

标识语汉英双语平行语料库的设计与创建

李广伟,戈玲玲^①

(南华大学 外国语学院,湖南 衡阳 421001)

[摘要] 规范的双语标识是文化软实力的重要体现,很大程度上反映了所在城市及国家的开放程度和管理水平,对于提升城市的国际形象具有重要的现实意义。文章在梳理国内外标识语研究现状的基础上,重点阐释了自建标识语汉英双语平行语料库的研制过程。在语料库设计方面,文章探讨了语料库建库目的、语料规模、语料选取标准、代表性和取样策略等五个方面的内容;在语料库创建方面,则对语料采集、语料的电子化、语料标注和语料平行等几个方面进行了讨论,旨在为从事相关翻译研究和实践的研究人员提供研究思路和借鉴。

[关键词] 标识语; 平行语料库; 设计与创建

[中图分类号] H314 **[文献标识码]** A **[文章编号]** 1673-0755(2015)01-0122-04

标识语(亦称公示语),是指在公共场所向公众公示须知内容的语言,包括标识、指示牌、路牌、标语、公告、警示等^[1]。标识语作为一种特定功能的文本形式,应用历史悠久,在世界各地民族的社会、文化、经济发展进程中都发挥着规范、协调、保障、稳定的功能机制^[2]。随着中国的国际影响日益扩大,国际交往日益频繁,标识语在全国各地的翻译与运用直接向世界展示了中国的对外交流水平以及现代化城市建设风貌^[3],同时它也是文化软实力的重要体现。国际化、现代化、城市化、信息化、社会化的迅猛发展对标识语翻译研究提出了更为迫切的要求。

语料库方法应用于译学研究,直接推动了翻译学研究方法的革新^{[4][5]}。语料库技术在标识语翻译研究中的运用不但可以更好地揭示出汉英标识语在词频、词长、句长及句法等方面的不同特点,而且可以揭示标识语翻译的语言特征及内在规律,从而推动标识语翻译研究的不断深入。

本文拟在分析国内外标识语研究的基础上,探讨“标识语汉英双语平行语料库”的研制过程,旨在为从事相关翻译研究和实践的研究人员提供研究思路和借鉴。

一 国内外标识语研究现状分析

在西方,标识语翻译研究相对发展较早,1959

年 Vinay & Darbelnet 两位学者就开始对公示语进行研究^[5]。1997年, Landry & Bourhis 提出了“语言风貌”的理念,为公示语研究提供了一个新的观察视角^[6]。20世纪90年代,针对安全标识在工作场所的重要性,欧盟发布了安全标识指令^[7]。发展至今,标识语研究已经逐步走向成熟和规范。

而在国内,针对标识语的英译研究则起步较晚。2002年北京第二外国语学院对汉英公示语翻译研究课题进行立项资助,并成立了公示语翻译研究中心。但严格来说,我国公示语翻译有规模、有组织、比较正式的研究应该从2005年北京首届全国公示语翻译研讨会的召开算起^[8]。发展至今,公示语翻译研究已成为国内翻译研究的大热门,其研究成果总量大大超过了新闻翻译、旅游翻译、法律翻译等应用翻译门类^[9]。截止目前,成果大体可归为5个方面:(1)注重对现状的调查分析。主要体现在对某一城市进行语言事实的分析,并就发现的问题提出对策^[10]。(2)理论与实践相结合。主要体现在研究者从不同视角,运用不同理论对公示语翻译进行探讨。所用理论主要包括文本类型理论^[11]、语用学理论^[12]、生态翻译学视角^[13]等。(3)研究成果的及时总结与述评。研究成果的总结主要包括《公示语翻译》^[14]和《公共标识英语纠错指南:实例、辨析和

[收稿日期] 2014-09-28

[基金项目] 湖南省教育厅科研项目“基于语料库的标识语英译规律研究”资助(编号:12C0368);教育部社科项目“基于汉英双语平行语料库的外宣翻译规范研究”资助(编号:14YJC740049)

[作者简介] 李广伟(1980-),男,山东菏泽人,南华大学外国语学院讲师。

^①南华大学外国语学院教授,博士。

解决方案》^[15]及对公示语研究的阶段性成果进行述评^[16]。(4)基于语料库的公示语研究。北京第二外国语学院公示语翻译研究中心开发研制了当今国内最大的公示语——“汉英双语标识语料库”^{[4][74]}。为了帮助大街小巷“说”一口更加地道的英文,上海市也于2012年启动了“公示语英译语料库”建设(<http://shanghai.eol.cn/>)。(5)双语标识规范化研究。目前,主要大城市,尤其是北京、上海、广州等地方政府相继颁发了规范双语标识的相关文件(如北京市政府颁发了《北京公共场所双语标识英文译法》;上海市语言文字工作委员会印发了《上海市公共场所中文名称英译基本规则》;广东省质监局批准发布了《公共标识译法规范》),其标识语翻译已经走上了规范、有序的良性发展的道路^[17]。

尽管如此,我国的标识语翻译研究仍然存在诸多问题。丁衡祁将2005年以前的问题归为三类:(1)硬性问题,即拼写错误、语法错误、字母颠倒等。(2)软性问题,包括语言知识、文化差异、翻译水平方面等。(3)隐形问题,包括“矢不对的”、语气不当等^[1]。截止目前,在全国很多地方上述三类问题不但不同程度的存在,而且还存在着中外信息互不一致,缺少规范和标准,各自为政,重复研究普遍,研究范围狭窄等问题。正如杨永林所指出的既缺少“宏观指导性”的研究,也缺乏“微观实证性”调查,更缺少国际化的规范与标准,不利于触及问题本质^[6]。由于社会重视程度不够,就连北京第二外国语学院建设的“全国公示语翻译语料库”也基本处于停滞状态,网站上的信息2008年后基本没有更新过^[16],而且语料没有进行深层标注,无法满足标识语翻译动态发展的强烈需求。

正如吕和发(2011)所说,一个景区、一座城市、一个国家,如果没有完善的静态公共服务体系即公示语体系,那么这个国家再“发展”,这座城市再“发达”,也只能是大理石和混凝土的“都市丛林”,绿水青山的“僻壤乡野”^{[14][3]}。

二 研究设计

本研究所建标识语汉英双语平行语料库(Chinese-English Parallel Corpus of Public Signs,简称CEPCPS)为双语专门语料库。它是指“关于特定主题文本的集合”,收集包括原文与译文的专题文本^[18]。该语料库始建于2013年3月,目前已初步建成,其对语言研究、翻译研究、翻译教学都具有极大的实用价值。以下着重探讨CEPCPS的研制,主要对该语料库的设计与创建步骤进行阐述。

(一)CEPCPS的设计

语料库的总体设计或研制方案,是决定今后语料库研究和应用的关键^[19],它不但是语料库建设的前提,还直接关系到语料库的价值和使用前景。在语料库设计方面,本语料库主要借鉴了北京外国语大学中国外语教育研究中心王克非教授汉英平行语料库的成功经验,主要考虑设计目的、语料规模、语料选取标准、代表性和取样策略等几个方面:

1、建库目的

语料库的设计应该围绕其建库目的即打算使用该语料库来做什么进行^[20]。本语料库将应用于语言研究、翻译研究及教学研究。我们将重点研究标识语语言在语篇、修辞、词汇等语言层面上的特点及翻译策略,探求汉语标识语的英译规律。同时为基于语料库的翻译教学提供真实而又多样化的资源。

2、语料规模

本研究拟创建一个动态、开放的英汉标识语双语平行语料库,初期库容约为100万字/词(为方便统计,汉语以字为单位,英语以词为单位)。随着时间的推移和研究的深入,研究人员将及时更新语料,把符合条件的文本不定期的录入语料库,以反映标识语的动态发展及语言的历时变化。

3、语料选取标准

标准问题实质上就是语料质量问题,即要保证源语言本身的质量、翻译的质量和相应的规范,还要有利于后期整理加工、收录和保存^[21]。在语料选取标准方面,主要考虑如下:(1)原文与译文必须有一一对应的翻译关系。(2)译文符合某种规范,如北京标准、英国标准或美国标准,且语料文本具有同质性。(3)以原语文本为基准,允许一对二、一对多等情况的存在。(4)所选文本具有代表性。(5)选取质量高的语料文本。

4、代表性

语料库的代表性是指语料库所收录的语料在多大程度上代表某一语言应用的种类或文本。语料库代表性是每个建设语料库的研究人员必须考虑的问题,它直接影响语料库研究的合理性和可靠性^{[4][44]}。

为确保语料的代表性,语料的选取要覆盖一定的范围,为此,研究人员选取了杨永林^[22]“千百十”项目中十个服务行业(即:交通服务、购物场所、餐饮服务、旅游景点、休闲场所、金融行业、通讯服务、医疗服务、文化设施及其他)。所选语料不但包含双语路标、路牌、广告口号、宣传语还包含环境保护、危机管理、安检防恐、无障碍设施、规约制度、应急救援等方面的标识语。

5、取样策略

在语料抽样方式上,本语料库参照了梁茂成^[23]等人的标准,将简单抽样与分层抽样相结合,以确保语料的代表性。简单抽样,是指按照设定标准,对某一类文本划定一个抽样的范围,如抽取某年出版的标识语专著上的所有有关标识语文本。分层抽样是根据标识语的功能类别或领域先进行分类,然后按照分类抽取文本,如按照标识语的六种突出功能“指示性、提示性、限制性、强制性、劝导性及宣传性”进行抽取。

(二) CEPCPS 的创建

译学语料库建设是语料库翻译学研究至关重要的第一步,而且技术含量高,主要涉及语料采集、语料的电子化、语料标注和语料平行等工作^{[4]19}。

1、语料采集及语料电子化

本语料库的语料主要来源于网上的电子文本、以纸质形式出现的印刷品、书籍或宣传页和实地拍摄的汉英标识语图片。

首先,纸质出版物通过扫描实现电子化,保存为 PDF 格式,再利用 PDF 编辑器—ABBY FineReader 11 将其转换为 word 文档,并进行人工校对。该编辑器拥有全球最通用的 ABBYFineReader 11 OCR 文字识别软件等功能,具有独一无二的简体中文、繁体中文和英文文本识别精确度和版面还原效果。

其次,图片语料提取其中文字部分,去掉非语言因素,并且不保留图片。针对直接从网上下载的语料,先把其转化为 word 文档,并去掉多余的空格、空行、断行、多余的语言符号及一些不必要的图形或符号。

最后,为确保语言质量,课题组首先排除一些翻译质量低劣或重复的原始语料,再请精通双语的专家和翻译教师进行二次审阅,然后将所有电子化的语料按照标识语在实际应用中 6 种突出的应用功能进行分类存储。

2、语料标注

标注问题始终是语料整理的最基本问题,平行语料库也不例外^[21]。CEPCPS 的语料标注参照 Cristina Monti^[24]的分类标准,对标识语从语言层次(如词性)副语言层次和超语言层次(如日期、来源等)进行标注,分为篇头信息标注和篇体信息标注两个部分。在标注前,首先检查语料是否为清洁文本;然后利用“文本编辑器”对文本进行批量整理,并进行人工校对,在文本格式下检查语料中的差错疏漏,然后进行篇头信息标注。标注内容如表 1 所示:

表 1 CEPCPS 基本篇头标注内容	
标记内容	标记
篇头	<HEAD>...</HEAD>
功能	<FUNCTION>...</FUNCTION>
出版信息	<PUBLISHER>...</PUBLISHER>
领域	<FIELD>...</FIELD>
时间	<TIME>...</TIME>
词性	<WORD CLASS>...</WORD CLASS>
来源	<SOURCE>...</SOURCE>
字数	<SIZE>...</SIZE>

表 1 显示,在篇头信息标注方面,标记符号采用国际上通用的尖括号<>,将标记内容置于尖括号内,其中起始标记为<……>,结束标记为</……>。标注内容主要涉及篇头、功能、出版信息、领域、时间、词性、来源、字数等 8 个标注项目,如<FUNCTION>“功能”用</FUNCTION>进行标识。

CEPCPS 的篇体标注包括段落标注、词性标注两个部分。段落标注标明段落的起始、结束及编号,词性标注标明英汉语料的词性或词类,以方便统计分析标识语的词汇密度及不同词性词汇的频率,从而揭示标识语特征及英译规律。在段落标注方面,本语料库应用新西兰奥克兰大学 MICHAEL Barlow 开发的 ParaConc 软件进行自动附码。在词性标注方面,鉴于英汉语言之间的差异,如汉语以字为单位,汉字之间没有空格,英语以词为单位,词与词之间有空格。我们分别采用 CLAWS7 及中国科学院计算机技术研究所开发的软件 ICTCLAS3.0 版对 CEPCPS 的英汉语料分别进行词性的自动附码处理,并进行人工校对。

3、语料平行对齐

语料之间的平行对齐是平行语料库建设的核心。双语语料之间的平行对齐一般包括篇章、段落和语句层面的对齐^[25]。篇章和段落层面的对齐可以采用软件自动实现。首先运用 ParaConc 实现双语语料在段落层面以及大部分源语文本和目的语文本语句之间的对齐,再进行人工检查;其次利用 ParaConc 内置的分割对齐单位和合并对齐单位等功能,实现双语语料的句级对齐处理。由于标识语包含各类文本,需要人工辅助对齐,具体而言,原文一句译成一句的,按一句对待;原文一句译成数句按一句对待;而针对一些类似“禁扔废弃物”(No Littering)、“禁止吐痰”(No Spitting)等强制性标识语仍按句子对待;而对一些诸如“急救”(First Aid)、“旅

游服务”(Travel Service)、“售票处”(Ticket Office)等指示性标识语,则按词组对齐处理,以保证后期信息提取。

最后,语料对齐后,将文本转换成XML格式,中英文分别存储,中英文同名,但以末尾处的“.CPS”和“.EPS”进行区分,“.CPS”代表汉语语料,“.EPS”代表英语语料,以方便语料库平行检索。

[参考文献]

- [1] 丁衡祁. 努力完善城市公示语,逐步确定参照译性文[J]. 中国翻译,2006(6):42-46.
- [2] 王颖,吕和发. 公示语汉英翻译[M]. 北京:中国对外翻译出版社,2007.
- [3] 吴贇. 一部公示语翻译的主臬之作——《常用标志英文译法手册》评介[J]. 外语电化教学,2013(4):76-80.
- [4] 胡开宝. 语料库翻译学概论[M]. 上海:上海交通大学出版社,2011.
- [5] 皮德敏. 公示语及其汉英翻译原则研究[J]. 外语学刊,2010(2):131-134.
- [6] 杨永林,刘寅齐. 双语公共标识文本的信息性研究——来自北京地区的报告[J]. 外语研究,2008(8):10-14.
- [7] 杨永林,赵珊. 双语标识译写研究——提示标识篇[J]. 外语学刊,2011(5):85-89.
- [8] 杨永和. 我国新世纪公示语翻译研究综述[J]. 外语教学,2009(5):104-108.
- [9] 邹彦群,满颖. 公示语翻译研究十年综述[J]. 上海翻译,2011(4):27-30.
- [10] 戴宗显,吕和发. 公示语汉英翻译研究——以2012年奥运会主办城市伦敦为例[J]. 中国翻译,2005(6):38-42.
- [11] 牛新生. 公示语文本类型与翻译探析[J]. 外语教学,2008(3):89-92.
- [12] 王永泰. 社会视角及语言视角下的公示语翻译[J]. 中国科技翻译,2009(4):42-44.
- [13] 束慧娟. 生态翻译学视角下的公示语翻译——以上海世博会主题标语为例[J]. 上海翻译,2010(2):39-42.
- [14] 吕和发,蒋璐. 公示语翻译[M]. 北京:外文出版社,2011.
- [15] 王欣,吕婷婷,李宁. 公共标识英语纠错指南:实例、辨析和解决方案[M]. 北京:中国书籍出版社,2010.
- [16] 李增垠. 二十年来的国内公示语英译研究综述[J]. 中南大学学报:社会科学版,2013(2):237-242.
- [17] 宋德富,张美兰. 小标牌大英文英汉公示语即查即用手册[M]. 北京:中国水利水电出版社,2011.
- [18] 李德超,王克非. 新型双语旅游语料库的研制与应用[J]. 现代外语,2010(1):46-54.
- [19] 王克非. 双语对应语料库研制与应用[M]. 北京:外语教学与研究出版社,2004.
- [20] 刘彬,戈玲玲,李广伟. 进口药品商品名英汉双语平行语料库的结构及其应用[J]. 南华大学学报:社会科学版,2012(5):121-124.
- [21] 崔卫,李峰. 俄汉-汉俄平行语料库的构建设想与应用展望[J]. 中国俄语教学,2014(2):1-5.
- [22] 杨永林,程绍霖,刘春霞. 北京地区双语公共标识的社会学调查——理论方法篇[J]. 语言教学与研究,(2007):1-6.
- [23] 梁茂成,李文中,许家金. 语料库应用教程[M]. 北京:外语教学与研究出版社,2010.
- [24] Monti C, Bendazzoli C, Sandrelli A, Russo M. Studing Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus.)[J]. Mata, 2005: 50.
- [25] 胡开宝,陶庆. 汉英会议口译语料库的创建与应用研究[J]. 中国翻译,2010(5):49-57.

The Design and Compilation of Chinese-English Parallel Corpus of Public Signs

LI Guang-wei, GE Ling-ling

(University of South China, Hengyang 421001, China)

Abstract: Being an embodiment of soft cultural power, the standardized public signs both in English and Chinese, to a large extent, are the reflection of the management level and the international image of the city, or even the whole country. This paper introduces the design rationale and the process of compiling the Chinese-English Parallel Corpus of Public Signs on the basis of a critical review of the existing study of public signs. The design and compilation of corpus includes the purpose, scale, criteria, representativeness, sampling, data collection, digitalization, tagging, and alignment respectively. The purpose of this paper is to provide a reference for researchers.

Key words: public signs; parallel corpus; design and compilation