DOI:10.16800/j.cnki.jqustss.2012.04.022

● 语言学研究

基于类比语料库的中国网站英语旅游文本语言分析①

○ 康宁
（青岛科技大学　外国语学院，山东　青岛　266061）

［摘　要］基于对自建的英语旅游文本类比语料库，通过与英语国家网站旅游文本的相互参照和比较，以及对中国网站英语旅游文本的语言分析，发现中国网站英语旅游文本的标准化类符/形符比（STTR）与平均词长均低于英语国家网站旅游文本；中国旅游网站英语文本句子长度长于英语国家旅游网站英语文本。此外中国网站旅游文本的词汇密度大于英语国家网站旅游文本。从实词类词频上看，中国网站英语旅游文本的名词、形容词、副词均显著超用，而实意动词、第一人称代词（we, our, us）和第二人称代词（you, your）呈现显著少用的特点。该发现有助于提高和改善中国语境下的英语旅游文本语言质量，从而推动中国旅游资源的对外有效传播。

［关键词］英语旅游文本；类比语料库；语言特征
［中图分类号］H315.9　［文献标识码］A　［文章编号］1671-8372（2012）04-0105-05

Linguistic analysis of English tourism texts on Chinese websites: a study based on a comparable corpus

KANG Ning
（School of Foreign Languages, Qingdao University of Science and Technology, Qingdao 266061, China）

Abstract: By analyzing and studying the tourism texts of Chinese websites and English spoken websites in the Comparable Corpus of English Tourism Texts, which is compiled by the author and his research team, we found that the STTR of English tourism texts on Chinese websites is lower; the average word-length is shorter while the average sentence-length is longer. It is also noted that the lexical distribution of English texts on Chinese websites are denser than on English spoken websites. Moreover, Chinese tourism websites employ more nouns, adjectives and adverbs but less lexical verbs, the first person and second person pronouns. It is hoped that these findings will contribute to the improvement of English tourism texts on Chinese websites.

Key words: English tourism texts; comparable corpus; linguistic features

一、引言

旅游业被称为“朝阳产业”，在全球的发展方兴未艾。我国国土广阔，旅游资源丰富，据世界旅游组织预测，2020年中国将成为世界第一大旅游目的地国家[1]。为对外推介自己的旅游产品和服务，各个国家和地区借助互联网的优势，纷纷建立旅游门户网站，我们国家也不例外。我国的旅游网站除了中文版外，大部分都有多语种超链接，其中以英语为主，目的就是向国外游客介绍本地旅游资源，吸引外国游客前来参观。在这些英文版网站中，除了影音资料外，英语文本的质量对于实现这一目的将起到关键作用。本文的出发点是在语言层面分析我国（不包括港澳台地区）旅游网站英文

①课题组组长：康宁；成员：姜德杰，杨一秋，房钰，张煜，张其云，常艳。
［基金项目］山东省高等学校科技计划项目（12WG24）
［收稿日期］2012-09-03
［作者简介］康宁（1972-），男，河北丰润人，青岛科技大学外国语学院副教授，文学博士。

版的语言质量，旨在发现问题和不足，从而为中国网站旅游文本的英语写作及英译提供一定意义上的帮助。

二、研究方法与设计

本项研究采用基于语料库的研究方法，对量化数据进行定性分析。基于语料库研究的最大优势在于使用真实语料描述语言特征，从语言现象入手，能够更加客观准确地揭示本质。我国学者桂诗春指出，语料库研究方法的特点在于它通过大量语料从纵向找寻重复出现的语言形式，以提高对语言系统的洞察力，从而为社会实践服务[2]。

为此，我们建立了英语旅游文本类比语料库（comparable corpus of English tourism texts）。类比语料库是由不同语言的文本或同一种语言不同变体的文本所构成的语料库，包括两个或两个以上的子语料库[3, 4]。类比语料库与平行语料库（parallel corpus）不同，其子库之间并无翻译关系，类比性在于所采语料相同或相似的平衡性和代表性[5]。本文所建语料库为单语类比库，包含两个子库，即英语国家网站英语旅游文本库和中国网站（不包括港澳台地区）英语旅游文本库。为保证语料的可比性，两子库的取样框架趋于一致，取样均为完整篇章样本，单个样本大小为200—800个英文词。目前两子库规模分别达到136万词和53万词。为便于提取研究数据，我们利用CLAWS4软件对语料库进行词性附码标注，并对标注结果做人工抽检，以尽可能保证标注的准确性，同时利用WordSmith4.0、AntConc3.2等语料库统计软件检索数据，对数据进行分析、描写，并利用卡方检验（X^2）语言差异是否具有显著性。

研究中，我们以英语国家网站英语旅游文本子库作为参照语料库，对中国旅游网站的英语文本的语言特征、语言质量加以量化分析。分析内容主要包括：文本总体统计特征、词汇密度与覆盖面、词频统计等。通过对比分析，找出中国旅游网站的英语文本与以英语为本族语旅游文本在语言层面上的差距和不足。

三、数据统计与分析

1. 文本总体统计特征

语料库文本总体统计特征一般包括文件的字节数、形符数、类符数、类符/形符比（TTR）、标准化类符/形符比（STTR）、平均词长、平均句长、段落数，等等[6]。首先，我们利用WordSmith工具统计出两个子库的形符数、类符数、类符/形符比及标准化类符/形符比（见表1）。

表1　形符数、类符数、类符/形符比与标准化类符/形符比

	中国网站语料	英语国家网站语料
形符	532969	1368318
类符	17988	38065
TTR	3.38	2.78
STTR	41.02	44.21

根据Baker的文学翻译理论，类符/形符比值的高低与写作者的词汇丰富度和多样性成正比关系[7]。类符/形符比与语料库容量的大小有关，但由于本研究两个子库的大小不同，单纯比较两库的形符/类符比没有意义。因此我们需要将二者的比率做标准化处理（这项工作由WordSmith软件自动完成计算），即对语料库中文本每1000词的形符/类符比都进行计算，最后计算出平均形符/类符比[8, 9]。从表1的统计结果可以看出英语国家网站文本标准化类符/形符比高于中国网站文本，这表明英语本族语的旅游文本在用词方面变化多，用到的词汇量要大于中国网站英语旅游文本。这一点并没有超出我们的预料，毕竟中国网站的旅游文本多是由中国作者或译者完成，相对于本族语作者来说，其英语熟练运用程度还是有很大差距的。

其次，我们统计了两个子库的平均词长。结果显示，英语国家网站文本的平均词长为4.91，而中国网站文本为4.52。从图1可以看出，2–、4–、5–、6–字母单词在中国网站英语文本语料库中的覆盖率要高于英语国家网站文本；而其他长度的单词覆盖率则低于英语国家网站文本。

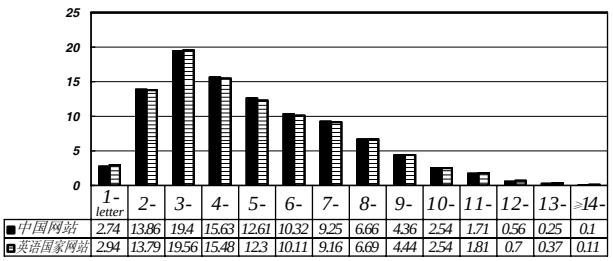

图1　中国网站语料库与英语国家网站语料库词长覆盖率对比

为验证不同长度单词在两个语料库中使用的频数是否具有显著差异和超用、少用趋向，我们计算各种长度单词出现频数与语料库出现的所有单词（即形符数）之间的卡方值，结果如表2所示。

表2　中国网站语料库与英语国家网站语料库词长对比

词长	在中国网站语料库分布频率	在英语国家网站语料库分布频率	卡方 X^2值	P值		
1-字母词	14621	40115	51.9624	0.000	***	-
2-字母词	73872	188307	1.5021	0.220		+
3-字母词	103404	267076	6.2433	0.012	*	-
4-字母词	83312	211341	6.7922	0.009	**	+
5-字母词	67222	167914	34.8738	0.000	***	+
6-字母词	55075	138040	20.8897	0.000	***	+
7-字母词	49234	125107	2.5490	0.110		+
8-字母词	35501	91282	0.3777	0.539		-
9-字母词	23242	60571	5.1842	0.023	*	-
10-字母词	13513	34729	0.1050	0.746		-
11-字母词	9121	24771	23.1676	0.000	***	-
12-字母词	3007	9502	101.6560	0.000	***	-
13-字母词	1326	5061	169.8337	0.000	***	-
≥14-字母词	519	1502	5.7514	0.016	*	-

注：P≤0.05，*表示显著差异，-表示少用，+表示超用

与英语国家旅游文本相比，中国网站英语文本中1-、3-、9-、11-及以上字母长度单词的使用频率显著偏少，而显著超用的只有4-、5-、6-字母的单词，其他不同长度单词的使用频数，在两个语料库中不存在显著差异。结合图1和表2数据，我们可以看出，中国网站英语文本从整体趋势上看多使用短单词，而英语国家网站文本则多使用长单词。这在一定程度上表明中国旅游网站英语文本更倾向于使用简单词汇。这可能与作者和译者的英语水平有关。

在平均句长方面，中国网站文本为19.23词，英语国家网站句长为15.44词。这组数据最初令我们感到意外，从英语熟练角度讲，中国作者和译者理应多使用简单句式，以短句为主，在句长方面应比本族语作者用的短。通过分析具体实例，我们发现这可能与译者（作者）受汉语无明显句子标记、大量使用流水句的影响有关，亦即受到母语语法负迁移的影响较大[10]。

例1：Due to the rich tourism resources and beautiful environment, in 2006, the National Tourism Administration and the United Nations World Tourism Organization named Dalian as "The Optimum Tourist City", this title further expands the reputation of Dalian both at home and abroad, thus attracts more tourists to travel and visit Dalian.（51 words）

例1选自大连市旅游局主办的大连旅游网。该句用词达51个单词，句式结构明显受到汉语造句规则的影响，语法上出现错误，用词啰嗦重复。试改写如下：

Rich tourism resources and beautiful environment earned Dalian a title of "The Optimum Tourist City" in 2006, from the National Administration and UN World tourism Organization. This has made Dalian well-known both at home and abroad, and attract more and more visitors.

例2：Luoyang lies in the west of Henan Province, crossing two banks of the middle reach of Yellow River and is one of the first historical cities recognized by the Council of State and famous ancient capital of some dynasties.（39 words）

上面例子选自新华网英文版对洛阳市的介绍，虽然行文没有语法错误，但冗长拖沓，与崇尚简洁的英语背道而驰，带有明显的中式英语痕迹。试改写如下：

An ancient capital of several dynasties, Luoyang lies in the west of Henan Province, crossing two banks of Yellow River in the middle reach. It is one of the first cities recognized by the State Council as a “Famous Historical City”.

2. 词汇密度

国内外学者利用不同的计算方法计算词汇密度，如Ure①、Holliday②、桂诗春③。本文采用的是Halliday提出的计算方法。Halliday把句子中的词汇分为两大类，即语法项和词汇项。前者又称功能项，是指在句中起到限定作用的词语，其中包括冠词、代词、介词、连词、部分副词和限定性动词；后者是指实词，如名词、动词、形容词、大部分副词、人称代词、数词等，篇章中实词传递了大部分信息。词汇密度反映的是篇章中实词所占的比例，比例越高，信息量越大，语篇难度也会越高。中国网站语料与英语国家语料的词汇密度对比见表3。

①参见 Ure J:《 Lexical density and register differentiation 》; Perren, G, J L M Trim (Eds.):《 Applications of Linguistics 》（Cambridge: Cambridge University Press. 1971）。

② 参见Halliday M:《 The Language of Science 》（Beijing: Peking University Press, 2007）。

③参见桂诗春：《基于语料库的英语语言学语体分析》（北京：外语教学上研究出版社，2009:29）。

表3　中国网站语料库与英语国家网站语料库词汇密度对比

	中国网站语料	英语国家网站语料
总词频数	532969	1368318
实词词频数	316817	729388
词汇密度（%）	59.44	53.31

计算结果显示，英语国家旅游网站英语文本词汇密度为53.31%，而中国旅游网站英语文本词汇密度则达到了59.44%。这一方面表明，中国旅游网站英语文本包含了大量的信息，这与中国旅游资源包含大量的历史、文化信息，而不仅仅是自然风光的特点有关。另一方面也说明，如此大量的信息，如果处理不当，会给英语读者造成阅读障碍，影响旅游资源的传播效果，如：

例3：The exhibition can be divided into seven sections which emphatically reveal the prosperity of the Prehistoric Age; the Zhou; Qin; Han; Wei; Jin; North and South; Sui; Tang; Song; Yuan; Ming and Qing dynasties.

本句介绍的是陕西省博物馆按朝代分区展览的情况，文中涉及了中国主要的朝代名称。而多数英语国家读者对中国历史知之甚少，原句对他们来讲意义不大，所以若在每个朝代名后面加上起止时间，效果会更好一些。

此外，如此高的词汇密度也和文本创作者的语言水平有关，中国网站英语旅游文本的写作或翻译受制于作者或译者的英语水平。从语料库中我们发现，中国网站英语文本有很多带有明显中式英语痕迹，如冠词、介词、连词等语法词用得少，名词、动词、形容词罗列堆砌现象较为常见。

例4：Guilin's *unique natural* and *man-made* landscapes provide abundance of options for *never-ending* and *unforgettable lifetime* journey.

该句中使用了6个形容词（斜体所示），尤其是最后3个，让英语读者感觉华而不实，虚张声势，影响了旅游文本的宣传效果。如果把这3个词换成your (journey)，不仅用词简洁，更能拉近读者与文本的距离。此外，这句话中漏掉了两个冠词，即abundance前的“an”，和never-ending前的“a”，冠词的少用（underuse）和误用（missuse）一直是中国英语学习者难以克服的问题[11, 12]。

3. 词频统计

词频是指某一词项或某一类词汇在语料库中出现的总频数，统计词频能够为语篇的文体或语体特征提供重要参考信息。本文以名词、实意动词、形容词、意义副词、人称代词（we，us，our，you，your）作为研究对象，对比分析上述词类在中国网站语料库和英语国家网站语料库中的词频分布差异，并利用卡方检验两者之间是否具有显著性差异。统计结果如表4所示。

表4　中国网站语料库与英语国家网站语料库实词词频对比

词类	在中国网站语料库分布频率	在英语国家网站语料库分布频率	卡方X^2值	P值		
名词	161543	386276	809.0792	0.000	***	+
实意动词	44610	130674	638.0118	0.000	***	–
形容词	72111	109192	13696.1758	0.000	***	+
副词	26702	60616	294.5666	0.000	***	+
人称代词	4530	19020	914.5649	0.000	***	–

从表4可以看出，上述词类的使用频次在两个语料库中都存在显著差异。其中在名词、形容词和副词的使用上，中国旅游网站英语文本远远多于英语国家旅游网站的英语文本。通过分析，我们认为这是由于中国旅游网站英语文本的作者和译者深受汉语旅游文本语言特点以及汉民族的思维方式、文化传统和审美情趣的影响。汉民族有着独特的社会历程和文化传统，汉语言对景物的描写大多文笔优美，用词凝练、含蓄，景物刻画不求明细，讲究音韵和美、“情景交融”，追求意象的朦胧之美。“重整和，重中和”是我们汉民族人文思想和艺术审美观的特点。然而，英语却不同，在描写景物时，客观具体，重理性、重写实、重形象，而非意象的直观感，最忌华而不实，累赘堆砌，追求流畅自然之美。因此在用英语创作旅游文本或英译汉语旅游文本时，一定要符合英语本民族旅游文本的特点，符合英语国家本民族的传统和习惯。

在实意动词的使用上，中国网站的英语旅游文本明显少于英语本民族的旅游文本。通过对比两类语料库，我们发现英语国家网站英语旅游文本中有很多关于旅游活动的介绍，涉及动作描写较多，而中国网站英语旅游文本大多是静态描写某一景点或景区。从这方面来讲，中国旅游网站在对外宣传旅游资源时，不妨借鉴英语国家旅游网站，多做一些旅游活动、节日方面的介绍。

此外，中国网站英语文本在人称代词（we，us，our，you，your）的使用上也明显少于英语本民族旅游文本。究其原因也是受到汉语原文本的

写作风格影响。汉语旅游语篇往往把某一景物作为描写对象，大量笔墨集中于景物刻画，涉及游客行为时，多用visitor，traveler等词，显得生硬；而英语旅游文本则在介绍旅游资源的同时，往往以主人身份（we，us，our）向游客（you，your）发出邀请、给出建议，有时提出警示。这往往能够拉近读者与文本的距离，让读者深受感染，从而达到良好的宣传效果。因此中国旅游网站英语文本的作者或译者应当摆脱汉语旅游文本写作风格的影响，多学习英语本民族旅游文本的语言表达方式和行文特点。

四、结语

本研究基于英语旅游文本类比语料库（comparable corpus of English tourism texts），以英语国家旅游网站英语旅游文本子库作为参照，对中国旅游网站英语旅游文本的语言特征进行了分析。研究发现中国网站英语旅游文本的标准化类符/形符比（STTR）低于英语国家网站旅游文本，词频表中前2000词的词汇覆盖率则明显高于英语国家网站旅游文本。这表明中国作者和译者的词汇量小，用词变化不多，且用词较为集中。从平均句长上看，中国旅游网站英语文本句子长度长于英语国家旅游网站英语文本，表明作者和译者受汉语无明显句子标记、大量使用流水句的影响，亦即受到母语语法负迁移的影响较大。此外中国网站旅游文本的词汇密度大于英语国家网站旅游文本。从实词类词频上看，中国网站英语旅游文本名词、形容词、副词均显著超用，而实意动词、第一人称代词（we, our, us）和第二人称代词（you, your）呈现显著少用的特点。

笔者认为，本研究中自建的类比语料库以英语国家网站旅游文本为标杆，有助于分析中国网站英语旅游文本的语言特征、文体特点，找出语言表达中存在的问题，从而提高中国旅游网站英语文本的语言质量，充分发挥旅游网站在对外宣传中的效果。

［参考文献］

[1] Word Tourism Organization.Tourism 2020 vision[M]. vol.3.2000.
[2] 桂诗春. 基于语料库的英语语言学语体分析[M]. 北京：外语教学与研究出版社，2009：10.
[3] 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，2004：7.
[4] Baker P. Using Corpora in Discourse Analysis[M]. London：Continuum, 2006： 89.
[5] 梁晓鹏，康宁. 旅游文本翻译研究的语料库途径[J]. 青岛科技大学学报（社会科学版），2010（4）：115-117.
[6] 杨惠中. 语料库语言学导论[M]. 上海：上海外语教育出版社，2002：153.
[7] Baker M. Towards a methodology for investigating the style of a literary translator [J]. Target, 12（2）：241-266.
[8] Scott M, Tribble C. Textual Patterns：Keyword and Corpus Analysis in Language Education[M]. Amsterdam： Benjamins, 2006.
[9] McEnery A M, Tono Y, Xiao Z. Corpus Based Language Studies[M]. London：Routledge, 2006.
[10]Nitschke S, Kidd E, Serratrice L. First language transfer and long-term structural priming in comprehension [J]. Language and Cognitive Processes, 2010, 25（1）： 94-114.
[11]杨梅. 中国学习者英语冠词替代误用研究[J]. 天津外国语学院学报，2011（1）：67-74.
[12]赵哲. 非英语专业大学生英语定冠词的使用：关联理论视角[J]. 外语教学，2010（1）：40-44.

［责任编辑　祁丽华］

（上接第49页）

象能否满足人的审美需要的属性，审美价值是人对现实的审美关系的对象性属性。如上所述，毛泽东在《关于正确处理人民内部矛盾的问题》和《在中国共产党全国宣传工作会议上的讲话》之中就已经完整地阐述了真善美的价值问题。我们完全可以在马克思主义创始人的原则基础上，在毛泽东的设想的基础上，来构建马克思主义文学批评中国形态的艺术价值论，继承和发扬中国传统美学思想的“美善相乐”、“比德说”等观点，把真、善、美统一起来，形成马克思主义文学批评中国形态的真善美统一的价值理论体系。

因此，马克思主义文学批评中国形态的理论形态大致上就是如上所述的艺术生产论、审美意识形态论、艺术掌握论、生活源泉论、艺术辩证法、艺术价值论这样几个部分。马克思主义文学批评中国形态只有在这样一些根本性的理论问题上弄清楚了，才可能进行正确的、科学的、有效的文学批评实践。

［参考文献］

[1] 马克思恩格斯全集：第42卷[M]. 北京：人民出版社，1979.
[2] 陆梅林. 马克思恩格斯论文学艺术（一）[M].北京：人民文学出版社，1982.
[3] 毛泽东论文艺[M]. 北京：人民文学出版社，1983.
[4] 马克思恩格斯全集：第46卷（上）[M]. 北京：人民出版社，1979.

［责任编辑　王艳芳］