日中口译平行语料库的设计与建设

庞　焱
（广东外语外贸大学・广州・510420）

内容提要：口译研究的总体发展路径正在向跨学科的多元式研究扩展；口译研究的方法由原来的主观推测和实践经验总结向以数据为基础的客观描写和实证分析转变。由此，基于语料库工具的口译研究成为了越来越重要的研究手段。本文介绍了用于日中口译研究的日汉口译平行语料库的设计和建设，期待以此为工具，为今后日中口译研究的深入展开打下夯实的物质基础。

关键词：日语；口译；平行语料库；语料库翻译学
中图分类号：H0-0　　**文献标识码：**A　　**文章编号：**1672-0962(2012)03-0029-04

一、语料库定义

语料库（corpus 或 corpora，corpuses[复]）是指按照一定的语言学原则，运用随机抽样的方法，收集自然出现的连续的语言运用文本或话语片段而建成的具有一定容量的大型电子文库[1]。而斋藤认为：语料库或电子语料库是指具有明确的研究目标，并能被电脑使用的书面语或口头语的话语文本集合[2]。因语料库可根据话语资料的收集方法做不同分类，且收录了用于语言研究的经整理的文本，因此语料库具有不同结构和设计，同时，收录在语料库中的文本已做相关标识。从其本质上讲，语料库实际上是通过对自然语言运用的随机抽样，以一定大小的语言样本代表某一研究中所确定的语言运用总体。

从不同的观点来看，语料库具有狭义和广义两种定义。如果从语言学角度来广义地解释地话，所谓语料库是指：能够用于语言研究，并在现实中被使用的口语或书面语的话语文本集合体。如果从语料保存手段来看，随着当今电脑的普及，所谓的“语料库”指的是“能够使用电脑处理的大量语言资料的状态，即通过机械可读形式 (machine-readable form) 处理的文本集合体。”狭义的定义仅限定于文本的收集方法，指为了语言研究为目的的，具有明确的语料库设计的文本集合体，这些语料库设计具体包括成为语言学研究对象的语言变种（language variety）等等。也就说，从广义的角度来看，收存了某个作家所有的文学作品的文本集合体可以称作语料库，与此相对，狭义的语料库指的便是抽取了某些样本的文本集合体。从附带的信息的角度狭义地来理解的语料库，指的是将话语文本附带上各种各样的数字化信息的文本集合。

综上所述，语料库的理解可从不同角度展开。本文使用赤野的定义：“代表了某种特定的语言、方言或其他语言的多样性特点，且设定了研究目标，储存在电脑中的、经过处理的口语或书面语的文本集合体。”[3]

二、语料库翻译学研究概况

语料库用于翻译研究最早可以追溯到 20 世纪 80 年代[4]，但通常认为，Baker[5] 的“Corpus Linguistics and Translation Studies:Implications and Applications”一文是语料库翻译研究范式（Corpus-based Translation Studies Paradign）开始建立的标志。1998 年，加拿大蒙特利尔大学主办的翻译研究季刊 META 出版了 Sara Laviosa 主编的基于语料库的翻译研究专号 META 43（4），从理论阐释和实证研究两方面宣告基于语料库的翻译研究已经成为一个新的翻译研究范式。以 Baker 为标志，语料库翻译学可划分为前语料库和基于语料库的两个时期，前者是指大规模机读翻译文本用于翻译研究之前，通过人工采集原文和译文文本，并对于翻译有关的语言现象进行对比、分析和统计的时期。基于语料库的翻译研究范式产生以来，其研究范围覆盖了从翻译过程到翻译产品的各种翻译现象，特别是翻译共性（Translation

收稿日期：2012-01-05　＊基金项目：本研究得到：“广东省普通高校人文社会科学研究基地重大项目”成果 / 资助，项目名称“口译规范描写及其在口译教学和评估中的应用”以及广东外语外贸大学校级青年项目资助，项目名称“口译研究方法探索——日语口译语料库的开发与建设”。
作者简介：庞焱（1971 －），女，广东外语外贸大学东方语言文化学院日语系副教授，高级翻译学院翻译专业博士研究生；研究方向：日语语言学，口译研究。

Universal）、翻译过程（translation process）、翻译转换与规范（translation shift and translation norms）、译者文体（translator's style）、翻译教学等诸多方面。新的研究方法促进范式的形成和发展，新的研究范式又能带来研究思路的更新和研究重点的转移。国际上基于语料库的翻译研究发展较快，目前此领域中以对翻译共性和译者文体的讨论最为突出。国内起步较晚，规模较大的双语对应语料库有北京外国语大学研制的 3000 万字 / 词的"通用汉英对应语料库"等。

语料库翻译学当前面临三大理论和应用的研究课题[6]：

（1）大范围的翻译调查，包括翻译教学、翻译文体的考察，以及对应词搭配频率等统计数据的检索和分析。

（2）自动翻译研究，将开展了半个世纪的机器翻译与语料库翻译结合起来，以期取得新的实质性突破。

（3）更广泛更有效地描写性翻译研究，包括翻译规范的研究和翻译普遍特征或共性的研究。

Mona Baker 是最早进行语料库翻译学研究的学者之一。她在"语料库语言学和翻译研究"中对这两者的结合做了初步阐发。在大范围翻译调查方面，如翻译问题的考察，她也率先从语料库角度探讨译者的文体特征特别是从类符——形符比、平均句长及词项使用特点等方面加以分析。

关于第二个研究指向，主要是将基于规则的研究方法 (rule-based approach) 同基于语料的研究方法（corpus-base approach）相结合，为自动翻译寻找出更便利可行的途径。

值得关注的是第三点，即语料库基础上的描写性翻译研究，特别是有关翻译普遍性问题的探讨。Baker 在前人研究的基础上，提出了翻译普遍特性（universal features of translation）的假设，主要内容是翻译文本中的（1）显化现象，（2）消歧和简化倾向，（3）泛化特点，（4）倾向于避免重复，（5）倾向于凸显目标语语言特征，（6）某些特征呈现特定类型的分布。此外 Laviosa 调查了英语翻译文本中的 4 种核心词汇运用模式；Kenny 通过对原文、译文的语义韵比较，发现译文语言有净化（sanitisation）现象，Overas 考察了英语 - 挪威语翻译中衔接层面上的显化现象；王克非据大型对应语料库探讨了译本扩增情况，Lavio 讨论译文与母语在词汇使用上的不同，柯飞通过语料库考察，发现了翻译过程中对原文的模仿可能使译文变得复杂化、冗长化，Xiao & McEnery 发现在"体"标记的使用上，汉语译文比汉语原文多出约一倍；Ebeling 比较了英语和挪威语存在存现句使用上的特点；maia 以双语对应语料库观察英语和葡萄牙语在人称主语使用频率上的差异等等。

三、本研究语料库的建立

口译研究的总体发展路径正在向跨学科的多元式研究扩展；口译研究的方法由原来的主观推测和实践经验总结向以数据为基础的客观描写和实证分析转变。有鉴于此，本文的主体方法可以定位为：以现场口译的源语 - 目标语平行语料库数据为基础的定量研究。

1. 现场口译语料的收集和选择

在现场口译的语料选择中，本研究选择同声传译，因为相对于交替传译而言，日译中时，由于时效性的影响，日语句子的长度和谓语等主要成分靠后的句型结构对译员理解发言人话语意思所起的作用更大更明显，观察和分析译员在有限的时间内如何权衡自己的口译策略和技巧，对本研究的目的达成将起到非常重要的启示和参考作用。

用于本研究的语料具有开放性和地域性特点。开放性是指语料来源于政治、经济、文化等相关的内容，并且根据会议召开的实际情况，每年都会及时增加相关语料入库。而地域性是指语料内容都是与广东省的具体情况相关。所选内容主要包括以下三个方面：

A．广东省政府召开的最高级别经济会议——广东省经济发展咨询会（广东省长洋顾问会议），包括 2005、2007 和 2009 年连续三次的会议录音。此类会议主要围绕广东省的经济发展展开讨论并提出相关建议，与会者均为世界 500 强企业的日本企业总裁，话题围绕技术创新、人才培养、知识产权保护、环境保护、国际合作等等展开。

B．广东 - 兵库县经济促进会会议录音。此会议主要也是围绕广东、兵库经济如何更好发展等献计献策而举办的大型经济研讨年会。

C．广东省包括广州市、深圳市等珠三角洲地区的其他国际性会议和论坛。包括南海经济座谈会、中日经济热点论坛、节能环保论坛、国际旅游峰会等等相关内容。

以上内容共计日语 10 万字词左右。

选择此类语料主要是出于以下几个方面的原因：一是因为保密等原因，大规模完整的口译语料的采集非常困难，而这些语料都是笔者在长年的同传现场使用录音笔从头到尾录音获得，虽然各场会议的内容不同，但各场会议性质一样，都是在正式场合下的会议口头语言表达形式，符合本研究的研究目的，可以看做是系统的同质性语料，都是客观上获得，语料的规模也比较大。二是由于地域性的限制。口译教学需要理论联系实际，因此笔者在从事口译教学的同时，通常会寻找机会从事一些力所能及的口译实践工作。基于地理条件等所带来的时间上的限制，笔者主要担当的是广州及周边城市举办的各类国际会议的同声传译。虽然会议在广东地区举办，但出席的日本嘉宾多来自于日本国内的各个地区和城市，尽管在语音语调上有所不同，但基本都是采用的日本标准语言讲话和讨论，对于将日语作为源语，并以研究源语长句和口译策略及技巧为研究目标的本研究完全没有影响。

2. 目标语语音的录制

由于录音条件和录音现场的限制，同声传译的录音往往只能录下现场或译员的单方面的声音，除非有专业电视台的录音录像设备，才能同时将会场发言人的发言和译员的译语

同步收录进来，但作为译员，大多数情况下根本无法拿到这种录音资料。基于此原因，笔者手头所拥有的也仅为源语录音，而无现场译员的声音。由此对用于本研究平行语料库的目标语文本的收集和整理产生了一定的困难，同时也对客观地获得真实的数据（real-life data），客观地分析实际口译情景下口译活动的各种情况的数据获得带来了一定的困难。

考虑到语料库建立时的实际困难，本研究语料库的建立采用的方法介于现场观察法和实验法之间。也就是说，本语料库中源语语料为现场录音语料，而目标语语料是在实验室请译员录音所获得。在实验过程中，笔者对源语的话语内容以及口译的现场给译员带来的认知处理压力两方面进行了控制。实验结果是对译员在口译过程中所采取的口译策略和技巧进行分析和比较。

要同时请到多名具有同传口译经验和水平的口译译员为自己的实验所用，是件非常困难的事情。一是人数上不允许，另外资金也是非常有限的。所幸的是在台湾辅仁大学翻译学研究所口译班二年级有 4 名大致符合本人研究要求的口译译员。之所以说他们基本符合本研究的实验要求，是因为这 4 名已经通过 2010 年 6 月辅仁大学翻译学研究所举办的口译专业考试，获得台湾地区社会口译资格，能独立在台湾地区从事口译活动的新进译员，具有十场以上的同传实践经验，所以从某种程度上来说，他们能代替活跃在社会实践中的社会口译群体；而同时他们又均是口译学员出身，刚经历过学生时代，身份背景以及水平大致等同但略超出笔者目前所教授的研究生和本科四年级同声传译课程的学生水平，又符合本研究所面向的学生群体。

语音录制的场所选定在辅仁大学翻译学研究所的同传教室进行。该教室由一个圆形会议桌形成，四面设有 6 个同传箱子，每个同传箱子均由玻璃窗户与教室隔开。坐在同传箱子中，译员能清晰地观察到教室的各个角落。此间教室平时用来讲授同传课程，但也经常举办一些小型国际性会议和讲座，拥有先进的同传设备，并能随时对发言人和译员进行多种要求的录音。所以这间平时用作教学的课室实际上也称得上是一件五脏俱全的小型国际会议室。

录制过程中，笔者会提前一个星期将每场会议的相关资料发放给译员，这些资料包括会议的议程、背景资料介绍、主办方和出席者名单，以及与会发言人其他发言稿等相关资料。笔者在上节中的第（2）点已经说明，用于本研究语料库中的语料的源语介于第（1）类和第（2）类之间，所以发放其他发言人的稿件目的仅仅只是帮助译员对会议相关知识有更具体的了解，而不会影响源语语料的特性。也就是说，笔者尽可能地还原了自己在承接每场口译任务当时的译前准备状态，尽可能保证译员工作的客观性。考虑到译员全部出身于台湾的成长背景，对于中国大陆特别是广东省的一些包括人名、地名和术语在内的特殊词汇的陌生性，笔者在发放资料后的第三天左右召开一次全体译员的碰头会，将发言中可能出现的相关词汇和会议涉及到的具有广东地域性特点的背景知识做一次系统的说明。考虑到连续作战给译员带来的疲劳，每次的语音录制时间控制在了 1 个小时之内，且中途 20 分钟左右会选择在短暂的休息后，再继续进行。发言人讲话时使用过的现场播放信息，比如 PPT 等资料也在录音现场得以还原，而且拥有录像的会议，也会在现场给予播放。且在录音阶段，笔者会在会议室的圆桌旁边安排 5、6 个学生扮演会议出席者进行旁听，以此营造一种开会的氛围。而每位译员因都处在专业考试前夕的紧张备战状态，非常珍惜每一次的练习机会，所以对于每场会议的录音，大家也都是全力以赴，认真对待的。基于以上这些过程，可以说，录音的目标语质量应该还是比较客观和真实的。

3. 文本整理

（1）目标语 word 文档的整理和标识：

整理：目标语的整理和标识方法基本等同于源语。先将 4 位译员的译语转写成 word 文档文本，然后再对应已划分的日语句子语意，将中文译文依次划分成独立句子。

标识：在此阶段，笔者将中文译员同传时停顿在两秒或两秒以上[①]的地方进行了标识并注明了精确的停顿时间，以作日后的非正常停顿研究之用。而且还根据研究的需要，将译员的每句话与每句话之间的间隔时间做了标识并注明了详细的时间。

以上的文本整理全部在 word 文档中进行。

四、本研究平行语料库的使用

1. 本语料库的构成

根据研究的目的、分析方法和观察的语言内容等，语料库的大小各有不同，也并非语料库容量越大约好。Biber 认为：在进行实际数据分析时，即便其依赖的对象只是一个只有 10 个文本的 1000 词左右的语料，只要这些语料能够足以表明研究者想要研究的语言特征，这也可称得上合适的语料库；而在调查单个语料库的一般语言特征的时候，即便准备一个含有 1 万词量的语料库，也未必能满足研究的需要。而在研究文本构成的论文中（使用的句子规模基本在 400-600 个左右。而且，真正意义上的现代计算机语料库并非以无限追求容量的扩张为主要目的。在语料库应用研究中，比容词量更重要的是如何根据研究者或用户的需要在语料取样中保持良好的平衡。“围绕某一可识别的文类与各种主题标准所提供的语料库材料，其构成应以用户需要为基础，即用户能够根据自己的学习和研究需要，通过汇集（语料库材料）或把语料库重新切割成各个微型语料库，获得自己的平衡和代表性”，所以，并不是说把大量电子文本简单堆放在一起就建成了语料库，一个语料库的设计和建成总是为了代表某一具体领域的语言运用或满足相应的研究目的。基于以上研究结果，笔者建设了以下内容的平行语料库用于本论文的研究。

另外本本语料库的标识工作全部通过手工作业完成。本语料库中的日语源语约为 10 万字词，中文译语约为 36 万字词，共有译员 4 人，共有 1018 个日语句子，4275 个中文句子。

表 2-1 本语料库的具体构成

	日语源语	中文译文
文字数	107，405	360，000
句子数	1018	4275

2. 本语料库的特征

根据本研究的研究目的以及按照以上的步骤完成的本研究用平行语料库具有以下几个特点：第一，本语料库源语语料属于即兴的独白式发言或对话，以及有一定准备的独白式发言或对话形式的国际会议发言，目的是用来研究会议同声传译中的日语长句存在哪些难句，以及译员应对这些难句时的相关策略，此语料库可以看成特殊目的语料库（special purpose corpus）。第二，本语料库由 4 个子语料库构成，即日语源语分别对应 4 个译员的中文译语，属于单向平行语料库。第三，为了保证研究目的和在同一环境下的公平口译，口译译员全部选择的是介于学生和职业译员水平之间的译员，且每位译员都是在同样的前提和条件下进行的同传口译工作，语料库中所整理的译文文本也是在此条件下整理而成，因此，日语源语和由各个译员的口译录音整理而成的中文文本具有实验性质。

3. 源语和目标语的对应方法

为了调查译员对日语源语长句中难句的口译策略，本语料库建设成了源语和目标语对应的平行语料库。源语和目标语的对应原则上是以句子为单位进行的对应。因为是通过录音进行转写，日语源语句子的划分是按照说话人意思表达的完整性来进行切割，从文本表达形式上来看的话，两个“。”间的句子看为一个独立句子。中文译员的句子切割方法基本同上。至于段落最开头部分的句子，便是将从发言人开始讲话的部分到第一个“。”之间的部分看成一个句子。但由于译员在翻译时，有漏译或者错译的情况，这时，便对源语和目标语进行对照整理；漏译的句子，在语料库中的译语处便为空白，而错译的句子，边根据源语语意以及目标语的上下文语境让其适当地与源语句子进行对应。

4. 语料库标识

根据本研究的需要，笔者对语料库进行了相关的信息标识。具体如下：

本语料库的源语语料中根据研究目的，对“という”结构进行了标识，而针对目标语语料，为了附和译员研究策略和技巧的研究目的的需要，在源语语料中对译员停顿长达两秒或以上的停顿之处进行了标识，标识记号为（P），根据此标识，笔者可以任意搜索到译员的非自然停顿之处，非常方便。

五、结语

本语料库的源语字词数量约为 10 万字词左右，目标语字词数约为 36 万字词，源语句数为 1018 个，目标语句数为 4275 个。虽然规模较小，但“此语料库的设计和建成代表了一个具体领域的语言运用或满足相应的研究目的”的，本语料库将成为日汉口译尤其是日汉同声传译研究的重要工具，为日汉口译研究的展开打下了夯实的物质基础。

注释

①参考名古屋大学语料库研究小组所定停顿时间。

参考文献

[1] 杨慧中主编 . 语料库语言学导论 [M]. 上海外语教育出版社，2002: 33.

[2] 斉藤俊雄，中村純作，赤野一郎 . 英語コーパス言語学基礎と実践 [M]. 研究者出版社，1997: 17.

[3] 赤野一郎 , 吉村由佳 .Corpus linguistics の現在の動向と問題点（1）コーパスとその構築 [J],Studies in English Liguistics & Literature, 1991.7(1-45).

[4]Laviosa, Sara. *Corpus-based TranslationStudies: Theory, Findings, Applications*[M]. Amsterdam, New York: Rodopi. 2002

[5]Biber,D. *Dimensions of Register Variation:a Cross-linguistic Comparison*[M].Cambridge University Press. 1995

[6] 王克非 . 语料库翻译学——新研究范式 [J]，中国外语 , 2006，(3): 5-6.

Design and Development of Japanese-Chinese Interpreting Parallel Corpus

Pang Yan

Abstract: With interpreting studies (IS) developing towards a broader scope of multi-disciplinary endeavor, its methodology has undergone a shift from assumption- and practice-driven inquiries to data-based empirical studies that are more objective and descriptive. Given this, corpus has become an increasingly important research method for interpreting. This article introduces the design and development of Japanese-Chinese interpreting parallel corpus, in the hope that it can lay a solid foundation for deepening research into Japanese-Chinese interpreting.

Key words: interpreting; parallel corpus; corpus-based translation studies