循证医学常见的研究设计类型分层分为 3 个类型:系统评价、随机对照试验、观察性研究;在此基础上增加 1 类不属于特定临床专科的“方法学研究”,共 28 个小类,平均每个细类语料含量约为 18 万词。

4.2.2 时效性 医学专业知识更新速度较快,语料来源应尽量接近现时间点,本研究语料选取自近 3 年发表或出版的循证医学论文、书籍或在线证据总结,可较好地代表当前循证医学发展水平,计划此后每年更新 100 万词。

4.2.3 权威性 选取影响因子较高、影响力较大的期刊论文,若需选取书籍或在线资料作为语料以及较权威、传播范围较广的书籍或网站,可有效保证语料库质量。

4.3 语料采集

语料采集繁琐且单调,单纯以人工“复制-粘

贴”的采集方法效率低且易出错,应选择合适的自动转换和提取工具软件辅助。绝大部分原始文档为 pdf 格式需转换为 docx 格式才能进一步处理。使用传统转换软件等存在错词、缺少空格和排版混乱问题,后期人工校对成本较高。可选用光学字符识别软件 ABBYY FineReader 12,其内置词典具备自动纠错功能,英语文本识别准确率较高,基本实现“所见即所得”,可降低后期人工校对工作量。转换后的 docx 文档需做进一步清理,例如删除作者和单位信息(可保留第一作者姓名作为检索识别标记)、参考文献、基金信息、利益冲突声明、版权信息和表格内短文本等,以保证检索精确性和效率。此外较短的图、表标题和注解等作为语料的参考价值有限也可视情况删除。所有语料最终转换为 txt 文档才能被标注软件识别,使用 Word 宏命令或批量转换工具即可实现。

4.4 降噪处理

4.4.1 单词拼写错误 由于语料库体量庞大,对单词、符号和格式等错误的纠正需有批处理思维以节省人工成本。例如单词拼写错误:来源于 *N Engl J Med* 等期刊语料的 pdf 字体设置不利于软件识别,识别结果常出现两个单词之间缺少空格情况,某些单词如“Randomized”中的“i”常未被识别或误识别为“l”。可直接使用宏命令批量处理,核心命令为:

```
> myFile = Dir (myPath & " \. docx") #设置文档路径
> txt = InputBox (" 被替换文本:") #设置被替换文本
> Re_ txt = InputBox (" 替换为:") #设置替换文本
> Loop #循环
```

Word 宏批处理的局限性是只能一次替换一组文本,若需一次批量修正多组错词则需反复操作,效率不高。可使用专用软件(如“Word 文档批量处理大师”等)同时替换多组错词。

4.4.2 特殊符号批量替换 应善用正则表达式构成批量替换的通配语法。例如软件转换后小标题和正文后会产生手动换行符,若直接替换为空格则会使标题和正文文本变为同一行。此时可先将所有小

标题前后手动换行符替换成段落标记,再将正文中手动换行符替换为空格。最后替换所有手动换行符为空格,即可获得符合要求的语料文本。具体操作步骤为:Word 中打开替换对话框,勾选“使用通配符”,采用正则表达式将“^l (<??????? >) ^l”替换为“^13/1^13”。上述正则表达式中“^l”表示查找手动换行符、“(<??????? >)”表示查找 7 个字符长度内容、“^13”表示替换为段落标记、“/1”表示替换后文本维持相应查找内容不变。由于大部分小标题字符长度较短,将“?”的数量定义为 5~30 个字符即可批量替换掉 95% 以上小标题前后的手动换行符,剩余少量较长标题再进行人工处理。

5 意义

5.1 概述

循证医学专业英语语料库构建本质上是从权威专业期刊、书籍和网站中收集海量真实语句,用高质量真实语言样本描述循证医学专业英语写作时所需的语言结构和要素,经计算机软件标注转换,形成一个多功能检索、词性匹配、图文并茂的数据库,目的是服务于循证医学论文写作。目前国内医学本科和研究生培养阶段缺乏双语、论文写作教材和 SCI 论文写作课程,影响从业后的循证医学 SCI 论文写作,研究者只能通过阅读文献获取零散语料^[9]。这些语料绝大部分难以转化为在论文写作过程中灵活运用的积极语料。公共语料库虽然体量庞大,但用户无法判断从中获取的语料是否适合循证医学 SCI 论文写作,因此开发循证医学专业英语语料库十分必要。

5.2 语料库范围专业且广泛

国内外医学专业语料库尚处于起步和发展阶段,已有语料库数量少、规模小、发展松散^[10]。对于临床医学而言,专业之间联系紧密、交叉多,针对每一个疾病系统专门建库局限性较大。医学科研论文可分为临床研究和基础实验 2 大类,循证医学是指导创造证据和利用已有证据指导临床实践的学科,研究范畴包含了所有临床专业^[11]。因此循证医

学专业英语语料库应将范围设为所有临床研究,不涉及基础实验。该范围选择可保证循证医学专业英语语料库的覆盖面既专又广,能较好地解决 SCI 论文写作时所遇多学科交叉问题。

5.3 多领域应用

5.3.1 循证医学 SCI 论文写作教学 目前国内高校缺少循证医学专业英语写作教学课程,主要原因是专业师资力量薄弱^[12]。具备顶尖英语水平的循证医学专业教师数量较少且大多以讲授循证医学或临床医学专业课程为主,难以顾及 SCI 论文写作教学。而英语专业教师如讲授循证医学 SCI 论文写作则缺乏科研实践经验,应用语料库可较好解决上述问题。教师检索语料库筛选能理解、易教授的真实语料,包括术语、语法、语篇结构等元素,用于编写教案、讲稿和课件,可大幅提高备课效率和教学深度。

5.3.2 写作教材编写 循证医学专业英语写作教材需包含大量实例,而语料库中的权威真实语料可作为实例来源。

5.3.3 语法研究 学生或研究者撰写循证医学 SCI 论文面临的最大困难是写作语法生疏。写作中所遇语法问题往往需要结合专业语境灵活判断,一般语法教材缺乏相关内容,例如动作抽象名称的冠词使用和主谓一致问题;特定循证医学术语的介词惯用搭配问题;不同语篇和语境下的时态和语态问题等。基于循证医学专业英语语料库中的海量权威语料开展语法研究有助于阐明和总结该类语法现象和规律。

5.3.4 辅助中英互译 随着循证医学在全国范围深入普及,目前已有大量权威循证医学英文论文、书籍和网站资料翻译为中文,将译文和原文用 ABBYY aligner 等句对齐软件处理后可用于继续建设循证医学中英双语平行语料库。配合 Trados、MemoQ 等计算机辅助翻译软件,双语平行语料库可成为大型循证医学中英互译记忆库,提高翻译准确性和效率。

6 结语

在实践过程中提出一套完整的循证医学专业英语语料库建库理论和实践模式,最终建成500万词

级语料库,可辅助科研从业者提升循证医学论文写作能力、突破语言瓶颈,对循证医学论文写作教学、翻译、教材编写和语法研究具有应用价值。语料库建设和应用有助于提高临床研究证据质量,促进我国本土循证医学证据的国际转化和国际影响力提升,助推循证医学服务大众健康。同时本次循证医学专业英语语料库建设的设计、过程和结果可为医学乃至其他专业领域语料库的开发提供理论基础和经验参考。

参考文献

- 1 刘晓东,李德凤. COCA 英语语料库在英汉商务翻译教学中的应用 [J]. 中国科技翻译, 2020, 33 (1): 29 - 32, 61.
- 2 李幼平,李静,孙鑫,等. 循证医学在中国的起源与发展: 献给中国循证医学 20 周年 [J]. 中国循证医学杂志, 2016, 16 (1): 2 - 6.
- 3 赵晨,田贵华,张晓雨,等. 循证医学向循证科学发展的内涵和思考 [J]. 中国循证医学杂志, 2019, 19 (5): 510 - 514.
- 4 Yao L, Sun R, Chen Y L, et al. The Quality of Evidence in Chinese Meta - analyses Needs to Be Improved [J]. Journal of Clinical Epidemiology, 2016 (74): 73 - 79.
- 5 徐昉. 非英语国家学者国际发表问题研究述评 [J]. 外语界, 2014, 35 (1): 27 - 33.
- 6 马鹤桐,任建松,王序文,等. 双语癌症平行语料库构建 [J]. 医学信息学杂志, 2019, 40 (9): 77 - 82.
- 7 周永称,范少萍,晏归来,等. 精准医学文本语料库构建研究 [J]. 医学信息学杂志, 2019, 40 (12): 41 - 47.
- 8 Kilgarriff A, Baisa V, Bu? ta J, et al. The Sketch Engine: ten years on [J]. Lexicography, 2014, 1 (1): 7 - 36.
- 9 余党会,彭俊,惠朝阳,等. 医学研究生开设 SCI 论文写作课程的几点体会 [J]. 基础医学教育, 2015, 17 (3): 269 - 272.
- 10 Lei L, Liu D. A New Medical Academic Word List: a corpus - based study with enhanced methodology [J]. Journal of English for Academic Purposes, 2016 (22): 42 - 53.
- 11 张鹏,高静,叶熊,等. 临床医学教育中循证医学思维方法的应用研究 [J]. 中国继续医学教育, 2018, 10 (28): 32 - 35.
- 12 李宝珠,沙玉霞,周浩月,等. 临床医学专业学生对循证医学双语教学的认知与需求调查分析 [J]. 安徽预防医学杂志, 2019, 25 (2): 84 - 87, 92.