

# 《基于双语语料库的汉英视点体对比研究》<sup>1</sup>序言

冯志伟

瞿云华博士的专著《基于双语语料库的汉英视点体对比研究》是一项以平行语料库为基础的双语言对比研究。作者从包含 18 部白皮书和 5 部文学名著的 500 多万字的汉英双语对齐的平行语料库 (parallel corpus) 中,抽取了 18101 条汉语视点体标记和英语对应语句,并人工标注了其中的 14568 条含有体义项的视点体标记和英语的对应语句,构建了汉英视点体平行语料库与汉语视点体和英语非视点体结构平行语料库各一个。从这两个经过人工标注的平行语料库 (annotated parallel corpus) 中,抽取了汉英视点体对应例句 12334 对,汉语视点体和英语非视点体结构对应例句 2234 对,统计了汉语各视点体标记所对应的英语视点体和非视点体类型与出现频率。在构建了汉英视点体标注平行语料库之后,作者进一步从前贤的研究成果中总结出汉语视点体的语法意义,从汉英视点体对应的实例数据中获取了确立汉语视点体标记语法意义的事实依据,运用 Smith 的双部理论<sup>2</sup>,从视点体对情状的聚焦位置差异出发对汉英各视点体之间的语义进行了全面而深入的对比分析,从理论上推导出了汉语视点体派生的语义条件,并从实际语料中归纳出汉英视点体转换所需要的语义条件,验证了从理论上推导出来的结果。最后,作者以认知语言学 (cognitive linguistics) 的原型理论为依据提出了汉语视点体的派生原则解释转换规则,运用派生视点体的合成规则对派生现象进行了形式化的描述。

可以看出,作者研究汉英视点体的方法是一种基于平行语料库的方法。这种方法与语言本体研究中经常使用的依靠语言学家的语感和个人的语言经验的内省方法截然不同,这是一种颇具新意的语言研究方法。

这种基于平行语料库的研究方法源远流长。在这个序言中,我愿意简单地回顾一下国内外平行语料库的发展的历史与现状。

我们知道,解读密码 (decipherment) 是古典文献研究的一个重要内容,历代学者们曾经依靠自己的聪明才智出色地解读了不少古代的铭文,或者通过铭文中已知的部分来解读铭文中未知的文字。Rosetta (罗塞塔) 石碑 (Stone) 上古代埃及文字的解读,就是使用平行语料库方法来解读密码的一次成功的范例。

Rosetta (罗塞塔) 石碑由上至下共刻有同一段诏书的三种语言版本,是用埃及象形文字 (Egyptian hieroglyphs, 又称为圣书体,代表献给神明的文字),埃及通俗文字 (Egyptian Demotic, 又称草书体,是古代埃及平民使用的文字) 和古希腊文 (Greek, 代表统治者的语言,这是因为当时的埃及已臣服于希腊的亚历山大帝国之下,来自希腊的统治者要求统治领地内所有的此类文书都需要添加希腊文的译版) 三种不同的文字写成的,刻于公元前196年,现藏大英博物馆。

<sup>1</sup> 此书于 2008 年 9 月由科学出版社出版, ISBN 978-7-03-022817-8。

<sup>2</sup> Smith, C., 1991, *The Parameter of Aspect*. Dordrecht: Kluwer Academic Publishers

在公元4世纪结束后不久，尼罗河文明式微，不再使用的埃及象形文字和埃及通俗文字的读法与写法都彻底失传了，虽然后来有许多考古专家与历史学专家极尽所能来研究，却一直解读不了这些神秘文字的结构与用法。直到1799年法国远征军在埃及的Rosetta（罗塞塔）发现了Rosetta石碑，才使埃及古代文字的解读工作获得了突破性的进展。Rosetta石碑独特的三语对照写法，意外成为解码的关键，因为这三种语言中的古希腊文是近代人类可以阅读的，利用这个关键来比对和分析碑上其它两种语言文字的内容，就可以了解这些失传的古代语言的文字与语法结构。学者们依靠已知的希腊文来解读未知的埃及象形文字和埃及通俗文字这两种埃及的古代文字，在1822年终于揭开了埃及古代文字的神秘面纱，成功地解读了埃及古代文字。

在许多尝试解读Rosetta石碑的学者中，19世纪初期的英国物理学家Thomas Young（托马斯·杨）是第一个证明碑文中曾多次提及Ptolemy（托勒密）这一人名的读音并且利用这个线索首先解读碑文的学者。法国学者Jean-François Champollion（让-佛罕索瓦·商博良）第一个理解到，一直被认为是用形表义的埃及象形文字，原来也是具有表音的作用，他的这个重大发现后来成为解读所有埃及象形文字的关键线索。也正是因为这一缘故，Rosetta石碑被学者们看成是探索古埃及的语言和文化的钥匙。

我们认为，Rosetta石碑上面的三种文字就像三个彼此对应的平行语料库（parallel corpus），Rosetta石碑也许就是世界上最早的平行语料库。Rosetta石碑的解读，是使用平行语料库解读密码的成功范例。可惜，这样成功范例当时在语言学研究中并没有得到推广，绝大多数语言学家仍然使用基于语感和个人语言经验的内省方式来研究语言。

这种情况，在上世纪90年代初才发生了明显的改变。

自从20世纪60年代初，第一代大型电子语料库——BROWN和LOB语料库建立以来，世界上许多国家和地区陆续建立起各种单语语料库（包括书面和口语，共时和历时语料库），但平行语料库的研制，则直到上世纪90年代初、中期才开始。比较早的平行语料库有英语—挪威语双语平行语料库和1992年建立的英语—意大利语双语平行语料库。

目前平行语料库的研制业已成为语料库研究的一个重点，平行语料库的研究正朝着不断扩大库的容量、深化加工和不断拓展新的领域等方向继续发展。

随着从事语言研究和机器翻译研究的学者对平行语料库重要性的逐渐认识，国内外很多研究机构都致力于平行语料库的建设，并利用这些语料库对形形色色的语言现象进行了深入的探索。

目前国际上已建立了很多外文的平行语料库。例如，

- 在加拿大人们收集加拿大议会辩论的英法双语稿建立了Hansard英法平行语料库。该平行语料库是最早建立的平行语料库之一，在平行语料库的研究中被许多学者广泛使用。
- Johanson等人在挪威奥斯陆大学（University of Oslo）建立了英语—挪威语平行语料库，包括一个核心语料库和一个增补语料库，1997—2001年又增加了德语、荷兰语、葡萄牙语的对应语料。
- 欧盟议会语料的多语语料库（包括11种欧盟语言）。
- Resnik等人在美国马里兰大学建立的Bible（圣经）九国语言的平行语料库
- JRC-ACQUIS 多语平行语料库。
- 捷克国家语料库（Czech National Corpus Czech-Other Languages）。
- TELRI（Trans-European Language Resources Infrastructure）多语言平行语料库（Plato的《理想国》多语言译本）。
- Scandinavian 语料库。
- ES-PC 英语—瑞典语平行语料库。

- PEPC 葡萄牙语—英语平行语料库。
- ET10-63 英语—法语平行语料库。
- CRATER 西班牙语—法语—英语平行语料库。

已建立的中文/外文平行语料库有：

- 英国伯明翰大学 (The University of Birmingham) 建立的中英对应语料库。
- Gao (高照明) 收集台湾 Sinorama 杂志文章建立的 Sinorama 中英对应语料库。
- 英国兰卡斯特大学 (Lancaster University) 的 Babel 英汉语料库 (544,095 词, 句子级对齐)。
- 北京外国语大学中国外语教育研究中心的通用汉英对应语料库 (约 3000 万汉字/英文词)。
- 北京大学汉语语言学研究中心的 CCL 汉英双语语料库 (233589 句对)。
- 北京大学计算语言学研究所的汉英/汉日双语语料库 (汉英句对齐语料: 200101 句对, 汉英词对齐语料: 10102 句对, 汉日句对齐语料: 20000 句对)。
- 哈尔滨工业大学的英汉双语语料库 (40-50 万句对, 在句子、短语、词汇三级实现双语对齐)。
- 中国科学院软件研究所的英汉双语语料库 (15 万句对)。
- 中国科学院自动化研究所的英汉双语语料库 (香港法律英汉双语对齐语料 31 万句对, 并从英汉双解词典中摘取例句 25000 个句子对)。
- 哈尔滨工业大学计算机学院语言技术研究中心面向奥运的中英日三语语料库 (220 余万字, 52227 个三语句对)。
- 东北大学的英汉双语语料库 (100 万词)。
- LDC 香港新闻英汉双语对齐语料库。
- 香港法律英汉双语对齐语料库 (31 万句对) 与在此基础上建立的双语法律信息系统 BLIS (21 万句对)。
- 内蒙古大学结合汉蒙机器翻译系统, 建立了近 20 万词的汉蒙对照政府文献语料库。
- 新疆大学建立的面向法律文档的汉维双语对齐语料库收集了 2000 对汉维句子 (句子级对齐)。

从国内外平行语料库建设情况来看, 国际上的研究一开始侧重于欧洲语言, 最早的平行语料库大多是西方语言之间的平行语料库, 90年代后期开始建设涉及到汉语的平行语料库。几年来, 平行语料库的研究价值越来越得到国内学者的关注。许多大学和研究机构开始进行汉语和外语的平行语料库建设, 现在已建成了为数不少的平行语料库。其中, 北京外国语大学中国外语教育研究中心建设的通用汉英对应语料库设计严谨, 有一定规模 (建成后约3000万词), 是研究英汉双语的难得资源。互联网上也有一些提供简单检索的英汉平行语料库, 如中科院计算所软件研究室开发的双语句对数据库, 有一定参考价值。目前, 在国内有藏语、蒙古语和维吾尔语等少数民族语言的语料库, 然而这些语料库只限于单语种。在我国涉及少数民族语言的双语语料库的探索工作还比较少。目前还没有见到规模较大的、经过深度加工的、以汉语为源语言的汉-民 (少数民族语) 双语语料库的报道。双语语料库由于涉及到两种语言的对应, 工作量很大, 需要投入较多的人力、物力和财力, 一般人不敢轻易从事。这种双语语料库的建设往往容易陷入步履维艰的困境。

尽管目前还存在这样或那样的困难, 国内外平行语料库的研制仍然取得了长足的进展。这些, 都为基于平行语料库的语言研究提供了非常有用的语言资源。

基于平行语料库的语言研究与基于语言学家的语感和个人语言经验的语言研究有着本质的不同。

在基于语言学家的语感和个人语言经验的语言研究中, 我国前辈语言学家曾经提出“例不十, 法不立, 例外不十, 法不破”的原则来衡量语言学规律是否正确。这个原则常常作为判断语言学家治学态度是否严谨的准绳, 广为流传, 为海内外语言学界所啧啧称道。然而, 在大规模真实文本的语料库中, 要找出十个例子来证明某个规律或者找出十个反例来推翻某

个规律是轻而易举的事，通过计算机的检索，人们在很短的时间内就可以找到成百甚至成千的例子或反例来作为论证的根据。在语料库语言学（corpus linguistics）的研制中，我们要处理成千上万的语言实例，也同样要处理成千上万的语言反例，十个例子或十个反例，在大规模真实文本的语料库中，只不过是沧海之一粟，难免会产生管窥蠡测之弊。从语料库语言学的观点看来，在语言研究中，还存在着另外一条原则：“例过十，法未必立，例外过十，法未必破”，语料库语言学要求研究者用概率和分布的科学计算来取代简单的枚举实例，不主张以几个或者几十个例子或反例就轻率地确立语言学的规律。因此，对于“例不十，法不立，例外不十，法不破”这一条我国语言研究中备受推崇、广为传播的治学原则，我们就有必要用新的眼光重新来认真地审视一番了，这也许有助于把传统的语言学研究由内省式的简单枚举转到现代科学的语料库语言学方法的原则上来，以推动我国语言研究的现代化进程。

当前国外语料库语言学正在迅速地发展，我们中国的语言学工作者在借鉴别人的成果和方法的时候，固然应该继承我国前辈语言学者的成果和方法，这是无可非议的，此外，我们还应扩大我们的视野，放眼世界，学习一些国外当代学者的成果和方法。我们既要以前辈学者为师，我们也要以国外的当代学者为友，这样，我们就有可能站在我国前辈学者和国外当代学者的肩膀上，看得更为辽远，看得更加清楚，从而把我国的语言学推向前进。

瞿云华博士的这本专著，为使用平行语料库来研究汉语和英语的对比进行了初步的尝试，取得了可喜的成绩。她不是仅仅依靠十个例子或者十个反例就轻率地确立语言规则，而是根据 500 多万字的汉英双语对齐的平行语料库来探索汉语动词和英语动词的视点体的规则，并试图把这些规则加以形式化，以便在机器翻译和语言信息处理中使用。这是很有意义的探索。希望作者在此基础上，更上一层楼，为我国语料库语言学的发展做出新的成绩。

2008 年 7 月 18 日，于北京