译文分析的语料库途径*

王家义
(湖南工程学院,湘潭 411104)

提 要: 译文分析的语料库途径是通过融合定量研究和定性研究,用特定的文本分析软件对翻译文本进行词汇、句法、语篇和修辞等层面的实证分析。本文探讨基于语料库的译文分析的可行性和实现途径,并以《茶馆》的英若诚译文和霍华译文为语料,对比分析两译文的用词特征。

关键词: 译文分析;语料库;用词特征
中图分类号: H315.9 文献标识码: A 文章编号: 1000-0100(2011)01-0128-4

Corpus-based Approach to Translation Analysis

Wang Jia-yi
(Hunan Institute of Engineering, Xiangtan 411104, China)

Integrating quantitative and qualitative methods and using particular software, corpus-based approach to translation analysis investigates translation versions across lexical level, syntactical level, textual level and rhetorical level. The thesis explores the feasibility and realization of corpus-based approach to translation analysis. A case study is carried out to investigate the lexical features of the two versions of *Cha Guan*.

Key words: translation analysis; corpus; lexical feature

1 基于语料库的翻译研究

语料库翻译研究是 20世纪 90年代以来兴起的翻译研究方法,其根本思想是用语料库语言学的工具、技术和方法对大量真实的翻译现象进行描述并从所描述的翻译"自身"的语言特征中寻找翻译现象固有的规律性特征。(胡显耀 2007: 214)语料库翻译研究已成为当今描述翻译研究领域中一种新的研究范式,在理论、描写和应用等层面对翻译研究和翻译教学以及翻译培训起着越来越重要的作用,并激发了众多学者对相关问题进行研究的浓厚兴趣。Mona Baker提出基于语料库的翻译普遍特征: 简略化、明晰化、规范化和平整化。之后,围绕这一主题的研究大量涌现,主要有 Baker(1993, 1996), Kenny(1998, 2001), Munday(1998), Laviosa(2002), Olohan(2004)等。他们纷纷检验 Baker的假设;以共时语料为研究对象,依靠计算机技术分析数据,如平均句长、类符形符比、词汇密度等,考察词汇多样性、信息负载等;把理论阐释和实证研究相结合,探讨翻译文本的特征。

近年来,国内基于语料库的翻译研究文章的发表量呈现快速增长趋势,语料库作为一种研究方法逐渐获得大家的青睐。这些研究主要集中在: (1)语料库与翻译研究综述和介绍,如廖七一(2000),丁树德(2001),刘康龙、穆雷(2006),王克非、秦洪武(2009); (2)语料库与翻译教学,如王克非(2004)、倪传斌(2005)、秦洪武(2007); (3)语料库与翻译普遍性,如黄立波、王克非(2006),吴昂、黄立波(2006); (4)语料库的建立和运用,如王克非(2004)、何莲珍(2007)。与国外研究相比,国内基于语料库的翻译研究中介绍性或评价性和建库方案的文章较多,实证性翻译研究较少。根据刘康龙、穆雷对 14类期刊的统计,只有 4篇论文,只占此类研究的 21% (刘康龙 穆雷 2006: 61)。"这个数字跟语料库作为实证研究的工具相比是极不相称的。"(Ibid.)语料库与翻译研究在国内还处于介绍和评价时期,基于语料库的翻译研究在各个层面都有待加强,译文分析的实证研究值得尝试。

2 基于语料库译文分析的实现

基于语料库的译文分析是采用定量和定性相结合的

* 本文系湖南省教育厅人文社科项目"基于语料库的译者风格研究"(08C214)的阶段性成果。

分析方法, 用特定的文本分析软件对翻译文本进行词汇、句法、语篇和修辞等层面的实证分析。

语料是语言分析的基础。目前, 国内与翻译相关的语料库有北京外国语大学通用汉英平行语料库, 南京大学英汉名著翻译语料库 (NUCECC), 北京大学计算语言研究所、清华大学智能技术国家重点实验室和中国科学院计算技术研究所共同开发的"面向新闻领域的汉英翻译系统"等。但大众能接触到的则很少, 无法满足语言学习与研究的需要。自建小型语料库因其内容更具针对性、即时性和新颖性而日渐受到语言学习者与研究者的重视。小型语料库的创建包括语料库的设计、语料的收集和语料库的预加工等过程。在创建自己的语料库前, 首先应根据该语料库的用途确定原则和方案。语料的收集主要有两种方式: 一是通过光电扫描或键盘输入制作电子文本; 二是利用网络上已有的电子文本, 将其转化为需要的格式。语料库的预加工主要包括语料的标识和语料的赋码。收集好的语料还要清除杂质和多余符号, 并统一语料的格式和存放方式。语料最好是每一个文本作为一个独立文件单独存放。这样, 研究时就可得出每个文本的统计特征及整个语料库的总体统计特征。以创建张培基译文库为例, 首先把张培基先生的译文通过扫描并保存为可用检索软件检索的纯文本文档; 为使语料库发生更大作用, 还应对语料作一定的标注, 如用 CLAWS对语料进行词性标注, 然后再按译文标题分类保存。

索引工具是基于语料库的译文分析的必备条件。小型译文库建立后, 根据研究目的, 研究者需检索语料库, 通过语料库的检索获得用来分析的相关数据。目前, 较常用的索引工具有 WordSmith Tools, Antconc, TACT, MicroConcord等。索引工具的基本功能包括词表生成、语篇统计、"带语境的关键词" (KWIC)索引、主题词提取、词丛统计等。词表功能主要用来创建语料库中词汇使用频率列表, 可用来研究语料库中的词汇类型, 确定语料库中常见词丛和比较不同文本特定词汇的使用频率。"带语境的关键词" (KWIC)索引主要是查询和统计某个或某类词汇或短语在指定文本中出现的次数。主题词提取功能把一个语料库中的词频与参照语料库中的对应词的词频进行比较, 以确定这个语料库与参照语料库在词频方面是否存在显著差异, 为研究语域差异、作家写作风格差异、学习者语言与本族语使用者语言间的差异提供数据。把这些功能运用于译文分析, 研究者就能获得译文在词汇、句法、篇章和修辞等方面的统计信息。如词汇方面, 研究者可进行用词变化、平均词长、词汇密度、常用词表等方面的分析; 句法方面, 研究者可对译文的平均句长、复杂句、缩略形式、标点符号等进行分析; 语篇方面, 研究者可进行词汇衔接的量化分析。翻译文本多层面、全方位定量分析, 为翻译批评提供可靠的量化依据。

3 《茶馆》两译文用词特征个案分析

本研究以《茶馆》的英若诚译文和霍华译文为语料, 对比分析两译文的用词特征。为了更好地比较两译文, 还选用英语本族语语料库 BNC作为参照语料库。

首先对英若诚译文和霍华译文进行扫描、校对并分别保存为纯文本文件, 然后用 CLAWS软件对两译文语料进行词性标注, 建立小型《茶馆》译文库 (包括英若诚译文子库和霍华译文子库, 以下分别简称为英译文库和霍译文库), 再利用 WordSmith Tools的相关程序对两译文子库进行检索, 得到译文在类符形符比 (type-token ratio)、平均词长、词汇密度、常用词表等相关信息, 通过这些信息考察《茶馆》两译文的用词特征。

3.1 用词变化

语料库语言学主要通过类符形符比考察文本的用词变化情况。类符 (type)是语料库中不同的词语, 形符 (token)是语料库中所有的词形。类符形符比在一定程度上反映了语料的用词变化。类符形符比值越大, 表明该文本使用的不同词汇量越大; 反之, 不同词汇越少。通过类符形符比值的大小可以比较不同语料库中词汇变化的大小。但由于在一定时期内语言的词汇量有限, 若语料库容量不断扩大, 形符数会持续增加, 而类符数却不一定会增加, 从而导致语料库容量越大, 类符形符比值反而越来越小, 因而不同容量的语料库的类符形符比不具备可比性。所以, 语料库语言学一般用标准类符形符比 (standard type-token ratio)衡量语料库的词汇变化, 即按一定长度 (通常是 1000个形符)分批计算文本的类符形符比, 再求平均值。下表是通过 WordSmith软件统计的《茶馆》两英译文语料库和 BNC语料库的类符形符情况。

表 1 各语料库的类符形符统计

	英译文	霍译文	BNC
形符	22, 211	22, 714	102, 467, 488
类符	3, 031	3, 040	166, 962
类符形符比	13. 65	13. 38	0. 16
标准类符形符比	42. 16	40. 99	41. 20

从表 1可知, 英译文的形符数为 22, 211, 霍译文的形符数为 22, 714, BNC语料库的形符数为 102, 467, 488。英译文的类符数为 3, 031, 霍译文的类符数为 3, 040, BNC语料库的类符数为 166, 962。再看标准类符形符比, 英译文的标准类符形符比为 42. 16, 霍译文的标准类符形符比为 40. 99, BNC语料库为 41. 20。从这 3个数字可以看出, 英译文的标准类符形符最高, 其次为 BNC和霍译文。这表明: (1)英译文用词范围更加宽泛, 表达方式更加生动; (2)霍译文用词范围相对狭窄, 但更接近本族与使用者用词变化。

3.2 平均词长

平均词长是西文文本中类符的平均长度。通常情况下, 平均词长较长说明文本中用的长词、常见文本中分别由 2、3、4、5个字母组成的单词较多, 平均词长在 4左右。如果低于 4,意味着文章语言比较简洁浅显。如果远高于 4,意味着文章语言比较复杂深奥。可见, 词长反映用词的复杂程度。3个语料库的平均词长分别为 4.19、4.35和 4.54。这表明, 3个语料库在平均词长方面接近, 总体用词难度没有什么区别, 霍译文用词复杂程度更接近本族语使用者, 英译文用词复杂程度略底于本族语使用者。

为了更详细地描写语料库各长度词的使用情况, WordSmith软件在作词频统计时会自动计算出各长度词在语料库中的使用频率。但如果语料库的容量不一样, 语料库的实际词长出现次数就不具可比性。所以, 我们采用每 1000词词长数, 即每 1000词不同长度的单词在语料库中出现的次数。这种方法可更客观地比较各长度词在不同语料库中的分布情况。表$_2$是各长度词在英译文、霍译文和 BNC语料库的每千词分布情况。结果显示, 英译文的每千词中 1-4个字母长度词均高于霍华译文, 而霍华译文每千词中 5-10个字母长度词均高于英译文。

5个字母词以上(含 5个字母)的词属于难度较大的词。我们把上表中各语料库中 5个字母词以上(含 5个字母)的词的每千词的出现频率相加, 得到如下数据: 英译文为 342.02/千词, 霍译文为 384.6/千词, BNC为 399.8/千词。这些数据反映 3个语料库使用难度词在总体上有一定差别: 霍译文用词难度大于英译文, 而两译文的用词难度又都小于本族语使用者。

表$_2$　各语料库每千词的不同长度词统计

	英译文	霍译文	BNC
1-letter words	37.8	37.3	47.1
2-letter words	150.3	141.9	172.5
3-letter words	227.1	207.3	203.9
4-letter words	242.9	228.9	176.7
5-letter words	121.1	129.1	107.7
6-letter words	87.4	105.1	79.4
7-letter words	63.1	64.4	72.9
8-letter words	37.1	44.4	51.3
9-letter words	18.1	20.1	37.5
10-letter words	9.1	16.4	24.5
11-letter words	3.7	3.2	13.8
12-letter words	1.62	1.2	7.2
13-letter words	0.7	0.5	4.0
14(+)-letter words	0.1	0.2	1.5

3.3 词汇密度

词汇密度指实词在语料库中占的比例, 其计算方法为: 实词÷总词数×100%。英语实义词指具有稳定词汇意义的词语, 包括名词、动词、形容词和大多数副词 4个词类; 功能词指不具备稳定词义或意义模糊而主要起语法功能作用的词语, 主要包括代词、介词、连词、冠词、助动词等词类。(胡显耀 2007:214)在具体统计中, 本文把名词、动词、形容词和副词 4类"具有稳定词义"的词类作为实词。篇章中的实词越多, 篇章的密度越大, 其传递的信息也越多。可见, 词汇密度可以反映篇章的信息量和难度。词汇密度偏高, 说明该篇章的实词比例较大, 因而信息量也较大, 难度也相应增加。

表$_3$　两译文子库词汇密度统计

	英译文		霍译文	
词性	形符	百分比	形符	百分比
n.	6,028	27.1	6,479	28.5
v.	2,803	12.6	3,210	14.1
adj.	1,121	5.0	1,377	6.0
adv.	1,031	4.6	1,762	7.8
总计	10,983	49.3	12,828	56.4

由表$_3$可知, 霍译文中名词、动词、形容词和副词所占的比列均高于英译文, 霍译文的词汇密度明显高于英译文。词汇密度的差异表明: 霍译文使用实词的倾向性明显高于英译文。英译文通过减少实词来增加译文的可读性, 而霍译文实词比例高, 使其译文传达更多信息, 客观上增加了译文的难度。

3.4 常用词表

WordSmith Tools提供的词表功能除了普通词频表, 还有按字母顺序随意改变次序的词表, 同时也提供语料库的各种基本统计信息。从表 4可知, 在 3个语料库使用频率最高的 10个词中, 有 6个相同: the, of, and, a, to, in。英译文和霍译文使用 9个相同词: the, of, and, a, to, in, a, you, I,各有 1个不同: 英译本中的 Lifa和霍译本中的 little。使用频率最高的 10个词在两译文中占的比例分别 18.69%和 18.47%,均低于 BNC的 13.14%。这说明: 两译文在使用频率最高的前 10个词的使用上基本一致, 词语选择和所占比列几乎相同; 与本族语使用者相比, 两译文呈现出高频使用最常用词的倾向, 这一现象验证了翻译文本的简化特征。

我们用同样方式统计两译文和 BNC中第 11-30个最常用词的情况。结果显示, 两译文在第 11-30个最常用词中共使用 11个相同的词: it, for, with, Liu, but, my, me, your, be, on, old。第 11-30个最常用词在各语料库中所占的比例分别为: 英译文 11.96%,霍译文 12.65%, BNC 12.01%。统计结果表明: 两译文在第 11-30个最常用词的用词选择上存在较大差异; 词语使用频率差别很小, 与本族语使用者相当。

表4 3个语料库常用词表统计

	英译文			霍译文			BNC		
序号	词汇	频率	%	词汇	频率	%	词汇	频率	%
1	The	813	3.66	The	625	2.75	The	6,074,315	5.93
2	A	517	2.33	To	564	2.48	Of	3,062,469	2.99
3	To	499	2.25	You	531	2.34	And	2,629,938	2.57
4	You	453	2.04	A	519	2.28	To	2,613,090	2.55
5	And	366	1.65	And	431	1.90	A	2,189,709	2.14
6	Of	365	1.64	Of	373	1.64	In	1,963,194	1.92
7	Wang	352	1.58	Wang	342	1.51	That	1,121,864	1.09
8	I	316	1.42	I	323	1.42	It	1,058,066	1.03
9	In	261	1.18	In	258	1.14	Is	997,109	0.97
10	Lifa	208	0.94	Little	230	1.01	For	890,134	0.87
总计		4150	18.69		4196	18.47		22,599,888	13.14

可以对《茶馆》两译文的用词特征概括如下: (1)英译文用词范围更加宽泛,表达方式更加生动; 霍译文用词范围相对狭窄, 但更接近本族与使用者用词变化情况。(2)霍译文使用实词的倾向性明显高于英译文; 英译文通过减少实词来增加译文的可读性, 而霍华译文实词比例高, 使其译文传达了更多的信息, 客观上增加了译文的难度。(3)两译文在使用频率最高的前 10个词的使用上基本一致, 词语选择和所占比例几乎相同; 与本族语使用者相比, 两译文呈现出高频使用最常用词的倾向。两译文在第 11- 30个最常用词的用词选择上存在较大差异, 但词语使用频率差别很小, 与本族语使用者相当。

4 结束语

本研究探讨基于语料库的译文分析可行性和实现途径, 对比分析《茶馆》两译文的用词特征。译文分析的语料库途径把定性研究与定量研究有机结合起来, 通过特定文本分析软件对翻译文本在词汇、句法、语篇和修辞等层面的倾向性特征进行量化分析。该途径具有如下优势: (1)它将翻译文本研究变得具体和可操作性强; (2)它将小规模的、人工的和局限于个别文本类型的研究变成大规模、系统和目标明确的研究; (3)研究结果具有客观性和说服力, 较好避免了传统点评式和感悟式的单一定量研究方法带来的主观性和随意性。

参考文献

丁树德. 浅谈西方翻译语料库 [J]. 外国语, 2001(5).

何安平. 语料库语言学与英语教学 [M]. 北京: 外语教学与研究出版社, 2004

何莲珍. 基于汉、英语平行语料库的翻译数据库设计 [J]. 现代外语, 2007(2).

胡显耀. 基于语料库的汉语翻译小说词语特征研究 [J]. 外语教学与研究, 2007(3).

黄立波 王克非. 翻译普遍性研究反思 [J]. 中国翻译, 2006(5).

廖七一. 语料库与翻译研究 [J]. 外语教学与研究, 2000(5).

刘康龙 穆 雷. 语料库语言学与翻译研究 [J]. 中国翻译, 2006(1).

倪传斌. 语料库数据驱动技术在科技翻译教学中的应用 [J]. 中国科技翻译, 2005(4).

秦洪武. 对应语料库在翻译教学中的应用: 理论依据和实施原则 [J]. 中国翻译, 2007(5).

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

王克非 秦洪武. 英汉语言特征探讨——基于对应语料库的宏观分析 [J]. 外语学刊, 2009(1).

吴 昂 黄立波. 关于翻译共性的研究 [J]. 外语教学与研究, 2006(5).

杨惠中. 料库语言学导论 [M]. 上海: 上海外语教育出版社, 2002

Baker, M. Corpus Linguistics and Translation Studies: Implications and Applications [A]. Baker, M., Francis, G. & E. Tognini-Bonelli (eds.). *Text and Technology: In Honor of John Sinclair* [C]. Amsterdam: John Benjamins, 1993.

Baker, M. Corpus-based Translation Studies: the Challenges that Lie Ahead [A]. Somers, H. (ed.). *Terminology, LSP and Translation: Studies in Language Engineering, in Honor of Juan C. Sage* [C]. Amsterdam: John Benjamins, 1996.

Kenny, D. Creatures of Habit? What Translators Usually do with Words [J]. *Meta XLIII*, 1998(4).

Kenny, D. *Lexis and Creativity in Translation — A Corpus-based Study* [M]. Manchester: St. Jerome, 2001.

Laviosa, S. *Corpus-based Translation Studies: Theory, Findings, Applications* [M]. Amsterdam: Rodopi, 2002.

Munday, J. A. Computer-assisted Approach to the Analysis of Translation Shifts [J]. *Meta XLIII*, 1998 (4).

Olohan, M. *Introducing Corpora in Translation Studies* [M]. London and New York: Routledge, 2004.

收稿日期: 2009- 06- 13 【责任编辑 李洪儒】