

壮族典籍多语平行语料库建设与应用研究

周秀苗

(广西百色学院国际教育学院 广西 百色 533000)

【内容摘要】文章综述了壮族典籍多语平行语料库的设计、建设和应用研究,该语料库具有库容大、开放性、便捷性等特点,为语言、文化、文学比较、翻译研究提供真实语料数据、技术支持等研究基础平台,还可以为其他少数民族典籍多语平行语料库建设提供借鉴和参考。

【关键词】壮族典籍多语平行语料库 设计 建设 应用

中图分类号:H128

文献标识码:A

文章编号:1007-9106(2017)10-0137-03

近几十年来,语料库、语料库语言学、语料库翻译学研究在国内外如火如荼,这一研究广泛应用于词汇、语法、语义、语言对比、词典编撰、二语习得、翻译、文学等领域中,并取得显著成效。为了应用或者研究需要,国内外先后建设了跨学科、多语言、内容多样、库容不同的语料库。目前,语料库特别是平行语料成为国内多语平行语料库建设与应用研究的热点,但语料库语料存在单语或双语语料库为主、多语语料库数量少以及民族典籍多语语料库建设滞后等问题。本研究首先构建壮族典籍(壮、汉、英)多语平行语料库,然后基于壮族典籍多语平行语料库检索数据,对壮、汉、英三种语言进行语言、文化、文学比较和翻译研究,以期为其他少数民族典籍多语平行语料库建设提供借鉴和参考。

一、多语平行语料库研究现状

就语料库建设而言,近五年,国际上较具代表性的多语平行语料库有:欧洲委员会联合研究中心研制的DGT-Acquis(2011)、ECDC-TM(2012)以及EAC-TM(2012)等三个语料库,语料主要涉及教育与文化、公共卫生和法律;比利时根特大学(Ghent University)(2011)研制的用于翻译学研究的荷兰语平行语料库(Dutch Parallel Corpus,

DPC)。这些语料库大多存在语料来源范围单一,主要用于语言识别、文档级对齐、专业术语提取等自然语言处理研究。具体信息见如下表:

在国内,多语平行语料库建设研究成果不显著。主要有:张姝、赵铁军等(2004)建设的英、日、汉三语平行语料库——“面向事件的多语平行语料库”,王成平(2012)建设了彝、汉、英三语平行语料库。就平行语料库应用研究而言,国内专家、学者主要在词汇、词典、语法、教学及翻译等领域开展

语料库名称	语料类型	语种与数量	用途
EAC-TM	文化与教育	26种欧洲国家语言	提供专业术语翻译服务
DPC	5种:教育、文献、文学、行政管理文件、国际通讯、新闻报道。	3个语种:英语、法语、荷兰语	翻译研究
ECDC-TM	公共卫生	25种欧洲国家语言	专业术语翻译服务
LLI-UAM 多语平行语料库	新闻(基于4个新闻文本语料库)	3个语种:阿拉伯语、西班牙语、英语	语言实际应用研究
DGT-Acquis	法律	23种欧洲国家语言	跨语言研究

* 本文为国家社会科学基金2015年度研究课题“中越跨境民族民间戏剧比较研究”(批准号:15XZW040);百色学院2014年度特色研究团队立项项目“桂西民族典籍译介研究团队”阶段性成果之一;2016年度广西中青年教师能力提升项目“南路壮剧及英译研究”(批准号:KY2016LX328)阶段性成果之一。

* 作者简介:周秀苗(1974—),女,广西百色学院国际教育学院教授,院长,主要研究方向为民族戏剧与翻译研究。

应用研究。词汇,如谢元花的语料库与词汇研究(2002);语法,如秦洪武、王克非(2009)基于对应语料库的英译汉语言特征分析;词典,如李德俊(2006)的基于英汉平行语料库的词典编写系统CpsDict;翻译与教学,如于莲江(2004)的基于语料库的翻译教学研究。

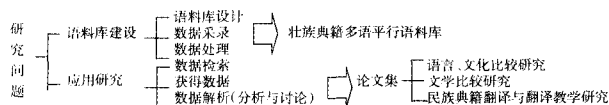
由此可见,国内多语平行语料库建设与应用研究主要存在以下问题:1.国内语料库语料内容不够丰富,主要为单语或双语语料库,多语语料库数量少。2.平行语料库多以英语为中心的双语语料库为主。3.国内多语平行语料库应用研究领域发展不平衡,主要集中于语言学、外语教学与词典学研究,较少用于文学和文化研究。4.国内民族典籍多语语料库建设滞后,不能适应文化“走出去”的要求。以上问题为本项目关于壮族典籍多语平行语料库建设与应用研究留下了空间,本研究具有以下意义:

1.学术价值:壮族典籍多语平行语料库建成可以对现有的壮族史诗、民歌、戏剧等优秀民族文化遗产进行数据采集和档案式保护。同时,为民族语言、文化、文学等领域的比较研究提供跨语言信息检索服务,也为翻译学和翻译教学研究提供真实语料与语言转换数据支持,通过课题相关研究成果,为民族语言文化、文学比较研究、民族典籍翻译学和翻译教学研究提供新思路、新方法。

2.应用价值:基于语料库检索数据进行以下研究,为壮、汉、英比较提供多学科新思路、新方法:语言、文化比较研究;文学比较研究;民族典籍翻译与翻译教学研究;为其他少数民族典籍多语平行语料库建设提供借鉴和参考鉴。

二、壮族典籍多语平行语料库前期设计思路

(一)整体研究思路如下图所示:



(二)语料库设计思路

1.文本子库。选取壮族史诗、民歌、戏剧等三大代表性典籍的壮、汉、英三种语言文本作为语料样本,建设以壮语版本为中心的单向平行语料库;壮族典籍文本子库,下设三个二级子库,每个二级子库均以壮语版本为中心,建立一对二,即壮语对汉语、英语的单向平行子库。

2.音频子库。建立与壮语文本相对齐的三大典籍的壮语音频子库。壮族典籍多语平行语料库音频子库,下设三个二级子库,每个子库内建立壮语与音频对齐的平行子库。

三、壮族典籍多语平行语料库的构建与应用研究

(一)壮族典籍多语平行语料库使用的是ELAN软件。ELAN全称是ELAN-Linguistic Annotator(语言学注解器),目前版本version 4.0.0。This manual was last updated on 2010-12-23。这是一个对视频和音频数据的标识进行创建、编辑、可视化和搜索的标注工具,旨在为标识提供声音技术以及对多媒体剪辑进行开发利用。虽然ELAN专门为语言、手语、姿势提供分析,每个人都可以用它来处理多媒体数据,如视频和音频,以便对其进行标识、分析和建档。ELAN(EUDICO语言注释)是允许创建、编辑、可视化以及搜索视频和音频数据的注释的批注工具。它诞生在荷兰奈梅亨的马克斯·普朗克语言心理学研究所,旨在为多媒体的开发和注释提供良好的技术基础。ELAN虽然是专门为语言分析、手语和手势设计的工具,但它可供任何从事多媒体,即需要视频和/或音频数据注释、分析报告和说明文档的人使用。

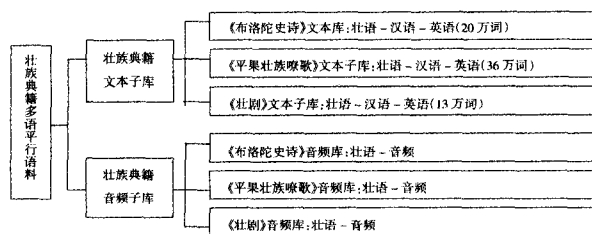
Elan的音频文件导入:通过Coll Edit录得语音文件,切分编辑后即可输入Elan运用,但要经过Praat软件编辑才会显示语音波形。Praat是一款语音学软件,通过它也能采集语音。Praat的主要功能是对自然语言的语音信号进行采集、分析和标注,并执行包括变换和滤波等在内的多种处理任务。作为分析结果的文字报表和语图,不但可以输出到个人计算机的磁盘文件中,和终端的显示器上,更能够输出为精致的矢量图或位图,供写作和印刷学术论文与专著使用。此外,Praat还可用于合成语音或声音、统计分析语言学数据、辅助语音教学测试,等等。随着新版本的发布,Praat的功能和用途仍在不断扩展。

(二)建设了壮族典籍多语平行语料库

1.壮族典籍多语平行语料库软件包括:(1)文本子库。选取被列入国家非物质文化遗产名录的壮族史诗、民歌、戏剧,而且已经发表的三大代表性典籍:韩家权著的《布洛陀史诗》,周艳鲜、陆连枝著的《平果壮族嘹歌》中的《三月歌》《房歌》以及

周秀苗著的《北路壮剧传统剧目精选》中的代表性剧目《蝶姹》、《农家宝铁》、《侬智高招兵》、《七女与龙子》、《太平春》、《朱买臣》壮、汉、英三种语言文本作为语料样本,建设以壮语版本为中心的单向平行语料库,库容为69万词。(2)音频子库。建立文本子库中的《布洛陀史诗》、《平果壮族嘹歌》、《壮剧》所选取的文本内容相对应的三大典籍的壮语音频子库,容量约为20GB,已经录制完毕。

壮族典籍多语平行语料库结构如下:



2. 壮族典籍多语平行语料库具有以下三大特点:(1)视角新。以被列入国家非物质文化遗产名录的壮族典籍《布洛陀史诗》、《平果壮族嘹歌》、《北路壮剧传统剧目精选》为研究对象,便于民族典籍的保护与弘扬。采用壮族史诗、民歌和戏剧等三大代表性典籍的壮、汉、英三种语言语料为样本,建设多语平行语料库,这种三语平行的语料库对壮族典籍来说尚是首次。(2)方法新。应用多语语料库、语料库语言学等理论,采用库助和库驱动研究相结合的实证研究方法对《布洛陀史诗》、《平果壮族嘹歌》、《北路壮剧传统剧目精选》进行民族语言、文化、文学比较研究和翻译研究,在研究方法上有所创新。(3)成果新。壮族典籍多语平行语料库具有库容大、开放性、便捷性等特点,是进行壮、汉、英语言、文学、文化比较研究和翻译研究的基础平台。

(三)开展壮族典籍多语平行语料库应用研究

基于壮族典籍多语平行语料库检索数据,笔者团队采用以下三个方法对语料库应用开展研究:首先,应用定量与定性研究相结合的方法,基于跨语言信息检索分析《布洛陀史诗》、《平果壮族嘹歌》、《北路壮剧传统剧目精选》中的语言文化现象,进行比较研究。其次,应用对比分析法,从语言、文化层面对语料进行对比研究和比较文学研究等。最后,以语料库翻译翻译学、语料库语言学理论为指导,应用语料库辅助研究和语料库驱动研究相结合的实证研究方法,进行翻译学与翻译教学研究。为此,笔者团队撰写、发表了语言、文化

比较研究、文学比较研究、民族典籍翻译与翻译教学研究论文8篇。《壮族嘹歌壮-汉-英三语平行语料库构建及其应用——中国壮族民歌与英国民歌认知模型比较之语料库研究》(百色学院学报,2015/01)、《语音与情感体验——中国壮族民歌与英国民歌语音情感效果比较》(广西教育学院学报,2015/06)、《北路壮剧亲属称谓文化内涵及英译策略研究》(社科纵横,2016/07)、《壮剧文化词及其英译对策研究》(百色学院学报,2016/06)、《壮剧修辞手法英译研究》(百色学院学报,2015/03)、《场域转换与文化对等:少数民族传统文化典籍跨语际实践探析》、《遗产译介话语建构与文化传播研究》、《民族典籍英译策略研究之以诗译诗》。

四、结语

壮族典籍多语平行语料库经过四年的建设,目前已经初具规模,但也存在以下不足:一方面是鉴于壮族典籍多语平行语料库建设是一个复杂的跨学科的研究工作,本语料库有待于进一步完善和扩充,而学界对民族典籍多语语料库建设重视不足,能借鉴的成果很少。另一方面是多语平行语料库应用研究领域发展不平衡,主要集中于语言学、教学与翻译研究,用于文学研究不多。今后可以再进一步深入研究的问题:基于壮族典籍多语平行语料库,加强文学角度研究。

参考文献:

- [1]周艳鲜,陆莲枝.平果壮族嘹歌(英文版)[M].广西师范大学出版社,2011.
- [2]韩家权.布洛陀史诗(壮汉英对照)[M].广西人民出版社,2012.
- [3]周秀苗.北路壮剧传统剧目精选(壮汉英对照)[M].广西人民出版社,2014.
- [4]O'Keeffe, Anne McCarthy, Michael. The Routledge Handbook of Corpus Linguistics [M]. London: Routledge, 2012.
- [5]Ho, Yufang, et al. Corpus Stylistics in Principles and Practice [M]. London: Continuum Publishing Corporation, 2011.
- [6]秦洪武,王克非.基于对应语料库的英译汉语言特征分析[J].外语教学与研究,2009(2).
- [7]于连江.基于语料库的翻译教学研究[J].外语电化教学,2004(02).
- [8]谢元花.语料库与词汇研究[J].外语教学,2002(01).
- [9]Braschler, M. & P. Schuble. Using Corpus-Based Approaches in a System for Multilingual Information Retrieval [J]. Information Retrieval, 2004.