利用关键词分析旅游文本的文体特征：一项基于语料库的实证研究

菏泽学院　侯晋荣　　曲阜师范大学　秦洪武

摘要：语料库语言学为翻译研究开辟了新的研究方式，推动翻译研究由规范性向描述性转变。本文基于类比语料库对比分析了英语原创旅游文本和汉译英旅游文本的语言特征，描述了二者在词汇、句法、篇章上的差异，并借此对旅游文本的汉英翻译提出一些建议。研究认为，旅游文本的汉英翻译在时间表达、主位信息的处理上应符合目的语特征并且在文体上突出召唤性和口语化特征。

关键词：旅游文本；原创文本；翻译文本

作者简介：侯晋荣，讲师，主要从事语料库语言学与翻译研究。电子邮箱：helenhouhou@126.com　秦洪武，教授，主要从事语料库语言学与翻译研究。电子邮箱：qinhongwu@163.com

1　引言

自20世纪90年代以来，语料库开始越来越广泛地应用于翻译研究。根据贝克（Baker，1993）的研究，翻译文本中存在显化（explicitation）、简化（simplification）和规范化（normalization）等共性。这些翻译共性得到了许多学者的支持。拉维奥茨（Laviosa，1998）发现了翻译文本句子短、词汇密度低且词汇重复率高；奥洛汉（Olohan，2001）则发现翻译文本的句法特征较为明显，如翻译文本中补语指示词“that”较多。与以往的定性研究方法不同，基于语料库的研究重视语言数据，重视定量分析，而且也不排斥定性分析。这就使得基于语料库的语言研究具有其他方法无法替代的优势。本文的研究也基于语料库，即英语旅游文本类比语料库（以下称CCETT）[①]，比较分析英语原创旅游文本和汉译英旅游文本（以下简称原创文本和翻译文本）在词汇、句法、篇章等方面的特征。

2　旅游文本类比语料库

为本项研究而创建的旅游文本类比语料库库容约12万词，由原创文本和翻译文本两个子库构成，每个子库包括人文和自然两个部分，各部分比例大致相等（见表1）。库中语料全部来自英美和中国旅游官方宣传网站，可以保证其代表性和准确性。

	原创文本		翻译文本	
	人文	自然	人文	自然
文件数	69	73	56	72
字数	30017	30293	30164	30196
比率	49.77%	50.23%	49.97%	50.03%
总计	60310		60360	

表1：CCETT语料构成

为满足不同的研究目的，旅游文本分别以生语语料和标注语料存放。一方面，旅游文本在文本清洁后，以UTF-8编码形式单独存放，组成生语语料库。另一方面，文本采用英国兰卡斯特大学研发的“CLAWS”标注软件进行词性标注，组成标注语料库。

研究使用的参照语料来自美国国家语料库（ANC）中命名为“berlitz”的旅游文本子库[②]。该库含有179个文件，约100万词，汇聚了旅游手册中的文本信息，介绍旅游景点的人文历史、景点概要、参观时间、出行信息等，主体内容与自建旅游文本类比语料库一致。库中文本采用宾夕法尼亚大学的标注集“Penn tagset”进行词性标注。为便于统计和对比，我们对该库语料用“CLAWS”重新进行了标注，作为参照语料库使用。

3　数据统计与分析

3.1　基本数据

标准化类符/形符比、词长、句长是最能体现文本特征的三个参数。标准化类符/形符比可以体现不同长度的文本中词汇量的多少；词长可以反映词汇使用的难易程度；句长可以体现篇章的难易程度。这三个参数可利用“WordSmith Tools”对文本

进行批量分析获取。表2是对这三个参数的描述。

文本 / 数据	翻译文本	原创文本	参照文本
标准化类形符比	43.70	44.75	48.40
词长	4.71	4.84	4.81
句长	20.06	19.64	16.84

表2：CCETT基本数据统计

表2显示，翻译文本标准化类符/形符比最低，平均词长最低，平均句长最高。标准化类形符比低说明翻译文本的词汇使用不如原创文本和参照文本丰富。平均词长最低，说明翻译文本中词汇使用相对简单，多使用短小词汇。原创文本和参照文本平均词长相差不大，仅0.03，说明二者在词汇使用难易程度上相差不大。平均句长翻译文本最高，参照文本最低，二者相差3.22词，翻译文本与原创文本平均句长相差0.42词。翻译文本句子最长，说明在翻译过程中存在句子扩张的趋势。

对比分析显示，原创文本和参照文本凸显相同的特征：标准化类符/形符比较高；平均词长高，平均句长低。而翻译文本呈现与之相反的特征：标准化类符/形符比较低，平均词长低，平均句长高，这些数据说明翻译文本具有简化、显化的特征。

3.2 高频词汇对比分析

拉维奥茨把高频词汇（high frequency words）定义为在语料库中词汇出现频率超过0.10%的所有词汇（1998）。高频词汇累计率反映文本中词汇使用重复率和词汇使用丰富程度。表3汇总了各个文本中高频词汇累计频次和累计比率。

翻译文本		原创文本		参照文本	
频次	频率	频次	频率	频次	频率
29558	48.97%	27987	46.41%	442516	43.01%

表3：CCETT高频词汇累计率

翻译文本高频词汇比率最高，原创文本次之，参照文本最低。翻译文本中高频词汇重复率高，说明翻译文本的词汇不如原创文本和参照文本丰富，这是翻译文本

简化特征的具体体现。

	翻译文本			原创文本			参照文本		
次序	词	频次	频率	词	频次	频率	词	频次	频率
1	the	5375	8.90%	the	5232	8.68%	the	88342	8.58%
2	of	2486	4.12%	of	2156	3.57%	of	39309	3.82%
3	and	2054	3.40%	and	1808	3.00%	and	33083	3.21%
4	in	1422	2.36%	a	1275	2.11%	a	23864	2.32%
5	a	1226	2.03%	to	1217	2.02%	in	20933	2.03%
6	is	1129	1.87%	in	1146	1.90%	to	20505	1.99%
7	to	966	1.60%	is	806	1.34%	is	13521	1.31%
8	it	470	0.78%	s	603	1.00%	s	9121	0.89%
9	are	455	0.75%	for	501	0.83%	for	8348	0.81%
10	with	451	0.75%	was	457	0.76%	on	7888	0.77%

表4：CCETT高频词汇前10位统计

为便于观察高频词汇的具体使用情况，表4列出高频词汇前10位的使用频次和频率。三个文本中出现频次最高的三个词都是“the”, “of ”, “and”。除翻译文本中“it”之外，各个文本中出现频次最高的词汇都是定冠词、不定冠词、介词、连词和系动词。高频词汇统计可以呈现文本的总体特征，但词汇使用的具体细节需进行关键词对比分析。

3.3　关键词对比分析

在“WordSmith”的“Wordlist”中，设定翻译文本为观察库，原创文本为参照库，在P<0.001（具有显著统计学意义）水平上利用对数似然率计算关键值。鉴于人文景点和自然景点在表述时存在一定词汇差异，关键词的对比将分别进行，对比结果分别在表5、表6中呈现。表中正值表示词汇较多地用于翻译文本，负值表示词汇较多地用于原创文本。

	关键词	频次 翻译文本	%	频次 原创文本	%	关键值	P
背景	dynasty	96	0.32	1	/	123.02	P<0.001
	culture	32	0.11	5	0.02	22.13	P<0.001
指称	his/him/he	185	0.61	58	0.19	69.17	P<0.001
	we	14	0.05	1	/	13.39	P<0.001
	visitor(s) / guest(s)	31	0.10	67	0.22	−13.72	P<0.001
	you	30	0.10	123	0.41	−61.26	P<0.001
连词	and	1032	3.42	805	2.68	30.16	P<0.001
缩写	’t	4	0.01	20	0.07	−11.57	P<0.001
	’ll	5	0.02	25	0.08	−14.66	P<0.001
	’s	198	0.66	329	1.10	−33.84	P<0.001

表5：CCETT人文景点关键词对比结果

	关键词	频次 翻译文本	%	频次 原创文本	%	关键值	P
背景	dynasty	55	0.18	0	/	76.47	P<0.001
	culture	26	0.09	5	0.02	15.66	P<0.001
指称	he/his/him	89	0.29	18	0.06	49.47	P<0.001
	you/your	158	0.52	268	0.88	−28.38	P<0.001
	visitor(s) / guest(s)	23	0.08	96	0.32	−47.89	P<0.001
动词	called	48	0.16	15	0.05	18.30	P<0.001
	offer	3	/	25	0.08	−19.83	P<0.001
	get	5	0.02	30	0.10	−19.90	P<0.001
系动词缩写	is/was	780	2.58	522	1.73	52.30	P<0.001
	’ve	0	/	9	0.03	−12.51	P<0.001
	’t	8	0.03	34	0.11	−17.42	P<0.001
	’re	3	/	24	0.08	−18.67	P<0.001
	’ll	0	/	34	0.11	−47.26	P<0.001
	’s	117	0.39	274	0.9	−65.77	P<0.001

表6：CCETT自然景点关键词对比结果

3.3.1 时间表达对比

在人文景点和自然景点关于历史背景的描述中，“dynasty”和“culture”的使用最为显著。尤其是“dynasty”，成为翻译文本与原创文本对比最突出的词汇（人文景点关键值123.02；自然景点关键值76.47）。汉语中，存在着以历史朝代表示时间的现象，这也在旅游文本中体现出来，并在翻译文本中留下痕迹。

词汇在文本中不是孤立的，词与词之间存在“相互期待”（mutual expectancy）。词汇的这种组合关系（syntagmatic relation）需从搭配和类连接来阐述。搭配指文本中两个或多个词汇的共现（Sinclair，1996：79）。类连接指词汇在语法范畴上的相互关系（Firth，1957：12），是表示意义的实词与表示语法关系的虚词之间的连接模式，也是词汇构成篇章的必要条件。

“Antconc3.2”[③]的排列检索（collocate）可以呈现关键词左右的搭配。以“dynasty”为节点词，设定左边四个单词的检索范围，检索发现与节点词组合的词汇出现频次从高到低为“the”（126），“of”（46），“in”（29），“from”（22），“during”（16），“early”（7），“late”（5）。由此可以推断翻译文本中“in/from/during the ×dynasty” 或者“in the early/late of × dynasty” 经常出现，如例1：

1. The history of Buddhism at Jiuhua Mountain can be divided into five periods: founded in the middle of Tang Dynasty ; decayed from the late Tang Dynasty to the Five Dynasties; slowly developed in the Song and Yuan Dynasties; and greatly developed in the Ming and Qing Dynasties.（翻译文本）

原创文本中，“dynasty”仅出现一次，毫无疑问，在时间表述上，原创文本与翻译文本存在很大差异。检索发现，“century”在原创文本中出现55次，翻译文本中出现38次。二者虽然绝对数量差距不大，但在“century”左右的词汇搭配上稍有不同。原创文本中，除了“×th”（序数词），“the”和“of ”，“century”左侧出现频次较高的词汇是“early”，“turn”，“end”。而翻译文本中除了“×th”，“the”和“of”，仅“early”出现4次。这说明在使用“century”进行时间表达时，原创文本的表达形式较为丰富，如例2中的 “around the turn of ×century”，而翻译文本多用例3中的表达“in the ×th century”。 参照文本中，“dynasty”出现249次，其中29次指西方君主制时期的“王朝”，其余220次都是对收录的中国景点进行时间描述。“century”出现2521次，与“dynasty”出现的比例为11.5:1，并且“century”的搭配方式也较为灵活。

2. The British removed the wall around the turn of the 18th century.（原创文本）

3. It was first built in the 7th century B. C. when China was still divided into many small states.（翻译文本）

4. Desperate for work, perhaps 20 percent of the Bahamian population left to

take construction jobs in Florida between the turn of the century and World War I.（参照文本）

在时间表达上，翻译文本受汉语源语历史文化背景的影响，频繁使用朝代作为时间标记。对缺乏中国历史背景的目标读者而言，这种表达方式使翻译文本不易理解。经检索，原创文本和参照文本多使用公元纪年或者“century”来表达时间，表达方式简单统一，容易理解。

3.3.2 指称对比

汉语倾向于零回指（Li & Thompson，1979），也就是说汉语中的人称代词较少。翻译文本中第三人称代词“he/his/him”明显使用过多（人文景点关键值69.17，自然景点关键值49.47），这说明在翻译过程中，添加了人称代词，夸大了目标语特征。

5. The 5.5 hectares mausoleum includes three giant yurt halls which house coffins of the Khan, his① wife, his② son and his③ generals.（翻译文本）

例5中，位置②和③通过添加“his”回指“Khan”，使指称变得明确。翻译文本中人称代词过多使用一方面是由于回指引起，另一方面是由于篇章的叙事性决定的。回指的明示导致人称代词相对数量增加，而篇章的叙事性决定了人称代词的使用频率相对较高。

原创文本和参照文本注重加强与读者的交流，循循善诱，极少有大量叙事的成分，而以诱导型话语居多。“you”和“visitors/guests”是原创文本和参照文本中最显著的词汇。“you”设置面对面交流的场景，瞬间缩短与读者的心理距离，是以读者为中心的体现。“visitors/guests”给游客提供清晰明确的指示，对旅游活动有极大的指引作用。

6. It is one o’clock in Seattle. You are walking down the street. A lot of people are outside for lunch. You see a woman. She is holding a white and green paper cup. You see another person. He is also holding a cup in his hands. You see another and another. Everyone has a cup in their hands! What are they all drinking? You smell the Seattle air. It’s coffee!（原创文本）

7. Visitors can climb the stairs to a hot, enclosed lookout.（参照文本）

例6中通过“you”把一连串的动词“walk down”, “see”, “smell”联系在一起，呈现出一幅动感的画面，令游客如临其境。例7中有明确的指导性语言，指出游客可登上台阶把景观看得更清楚。

在指称方面，翻译文本中第三人称代词较多，体现了回指的显化和文本的叙事特征。原创文本和参照文本中第二人称代词和名词较多，体现了文本的指引特征，召

唤性较强。

3.3.3 连接词对比

关键词（表5）对比分析显示，“and”在翻译文本人文景点的描述中较多使用。“And”可以连接两个并列的名词、形容词、动词、不定式或者句子，对文章的衔接起着不可替代的作用。

8. Follow park signs to the harbor and①then to the visitor center on Spinnaker Drive. Get oriented here and②then go to the nearby Island Packers office and③inquire about boat schedules to the islands.（原创文本）

9. Whatever their origins, their culture evolved under the pressure and④ influence of foreign forces.（参照文本）

10. They wear helmets and⑤ armor and⑥ carry real bows and⑦ arrows, swords, lances, javelins and⑧ crossbows in their hands.（翻译文本）

11. And⑨ farther behind, high up on the mountain stands a screen of five peaks coloured by green trees and⑩ bamboos and⑪ marked by serene valleys and⑫ rocks of pleasing shapes.（翻译文本）

原创文本和参照文本提取的例8和例9中，“and”①连接两个动词不定式，②和③连接动词，④连接两个名词。“and”能够有效衔接句子成分，使动作连贯，或者附加信息。翻译文本提取的例10和例11中，“and”出现频次很高，⑤⑦⑧⑩连接名词，⑪连接动词，⑨用在句子开头，表示连续性。值得注意的是例10和例11中，连接名词和连接动词的“and”混杂在一起频繁使用，引起理解困难。究其原因是由于汉语源语引起的，汉语缺乏语法衔接手段，多依赖意义将句子连在一起，句子结构松散。汉语文本在英译过程中，这些松散的句子结构，尤其是并列结构需要连词衔接在一起，“and”使用就明显过多。

3.3.4 动词对比

自然景点关键词对比显示，翻译文本和原创文本在动词使用上有所不同。翻译文本多用“called”，原创文本多用“offer”和“get”。经检索，翻译文本中“called”出现48次，其中46次后面接有专有名词，如例12所示。也就是说，翻译文本中多使用被动语态“be called”来介绍景点名称。相比之下，原创文本中，景点名称常常置于句首主位，突出主题信息，如例13。

12. Founded in 1924, the area was first called Phoenix Mountain Park.（翻译文本）

13. Hanawi Falls is a popular stop on the Road to Hana on Maui's northern

shoreline.（原创文本）

“offer”在两个文本中出现频次相差悬殊，原创文本出现25次，翻译文本中仅出现3次。根据柯林斯高级英汉双解辞典（2009：763），“offer”最常用的意项为“提供（某物给某人）”。

14. Sellicks Beach in the south our pristine coastline offers abundant choice for the beach-lovers.（原创文本）

15. Longji’s terraced fields offer some of the most fantastic scenery in Guilin or indeed China.（翻译文本）

例14 “塞利克斯沙滩给沙滩爱好者们提供了足够的选择”。“offer”意为“提供（某物给某人）”，把景点与游客联系起来。例15“龙脊梯田呈现了广西境内乃至中国境内的奇异风景”。“offer”在这里意为“提供（服务、产品）”，其实翻译为“呈现”更为确切，表示的是景点一种单向的展示。根据柯林斯高级英汉双解辞典，这个义项并不常用，排在第七位。原创文本体现了景点与目标读者之间的互动，而翻译文本忽视了这一点，仅实现了景点与目标读者的单向信息传递。

原创文本中“get”出现30次，翻译文本中仅5次，频次相差6倍。与“go”和 “make”一样，“get”在英语中使用频繁，意义繁多，在旅游文本中主要指“获得”、“到达”。

16. When you are standing at its feet, and you get a great view of the Manhattan skyline from the island.（原创文本）

17. It takes about forty minutes to get to it.（原创文本）

18. If you take a bamboo raft trip fleeting through the river, you will get quite a view of the water and the mountain…（翻译文本）

19. As the field is not very high, tourists can easily get into the orchard to look around.（翻译文本）

例16至例19中，“get”在原创文本与翻译文本中的意思一致，使用“when you… you get”和“if you … you get”这样的表达方式为游客提供建议，介绍游客可以看到怎样的景色。“get”的另外一个意思是到达，主要也用于提供建议，游客如何到达，花费多少时间到达目的地。“get”在原创文本中使用较多而翻译文本中寥寥无几，与原创文本建议较多、翻译文本缺乏旅游建议有关。

3.3.5 正式程度对比

文本的正式程度可分成五个级别：僵化、正式、中性、非正式、随便（Quirk，1985）。关键词对比显示原创文本中多用“’t”, “’ll”等缩写，而翻译文本中缩写较少。系动词、助动词的缩写使文本偏向非正式，可以增加读者的阅读速度，便于获取有

效信息。

20. If you’re lucky, you'll see the occasional native deer grazing the meager grass that grows in small patches on the trail.（原创文本）

21. In summer, the traffic isn’t quite as appalling as on Cape Cod, but that’s not saying much.（参照文本）

22. Do not bathe in the hot spring if you do not feel very well or when you are hungry.（翻译文本）

例20使用“you’re”, “you’ll”使行文变得活泼易懂，文本倾向于非正式文体。例21同样使用缩写“isn’t”, “that’s”便于目标读者的注意力集中在主要信息上。例22中“do”与“not”分开书写，突出强调“not”，有明令禁止的意思，口气强硬，文本偏向正式文体。

原创文本和参照文本从目标读者的需求出发，在较短的阅读时间内传递尽可能多的信息。系动词助动词的缩写使目标读者把阅读重点放于主要信息获取上，文本简洁易读，符合旅游文本的文体特征。翻译文本缩写使用较少，文体较为正式，在一定程度上影响目标读者的阅读速度和旅游信息的传递。

4　结语

基于语料库的数据分析显示，旅游翻译文本呈现出显化和简化特征，且有扩张目的语某些语言特征的倾向。原创文本与翻译文本在词汇、句法、篇章等方面有较大差异。

词汇方面，关键词对比分析显示原创文本和翻译文本在时间表达上使用的词汇有较大差异。原创文本多用公元纪年和“century”来指代时间，简单易懂。翻译文本频频使用“dynasty”表述时间，可能会引起外籍目标读者时间理解的困难。在翻译过程中，译者应考虑不同目标读者的接受视野，在汉译英的过程中转换表达方式，使用更通行的时间表达方式。

句法方面，翻译文本中连词“and”使用过多，这是由于英汉衔接方式不同，导致译者过度使用该衔接手段，出现了不同于原创和目的语的“第三代码”（Frawley, 1984；Øverås, 1998）。另外，翻译文本多用被动“be called”引出景点名称，导致翻译文本被动语态使用较多。相比之下，原创文本中景点多出现在开头位置，突出主位信息。旅游文本汉译英过程中，译者应关注汉英句法的差异，使译文更符合目标语规范。

篇章方面，指称对比显示，翻译文本多用第三人称，还保留着明显的叙事风格；原创文本多用第二人称，意见建议较多，偏于指导风格。另外，翻译文本语言较为正式，偏向正式文本，原创文本系动词、助动词缩写较多，偏向非正式文本。旅游文本汉译英过程中，译者应考虑旅游文本的文体特征，突出其召唤性和口语化特征。

注释

① CCETT的全称是Comparable Corpus of English Tourism Texts。

② ANC (American National Corpus) 美国国家语料库约收录1.4亿词，包含口语和笔语两部分，其中笔语语料收录政治、旅游、科技、小说、信件和杂志等。旅游宣传文本以“berlitz”文件夹命名，语料可从网站http://www.americannationalcorpus.org/OANC/index.html下载。

③ “Antconc”是免费的语料库工具软件，由日本学者劳伦斯·安东尼 (Laurence Anthony) 开发，具有词语检索、生成词表和主题词三大功能。

参考文献

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In M. Baker, *et al.*(eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 233—252.

Firth, J. R. 1957. *Papers in Linguistics*. London: Oxford University Press.

Frawley, W. 1984. Prolegomenon to a theory of translation. In W. Frawley (ed.), *Translation: Literary, Linguistic and Philosophical Perspectives*. London: Associated University Press, 159—175.

Laviosa, S. 1998. The corpus-based approach: A new paradigm in translation studies. *Meta* 43 (4) : 474—479.

Li, C. N. & S. A. Thompson. 1979. Third-person pronouns and zero anaphora in Chinese discourse. In T. Givon (ed.), *Syntax and Semantics, Vol. 12*. New York: Academic Press, 311—335.

Olohan, M. 2001. Spelling out the optionals in translation: A corpus study. *UCREL Technical Papers* 13: 423—432.

Øverås, L. 1998. In search of the third code: An investigation of norms in literary translation. *Meta* 43 (4) : 557—570.

Quirk, *et al*. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair, J. M. 1996. The search for units of meaning. *Textus* 9 (1) : 75—106.

Sinclair, *et al*. 2009. *Collins COBUILD Advanced Dictionary of English*. Beijing: Higher Education Press.