ISSN 1009-5039
Overseas English 海外英语

zwwh@overseaen.com
http://www.overseaen.com
Tel:+86-551-5690811 5690812

皖南旅游资源翻译语料库的构建及其应用研究

安徽大学外语学院创新实验项目组

(安徽大学 外语学院,安徽 合肥 230601)

摘要:文章回顾了语料库的发展历程,基于对其发展的现状及国内翻译实践需要的分析指出了构建旅游资源翻译平行语料库的重要性。然后,以皖南地区主要旅游资源的文本材料为例,详细讨论了建库的过程,并对此语料库的应用途径做了初步探讨。

关键词:语料库;翻译;应用;皖南旅游

中图分类号:H315.9 文献标识码:A 文章编号:1009-5039(2011)09-0236-03

The Building and Application of the Translational Corpus of Tourism for Southern Anhui

Innovational Experimentation Project Team of SFS, Anhui University

(School of Foreign Studies, Anhui University, Hefei 230601, China)

Abstract: After reviewing the history of corpus, the importance in building a parallel corpus for tourism translation is explicated from the perspectives of the status quo of corpus and the translation need in China. Then how to build a translation corpus of tourism promotional materials centred on Southern Anhui is fully discussed as well as its various applications.

Key words: corpus; translation; application; tourism in Southern Anhui

计算机语料库诞生于20世纪60年代,以美国布朗大学的Nelson Francis和Henry Kucera建立的BROWN英语语料库为标志。此后,语料库蓬勃发展,数量不断增加,种类不断丰富。初期语料库多为单语语料库,主要用于语言研究。20世纪90年代初开始,多语种语料库开始发展起来,并且随着语料库用途的细分,特殊用途英语(English for Specific Purpose, ESP)模式的语料库逐步兴起。1995年,英国曼彻斯特大学科技学院翻译研究中心率先建立起了世界上第一个翻译语料库(Translational English Corpus,TEC)。计算机语料库对自然语言处理的不同方面(如话语识别、人机对话、信息提取、网页分类、机助翻译、文字处理等)的重要性和蕴藏的潜力,得到了国际计算语言学界的广泛认可[1];并且在语言研究的各个相关领域都得到了重视,催生了一门新的学科——语料库语言学,成为了语言学研究的一种方法和基础。

语料库在国内则起步较晚。1985年上海交通大学构建了JDEST学术英语语料库,成为中国语料库的先驱之作。此后建立的语料库多为单语语料库,用于语言研究和二语习得,例如中国英语学习者语料库、大学学习者口语语料库等,双语平行语料库基本仍在创建与开发之中。而且大型通用语料库和小型专用语料库之间发展尚不平衡。

在本例中,皖南地区的旅游资源(包括黄山、九华山、芜湖方特等)极具地方文化特色和多样性的特点。黄山是世界自然文化双遗产,有着丰富的物种资源和深厚的人文积淀,其中徽文化内容广博深邃,有整体系列性等特点;九华山乃四大佛教名山,历史悠久;芜湖方特主题多样,涉及面极广。这些特点都为这些宝贵旅游资源的翻译带来的了难题,造成景点介绍的翻译不够规范,文化现象的翻译不够准确传神等。然而,现有的大型通用语料库难以准确地体现旅游文本的文体特征、语篇功能等信息,无法满足旅游翻译研究与实践的检索统计需求。开发构建关于皖南地区旅游资源汉英翻译语料库(Translational Corpus for the Tourism of Southern Anhui,TranCoTSA),将大量高质量的实际应用中的汉英翻译语料经处理后整合起来,是一项有意义而先期未能实施的项目。

1 皖南旅游资源汉英翻译语料库的设计

杨惠中(2002:36)提出,语料库的设计和建设是在系统的理论语言学原则指导下进行的。特殊的皖南地区主要旅游资源汉英翻译语料库的构建,又因其特殊的应用目标和特定的语域,是一项复杂而艰辛的工作。因此,对其进行一个总体的设计和规划,是每项工作得以顺利进行的前提。具体而言,语料库的设计包括以下四个方面:

1)明确语料库的应用目标。一般地,语料库多应用于语言研究、翻译研究、教学研究和双语词典研编等[2]。本语料库的应用目标就是为旅游翻译工作者,尤其是皖南地区的旅游、宣传相关单位,在从事旅游翻译时提供翻译参考和依据,发挥优秀译作的借鉴作用;并为相关领域的语言和翻译研究提供重要的双语文本资源,客观准确地揭示旅游文本的文本功能和语体特征。

2)初步确定语料库的规模和收集范围。纵观语料库发展史,从初期的Brown语料库(100万词次)到目前的亿词级语料库,不同的规模适应了不同的需要。对待语料库的规模建库者应采取审慎的态度。Bowker & Pearson(2002:48)曾指出,千词级和十万词级的小型语料库在实践中被证明是可以满足特殊目的语言的研究需求的。本例立足小型专业语料库,根据特殊的应用目标,我们拟定TranCoTSA的规模为约50万词次。另一方面,语料的收集范围直接影响到语料的质量。确定收集范围,要保证语料的代表性、平衡性,以及本例中要求的一定的权威性。因此,我们将收集范围主要确定于皖南地区旅游部门的申遗报告、官方宣传资料、调查报告、公开出版的书籍等。

3)制定实施步骤。一般而言,其中包括语料的收集、预处

收稿日期:2011-04-25 修回日期:2011-06-20
基金项目:国家大学生创新性实验项目(项目编号:101035738)
小组成员:徐肖嵘,江婧婧,程思敏,张旸。

理、对齐、分析标注和建立 SQL 数据库等，下面一节将做详细说明。

4)统一技术标准。语料库中所有语料均需采用统一的技术标准，按照同样的方式进行编码或标记，以使得其能够独立于软件平台和具体的应用程序，具有较强的数据可交换性[3]。而保证技术方面的内部统一性，首先要确定的就是字符编码的问题。进行双语语料库建设的时候，中英文混排很容易出现乱码，为了实现最大的兼容性且最大程度避免乱码，建议采用国际通行的 Unicode 编码而不是 ASCII 编码或其他编码，同时必须确保中英文文本中分别使用纯粹的全角和半角字符，避免混杂。其次，为了确保语料库的可识别性和重复使用性，对语料进行科学系统的标注尤其重要。目前可供我们选择的标记方案有三种：TEI 文本编码标准、CES 语料库编码标准和自制方案。语料库建设者可以根据语料库的应用目标、规模等进行选择。

2 皖南旅游资源汉英翻译语料库的构建

1)语料收集。相对于单语语料来说，双语语料的获取，尤其是特殊语域双语语料的获取，则相对比较困难。在保证语料数量的同时，要注意以下问题：

首先，确保语言质量和翻译质量。语言质量和翻译质量决定了语料库辅助翻译价值的大小。官方双语报告、宣传材料和知名出版社出版的双语书籍经过了多次核对加工，以其作为语料来源有助于提高整体的语言质量和翻译质量。例如，我们收集获取的《黄山世界地质公园考察报告》就具有很高的代表性和权威性。

其次，进行适当抽样。由于小型专业语料库容量较小，易造成同一译者或同一出版社影响过大的情况，使得语料库失去参考意义[4]，因此需对不同来源的语料进行抽样，以保持不同译者和出版社的作品在语料库中的平衡。

2)语料预处理。在这一阶段，需要将不同载体不同格式的文本转换成统一格式的生语料文本。在收集的资料中，纸质文档需通过扫描仪录入计算机，并用 OCR 字符识别软件转换为可编辑的电子文档；同时，不同来源的电子文档中会混杂着很多多余空格、换行符和乱码等不必要的字符，也都需要在这一过程中对其进行消除、整理。虽说这部分工作繁琐、细致，大部分需依靠人工完成，但可以运用一些方法简化操作，尽量减少工作量。例如，利用以下一段 VBA 代码就能在 Word 中轻松实现选中文字中多余换行符的批量删除：

```
With Selection.Find
.Text = "^p"
.Replacement.Text = ""
.Forward = True
.Wrap = wdFindStop
End With
Selection.Find.Execute Replace:=wdReplaceAll
```

3)语料的标注和对齐。经过预处理的干净的生语料文本需要用预先规定的某种符号系统添加进适量的人工信息，才能被计算机程序识别、应用，并为语料库使用者提供所需的语言信息。语料的对齐、标注过程是一个十分艰辛复杂的工程，其中涉及语言的结构分析、标注体系的设计和计算机自动处理技术的应用等。根据不同的标注层面，王克非(2004:21)提出目前语料库的标注主要包括：中文分词、文本结构及文本来源、词性标注、句子结构标注和语义标注等。考虑到本例中 TranCoTSA 的特殊应用目标，无需对其进行词性、语义等语言信息的标注，但需要对中英文本进行分句对齐信息的标记。

国内外学者对自动双语语料对齐进行了大量的研究，如 Brown[5]、Gale & Church[6]、Chen[7]等都提出了相对有效的算法。经过多年的发展，算法不断优化的自动对齐技术在某些特殊文本的应用中已经可以达到较高的精度，如法律文本。但在本例中，旅游翻译由于面向对象的不同而导致了表达重点的不同，翻译中经常有省略信息和添加信息的现象，译文十分灵活，双语对应不严格。经过试验，自动对齐精度较低，需人工仔细核对或完全由人工进行对齐工作。

如前文所述，对齐后需对文本加以标记，以方便计算机识别。我们参照现有的标记体系开发了一个新的标记体系。这个标记体系基于目前普遍应用的 XML 语言，用一系列嵌套的标签标记文本，以期获得广泛的软件支持和跨平台支持，并且可以根据需求自由扩展。我们制定的标记集如表 1 所示：

表 1

标记内容	标记
标题	<title>...</title>
作者	<author>...</author>
语言	<language>…</language>
来源	<source>...</source>
句子	<s s_id="序号">…</s>
段落	<para p_id="序号">…</para>
未译信息	<miss>...</miss>
修订	<ttr>...</ttr>

以下是一个实例。首先定义了一个文档类型定义(DTD)：

```
<?xml version="1.0" encoding="unicode"?>
<! ELEMENT article (header, paras)>
<! ELEMENT header (title, author, language, source)>
<! ELEMENT title (#PCDATA)>
<! ELEMENT author (#PCDATA)>
<! ELEMENT language (#PCDATA)>
<! ELEMENT source (#PCDATA)>
<! ELEMENT paras (para+)>
<! ATTLIST para
p_id CDATA #REQUIRED>
<! ELEMENT para (s+)>
<! ELEMENT s (#PCDATA)>
<! ATTLIST s
s_id CDATA #REQUIRED>
```

并且在 XML 文档中加入引用创建好的外部 DTD 的语句：

```
<?xml version="1.0"?>
<! DOCTYPE article SYSTEM "sample.dtd">
<article>
<header>
  <title>...</title>
  <author>...</author>
  <language>English</language>
  <source>...</source>
</header>
<paras>
  <para p_id="01">
      <s s_id="01">sentence 1 in paragraph 1</s>
      <s s_id="02">sentence 2 in paragraph 1</s>
      <s s_id="03">sentence 3 in paragraph 1</s>
  </para>
  <para p_id="02">
      <s s_id="01">sentence 1 in paragraph 2</s>
      <s s_id="02">sentence 2 in paragraph 2</s>
      <s s_id="03">sentence 3 in paragraph 2</s>
  </para>
</paras>
</article>
```

3 皖南旅游资源汉英翻译语料库的应用途径

1)辅助翻译。在 TranCoTSA 相应的 Web 检索平台上输入

关键词，即可获取包含关键词的所有双语对译的句对和相关语篇信息，进行皖南旅游翻译知识的抽取。语料检索既可为译者提供不同来源的词句翻译参考，也可以帮助译者理解旅游翻译的结构特征和语体特征。此外，将 TranCoTSA 作为基于统计的和基于实例的机器翻译或机助翻译的支撑数据库，可以大大提高机器翻译或机助翻译在旅游翻译领域的效率。

2）对比研究。目前比较著名的语料库研究工具包括 Mike Scott 开发的 WordSmith 和 Michael Barlow 设计的 ParaConc。我们可以利用此类软件在语料库中的检索统计，对比分析英汉旅游文本在文体和功能上的特征、差异及共性，并基于上述对比分析的结果，结合现代翻译学理论，集中探讨汉语旅游文本翻译的基本原则和方法[8]，尤其是旅游文本中大量文化词语对于深入研究翻译中的异化和归化现象意义重大。

3）翻译教学。国内的语料库研究最初与外语教学联系密切，如最早建立的 JDEST 语料库，其最初目的就是为语言教学提供有关学习者语言运用和典型困难的可靠信息[9]。本例中我们可以基于 TranCoTSA，利用索引软件，共现动态语境，或进行文本等值概率的分析、译文风格特点的量化分析等，为实现翻译教学课程结构的科学化和规范化提供了保证；其次，利用计算机强大的功能进行快速、准确和复杂的检索分析可以实现教学理念的现代化，有利于培养学生的学习能力和创新精神[10]。

4 结束语

基于语料库的翻译实践和语言研究日益受到了国内外学者的关注，其前提和基础在于各类型高质量的语料库的建立。本文从理论到实践，完成了关于皖南地区主要旅游资源翻译语料库的构建，填补了该特殊语域语料库建设的空白。然而，在研究过程中我们发现，语料库的发展仍存在很多新的问题有待解决，如双语文本句级对齐算法的革新，语义、句法标注技术的完善等。但不可否认的是，随着语言研究的深入和计算机技术的发展，语料库因其定性与定量相结合的独特优势，必然拥有着广阔的前景。正如王克非（2004：15）所述，我们有理由相信，准确度更高、更为成熟的词性标注、句法分析和语篇自动分析技术将会不断得以应用；正确地进行语料分析将使我们受益良多，也终将为语言学和其他相关领域的研究开辟新的天地。

参考文献：

[1] 王建新.计算机语料库的建设与应用[M].北京：清华大学出版社，2005.

[2] 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.

[3] 常宝宝，詹卫东，张华瑞.面向汉英机器翻译的双语语料库的建设及其管理[J].术语标准化与信息技术，2003(1)：28-31.

[4] 麻丽莉，王祥兵.军事平行语料库的建立及其在军事翻译方面的应用[J].国防科技，2009，30(1)：38-41.

[5] Brown P F，Lai J C，Mercer R L.Aligning Sentences in Parallel Corpora [C]//Proc. of the 29th Annual Meeting of the ACL，1991.

[6] Gale W A，Church K W.A Program for Aligning Sentences in Bilingual Corpora[J].Computational Linguistics，1993，19(1).

[7] Chen S F.Aligning Sentences in Bilingual Corpora Using Lexical Information [C]//Proc. of the 31th Annual Meeting of the ACL，1993.

[8] 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，2010，26(4)：115-117.

[9] 邢富坤.语料库：值得教育技术学关注的新型学习资源[J].解放军外国语学院学报，2006(2)：65-68.

[10] 于连江.基于语料库的翻译教学研究[J].外语电化教学，2004，96：40-44.

（上接第 235 页）

How I Met Your Mother《老爸老妈浪漫史》

Desperate Housewives《绝望主妇》/《欲乱绝情妻》

The Secret Life of the American Teenager《青春密语》

Dirty Sexy Money《黑金家族》

The Big Bang Theory《生活大爆炸》

Women's Murder Club《灭罪红颜》

一些字幕组的翻译桥段，一时间更成为网上各大论坛里美剧迷津津乐道的话题。

比如《越狱》里面的名句"Preparation can only take you so far."准备的作用是有限的被翻译成了"谋事在人，成事在天"；"Be the change you want to see in the world."——"欲变世界，先变自身"。

又如美剧 Heroes《英雄》中开始旁白部分的翻译，则充满东方哲学归属气息：

And now it's up to us to figure it all out.

（现在该由我们自己判断。）

时至今日，人之为人，自斟自酌。

To decide what drives our actions. In each of us is the capacity.

（每一个人都有能力决定驱使我们行动的目的。）

决断之力，发自人心，显于外表。

While others know only self-interest? Isolating themselves in a world of their own making?

（而有些人却自私自利，把自我封闭在孤立无援的世界。）

自私之人，与世隔绝，独处一己幻境。

Some seek only love, even if unrequited, While others are driven by fear and betrayal.

（有人只追寻爱，即便没有回报，有些人却被恐惧和背叛驱使着。）

求爱之人，不计回报；沦陷之人，惧与叛者。

还有些字幕翻译结合当今网络流行语言，如 "I wasn't even listening. "被翻译成"我只是路过来打酱油的。" "I'm not a pervert"直接翻译成"我又不是'陈冠希'。"在博取欢乐的同时，赢得观众认可。

3 结束语

归化策略是语言与跨文化交际活动紧密结合的产物。由于文化背景的不同，翻译应使原文更"本土化"。受版权、商业利益等因素的影响，字幕组的存在和译品还存在诸多争议。如何使影视剧翻译雅俗共赏，还需要翻译研究人员不断探索和实践。

参考文献：

[1] Nida，EugeneA.Towards a Science of Translating[M].Leiden：E，J.Brill，1964：159.

[2] 郁文.从归化和异化理论看英文电影片名的翻译[J].山东水利职业技术学院院刊，2008(3).

[3] 贺瑛.东西文化背景下英文电影名的归化翻译策略[J].考试周刊，2008(32).

[4] 毛发生.两岸三地外语影片片名的翻译比较[J].西安外国语学院学报，2002(12).