

原文载《现代外语（季刊）》，第33卷，第4期，2010年11月

语料库语言学与中国外语教学

【编者按】为推动我国应用语言学研究的新发展，教育部人文社科重点研究基地——广东外语外贸大学外国语言学及应用语言学研究成功于2010年9月24日至25日成功地举办了“首届广外应用语言学论坛”，庆贺我国著名语言学家、应用语言学学科的开拓者桂诗春教授八十华诞。期间，举行了题为“语料库语言学与外语教学”的高层论坛，就语料库语言学的现状、发展前景及语料库的建设、共享、应用等展开了互动讨论。以下是根据专家发言，整理后的主要内容。

双语语料库的建设与用途

冯志伟 国家教育部语言文字应用研究所

1. 双语语料库的建设

我很赞同桂诗春教授的意见，积极推进语言资源的共享，语料库只有共享才能变成财富，如果把语料库的研究成果“藏诸名山，束之高阁”，只是一堆数据垃圾，必将自毁前程。桂诗春教授刚才提到宾西法尼亚大学的 Linguistic Data Consortium（我建议最好翻译为“语言数据联盟”，简称 LDC），是一个很好的供语料库语言学者进行交流互动的平台。在语言数据联盟和其他相关机构的帮助下，研究者们可以获得口语和书面语的大规模的语料。重要的是，在这些语料中还包括一些标注过的语料，如宾州树库（Penn Treebank），布拉格依存树库（Prague Dependency Tree Bank），命题库（PropBank），宾州话语树库（Penn Discourse Treebank），修辞结构库（RSTBank）和 TimeBank（我不知道 TimeBank 这个名称如何翻译为中文）。这些语料库是带有句法、语义和语用等不同层次的标记的标准文本语言资源。这些语言资源的存在大大地推动了人们使用“有监督的机器学习方法”（supervised machine learning）来处理那些在传统上非常复杂的自动句法剖析（automatic syntactic parsing）和自动语义分析（automatic semantic analysis）等问题。这些语言资源也推动了有竞争性的评测机制的建立，评测的范围涉及到自动剖析（parsing）、信息抽取（information extraction）、词义排歧（word sense disambiguation）、问答系统（question-answer system）、自动文摘（automatic summarization）等领域。

几年前由中国中文信息学会发起，在北京创建了一个“中文语言数据联盟”（Chinese Linguistic Data Consortium，缩写为 CLDC），是一个自愿组成的学术性社会团体，其宗旨是团结

中文语言资源建设领域的广大科技工作者，建成代表中文信息处理国际水平的、通用的中文语言和语音的资源库。欢迎语言学界的同仁积极参与 CLDC 的工作，促进语料库资源的共享。

目前单语语料库很多，已经取得了煌煌的成绩，但是双语并行语料库（parallel corpus）不容易获得，它的构建和加工是很困难的工作。我国还没有高质量的、大规模真实文本的英汉双语语料库，更没有成熟的、可共享的加工工具，最近公布的 2010 年国家自然科学基金重大项目中有一项就是“大规模英汉平行语料库的构建与加工研究”，资助强度大约是 50 万元左右，可见国家对于双语语料库建设的重视。这个项目是我和王克非教授在今年的社科基金评审会议上建议提出的，已经开始招标，希望大家积极投标，积极推进我国的双语语料库建设。

2. 如何将语料库语言学运用到外语教学，如何从语料库中挖掘知识？

我认为英汉双语语料库的最大用途就是推进英语教学，我们可以从双语语料库中抽取教材的原材料，帮助语言学习者提高对于真实语言材料的语感，从而编写出高质量的外语教材。有的外语老师冥思苦想地根据自己的语感来编写教材，费时费力，其实，如果依靠英汉双语平行语料库，就可以减轻搜集素材之困难，大大提高编写教材的工作效率。

另外，语料库中蕴藏着无比丰富的知识等待我们去挖掘，如果我们使用“文本数据挖掘”（text data mining）的技术，从语料库中挖掘知识，既可以挖掘语言学的知识，也可以挖掘非语言学的知识，就像从矿石中挖掘出黄金一样，这些知识可以弥补传统语言学的不足，克服研究者的主观性和片面性。我们在 text data mining 这个术语中使用 mining（挖掘）这个单词，而没有使用 extraction（抽取）这个单词，正是为了强调在从语料库中获取知识的时候，要开动脑筋，要经过一番“去粗取精，去伪存真，由此及彼，由表及里”的深思熟虑的功夫来加工数据，而不要被海量的数据所迷惑。数据就像矿石，我们的任务是从海量的数据中挖掘出隐藏在其中的有规律性的东西，把海量的、离散的“数据”（data）变为精炼的、系统化的“知识”（knowledge），从而把经验主义方法和理性主义方法紧密地结合起来。这种知识获取方法上的巨大变化，有可能引起整个语言学研究的“战略转移”（strategy transit）；我们中国的语言学家应当敏锐地关注“战略转移”问题，做出我们的应有的贡献，千万不要错过这个在语言学历史上千载难逢的良机。