中美旅游网站英语文本语料库的建设及应用

谷慧娟
(周口师范学院 外国语学院,河南 周口 466001)

摘 要:旅游网站的英语文本是专门用途英语研究的重要内容,语料库方法是切入旅游文本研究的一种实用方法。对自建小型中美旅游网站英语文本语料库的建设过程做了回顾,介绍语料来源、选取和分类标准、赋码过程;概述该语料库的整体情况,即平均词长、类符形符比、平均句长;讨论该语料库的应用前景,对促进网络旅游文本研究,提升旅游网站文本翻译质量的作用和意义。

关键词:网络旅游文本;语料库;英语
中图分类号:H31 文献标志码:A 文章编号:1671-9476(2016)01-0044-04
DOI:10.13450/j.cnki.jzknu.2016.01.011

经济、文化的全球化带动旅游业的发展。2012 年,旅游业收入占世界服务产业总额的 30%[1],旅游业经营的是经过开发的自然风光、人文景观和各种旅游产品及服务等,被称为绿色产业,已成为许多国家的主要收入来源,受到重视。对于旅游业而言,人们是否购买旅游商品很大程度上取决于能否事先有所了解,因而针对旅游资源和配套服务的宣传十分重要。当今,网络成为世界各国的旅游者查找获取旅游目的地信息的便捷途径,作为国际通用语的英语,成为网络旅游资料文本创作和翻译的主要语言。英语版的旅游网站或主页对于推介我国的旅游资源,扩大信息在网络上的传播,吸引更多的境外游客到中国旅游起着不可忽视的作用。因此,旅游网站的英语文本,逐渐成为一个专门的研究领域,受到多方面的关注。

语料库将大量真实语料汇聚起来,侧重语言使用的实际,被用来研究各类语言现象。与主要依赖研究者个体的语言直觉经验加上零星的语言实例的传统方法相比,更有助于揭示语言运作的深层机制。在旅游语言研究领域,研究者们利用语料库的优势分析英语旅游文本。李德超、王克非介绍了香港理工大学的双语旅游语料库的建设,该语料库属于较大类型,目前以收录香港地区旅游文本语料为主,既有网络文本,也有书面文本[2]。康宁自建了英语旅游文本类比语料库,用于研究中国旅游网站的英语文本与以英语为本族语的旅游文本在语言层面上的差异,但他对于文本语料的来源、内容构成、赋码过程等没有做详细说明[3]。侯晋荣用自建类比语料库来分析旅游文本的语篇及语言特征,她以表格形式简要介绍了人文和自然两类语料所占比例,而对语料库的其他情况没有加以说明[4]。本文将介绍中美旅游网站英语文本语料库的建设过程以及整体概况,并对其应用前景加以探讨。

一、中美旅游网站英语文本语料库建设

英语旅游文本通常提供关于目的地的食、住、行、游、购、娱信息,以及自然、地理、文化、风俗等知识,主要在于激发潜在游客的旅游行为,保障旅游活动的顺利实施。为了全面了解英语旅游文本的词汇、语言、文体等特征,我们选定网络文本为研究对象,并从 2012 年开始了小型语料库的建设。

研究所建的语料库为单语类比语料库,包括两

收稿日期:2015-08-20;修回日期:2015-12-04
基金项目:河南省教育厅 2013 年度科学技术研究重点项目指导计划"跨文化视角下河南省旅游资源的外宣翻译研究"(13B740272);2014 年度河南省科技厅软科学研究项目"跨文化交际视阈下河南旅游资源的外宣翻译研究"(142400411152);2014 年度河南省教育厅教师教育课程改革研究项目"高师英语专业基础阶段 CBI 教学模式研究及实践"(2014-JSJYLX-055)。

作者简介:谷慧娟(1974—),女,河南扶沟人,副教授,硕士,研究方向为语言学。

个组成部分,一类是美国旅游网站英语文本,另一类是中国旅游网站英语文本。建设美国旅游网站英语文本时,根据 2012 年美国国际贸易管理局下属的旅游业管理办公室(Office of Travel and Tourism Industries)发布的境外旅游者到访美国本土的统计数字[5],优先选取接待量排名在前 22 位的州,利用美国官方网站提供的推介旅游资源的平台,搜集下载信息。目前美国旅游网站英语文本的容量已达 130 438 个单词。中国网站英语旅游文本,选取境外游客接待量高的大陆知名旅游省市,利用各级政府旅游局官方网站,从其英文界面下载汇总文本作为语料来源。目前,中国旅游网站英文文本规模已达 126 074 个单词。

为保证语料的可比性,两类文本的取样框架趋于一致,语料的类别、内容、构成均相同。仿照 Maia 的做法[6]45,根据社会语域标准,旅游语料库文本的选取遵循分层抽样、结合均衡抽样原则,使入库语料具有一定的代表性,尽量同质,不致影响处理结果。内容包括旅游目的地的地理区划及简介,自然风光、历史古迹及典型旅游景点,风土习俗,美食购物娱乐,出入交通等,分别归入六个类别,即景点概貌、文化风俗、游览观光、购物餐饮、气候、交通,几乎涵盖旅游活动的各个方面,目前每个类别有 2 万~3 万个单词。

然后,对已建成的语料使用英国兰卡斯特大学 UCREL 研究小组开发的 CLAWS5(输出格式为:Horizontal)词性标注工具进行自动赋码。CLAWS5 标注集共包含 62 类词性,完全能够满足后期研究的需要。自动赋码完成后,针对中美旅游文本特点以及中英语言差异,进行了人工检查,修改错误赋码。对于中国网站英语旅游文本的词性标注,重点关注以-ng,-n,-an 等为结尾的专有名词的赋码。因为在英语中,-ing 是现在分词的词尾,有些过去分词会以-n 结尾,一些国名、地名之后加词尾-an 可以构成形容词,所以,中国网站旅游文本中含有-ng,-n,-an 的专有名词往往会被错误地标注为英语中的原形动词、不定式、分词或形容词。对美国网站旅游文本的词性标注检查,主要关注多义词的赋码,如 CLAWS TAGGER 把美国英语中的 fall 全部标注为原形动词-VVB,而 fall 在美式英语中可做名词,有"秋天"的意思,这时就需要人工介入,把词性赋码改为单数名词-NN1。以上两道步骤提高了词性标注的准确性,使语料库赋码的准确率达到 96%~97%,保证了后期研究的需要。

二、中美旅游网站英语文本语料库的总体特征

在完成了语料汇总和词性赋码之后,利用 Wordsmith 5.0 和 Antconc 3.2.4w 统计了两个部分文本的词长、类符形符比和平均句长,对文本的整体构成有一个概括性了解。

(一)词长

单词的长短与文本的词汇难易度有一定关联性。利用 Wordsmith 5.0,以字母个数为单位,统计语料库里不同长度单词的频率和总体平均词长,结果如表 1 所示。

表 1　不同长度单词在中美网站文本中的频率

词长	在中国网站的频率/占比	在美国网站的频率/占比	对数似然比	显著性(P 值)
1-字母词	3441　2.76%	3582　2.81%	0.3109	0.5771
2-字母词	19782　15.87%	18342　14.37%	-63.2588	0.0000
3-字母词	24663　19.79%	24968　19.56%	-1.1266	0.2885
4-字母词	17766　14.26%	19421　15.22%	26.4400	0.0000
5-字母词	14522　11.65%	16017　12.55%	28.0908	0.0000
6-字母词	11846　9.51%	13052　10.23%	22.2400	0.0000
7-字母词	12206　9.79%	11666　9.14%	-19.1600	0.0000
8-字母词	8725　7.00%	8121　6.36%	-25.8200	0.0000
9-字母词	5150　4.13%	5213　4.08%	-0.2500	0.6189
10-字母词	2972　2.38%	3321　2.60%	8.0100	0.0046
11-字母词	1991　1.60%	2106　1.65%	0.7230	0.3951
12 字母词以上	1558　2.15%	1830　1.85%	19.10	0.0000

由表 1 可知，中国网站 2 字母词、7 字母词和 8 字母词出现的频率较高，而 4 字母词、5 字母词和 6 字母词在美国网站文本的出现频率显著偏高。其中长度在 2～7 字母之间的词汇在美国网站文本中占 81.06%，在中国网站文本中占 80.87%。根据 Peter Norvig 对谷歌图书数据资源所做的统计，文本里英语单词的平均词长是 4.79 个字母，其中 80% 单词的词长在 2～7 个字母之间[7]，自建语料库的平均词长与 Peter Norvig 的统计结果一致，可以用来研究英语旅游文本。

经过统计，中国网站英语旅游文本的平均词长是 4.85 个字母，美国网站文本是 4.89 个字母，仅从数值判断，美国文本的单词平均略长于中国网站文本。又计算了词汇密度，即文本的实词数量在词汇总量中所占的比率。结果显示中国旅游文本的词汇密度是 0.59648，略低于美国文本的 0.59897。这两组数值说明美国网站文本的词汇难度要略高于中国网站文本。两位日本学者的研究指出，大英百科全书在线的平均词长是 4.32，而维基百科为 4.34[8]，他们结合其他易读性指标，认为大英百科全书在线的文本要比维基百科的文本简单。

(二)类符形符比

类符形符比反映了文本的词汇丰富程度，比值越高，用词越丰富多样；反之，用词较贫乏单调。由于两个子库所选文本主题类似，因此，其整体类符形符的比值可以反映中美旅游文本的词汇丰富程度。如表 2 所示，美国网站的类符形符比为 10.55，标准化类符形符比为 45.00，中国网站分别为 9.76 和 40.82，标准化类符形符比值反映出中国网站旅游文本的类符形符离散程度要大于美国文本。从这组数值判断，美国网站文本的词汇重复率少，相比较而言，中国网站文本的词汇重复率略高，用词表现得相对单调贫乏。

表 2 中美网站文本的类符形符比

	形符	类符	类形符比	标准化类形符比	标准化类形符比标准差
中国网站	126074	12307	9.76	40.82	56.58
美国网站	130438	13765	10.55	45.00	52.97

(三)平均句长

Butler 按长度把句子分为 3 类：1～9 个词长的为短句，10～25 个词的为中等长度句，25 个词以上的是长句[9]121。CLAWS5 词性标注工具有自动断句功能，因此，可以利用 AntConc3.2.4 索引工具有效统计文本的句子总数，进而计算平均句长。从统计结果观察，美国网站旅游文本的句长是 20.53，略大于 20 词，中国网站是 19.97，接近于 20 词。另外，Wordsmith 所计算的中美旅游文本平均句长分别是 17.66 和 16.67。据此判断语料库中所收录的旅游文本以中等长度句为主，美国旅游文本句子的平均长度要大于中国旅游文本。

综合观察比较单词长度、平均词长、词汇密度、类符形符比、平均句长这些指标，中国网站英语旅游文本表现出翻译文本的语言简化倾向，具体特征是词汇密度偏低，类符形符比的比值较小，句子较短[10]。相比较而言，语料库中所收集的美国网站英语旅游文本的用词更富于变化而多样，句子也相对更长一些。

三、中美旅游网站英语文本语料库的应用前景

目前，中国许多省市旅游局网站上英语旅游文本的语言质量亟待提高。通常中国网站的英语旅游文本是以汉语为源语翻译而来，网页中语言表述不地道、不恰当的例子比比皆是，错误频现。当潜在的海外游客通过互联网打开这些介绍中国旅游资源的网页，查找相关信息时，英语文本生涩难懂的语言难以激发阅读兴趣，将阻碍他们顺畅地浏览相关内容和获取资讯。而语言地道、生动、易读的旅游文本不仅能够有效地传播旅游资讯和中华文化，而且有利于旅游者获取相关信息，进而激起旅游兴趣，触发旅游行为。鉴于网站在宣传旅游资源中所起的重要作用，我国各级旅游部门应该重视网站英文版文本的质量，使旅游资讯得到有效的传播，促进涉外旅游业的发展。

本研究所建设的小型语料库可以用来帮助解决旅游文本汉英翻译中的语言质量问题，为改善和促进国内网站英语旅游文本的创作和翻译提供参考。

首先，网络中美英语旅游文本语料库建成后，可以利用 Wordsmith 5.0 或 AntConc3.2.4 等语料库统计软件，实现对文本的多角度定量分析。例如，生成旅游文本常用单词表，特别是中国历史文化常用词汇的恰当通用英语表达；描述、比较中美旅游文本的词类分布频率等。然后，以量化数据为依据展开定性讨论。例如，对词类的量化统计，有助于发现文本的词语运用模式和规律。其次，网站英语旅游文本的创作和翻译是一种特殊的文化交流活动，在英语旅游文本的翻译创作中，为达到等效的信息交流

目的,既要保证中英两种语言的恰当准确转换,重视表达的连贯性和清晰性,又要兼顾旅游语言独特文体风格的传达。因此,要细致分析语料库里的美国网站英语旅游文本,考察其在词汇用法、句式结构、语篇模式等方面的特点,探讨旅游文本与文学文本或应用文文本在文体上的差异,总结旅游文本的文体风格特征。最后,网络旅游文本语料库可以用作培训涉外旅游从业人员英语的资源库,还能为旅游英语教材的编写提供丰富多样的素材,促进专门用途英语的研究。

今后,要对所建语料库开展多方位的定量定性研究,为提升旅游资源的外宣翻译质量提供参考依据。翻译人员可以通过借鉴这些语料库研究的发现和结果,努力实现网站旅游文本措辞恰当,词汇搭配合乎英语的模式,少犯语法错误,减少用词冗余的现象,摆脱明显的翻译腔,尽量避免跨文化交际失误的产生。要想有效合理利用语料库,还要对之不断完善更新,将其建设成动态的语言。

参考文献:

[1]UNWTO. International tourism receipts surpass US $ 1 trillion in 2011[EB/OL]. [2015-05-07]. http://media.unwto.org/en/press-release/2012-05-07/international-tourism-receipts-surpass-us-1-trillion-2011.

[2]李德超,王克非. 新型双语旅游语料库的研制和应用[J]. 现代外语,2010,33(1):46-54.

[3]Kang Ning. Corpus-based stylistic analysis of tourism English[J]. Journal of Language Teaching and Research, 2011,2 (1):129-136.

[4]侯晋荣. 基于语料库的旅游文本语言特征及语篇分析[J]. 菏泽学院学报,2011,33(6):124-128.

[5]Office of travel and tourism industries[EB/OL]. [2015-06-30]. http://travel.trade.gov/.

[6]Maia B. Some languages are more equal than others: Training translators in terminology and information retrieval using comparable and parallel corpora[C]//In F. Zanettin, S. Bernardini &D. Stewart (eds.). Corpora in Translator Education. Manchester: St. Jerome, 2003: 66-80.

[7]Peter Norvig. English letter frequency counts: Mayzner revisited[EB/OL]. [2015-09-30]. http://norvig.com/mayzner.html.

[8]Adam Jatowt, Katsumi Tanaka. Is wikipedia too difficult?: comparative analysis of readability of wikipedia, simple wikipedia and Britannica, Proceedings of the 21st ACM international conference on Information and knowledge management[C]. New York, 2012: 2607-2610.

[9]Butler Christopher. Statistics in linguistic [M]. Beijing: World Publishing Corporation, 1991: 121.

[10]Mauranen A. & Kuyamaki P. Translation Universals: Do they exist? [M]. Amsterdam: Benjamins, 2004: 183-199.

On the construction and application of corpus of English texts from Chinese and American tourism websites

GU Huijuan

(College of Foreign Languages, Zhoukou Normal University, Zhoukou 466001, China)

Abstract: English texts from tourism websites are an important part of the study on English for special purposes. Corpus method is a practical way to approach the study of tourism texts. This paper reviewed the process of constructing a small-scale corpus of English texts from Chinese and American tourism websites, and introduced the source of texts, criteria of selection and classification, and tagging. The overall characteristics of the corpus, like average word length, type-token ratio, and average sentence length were investigated. Finally prospects for applying the corpus were discussed. The corpus will enhance the study on website tourism texts and help to improve the quality of translation.

Key words: website tourism texts; corpus; English