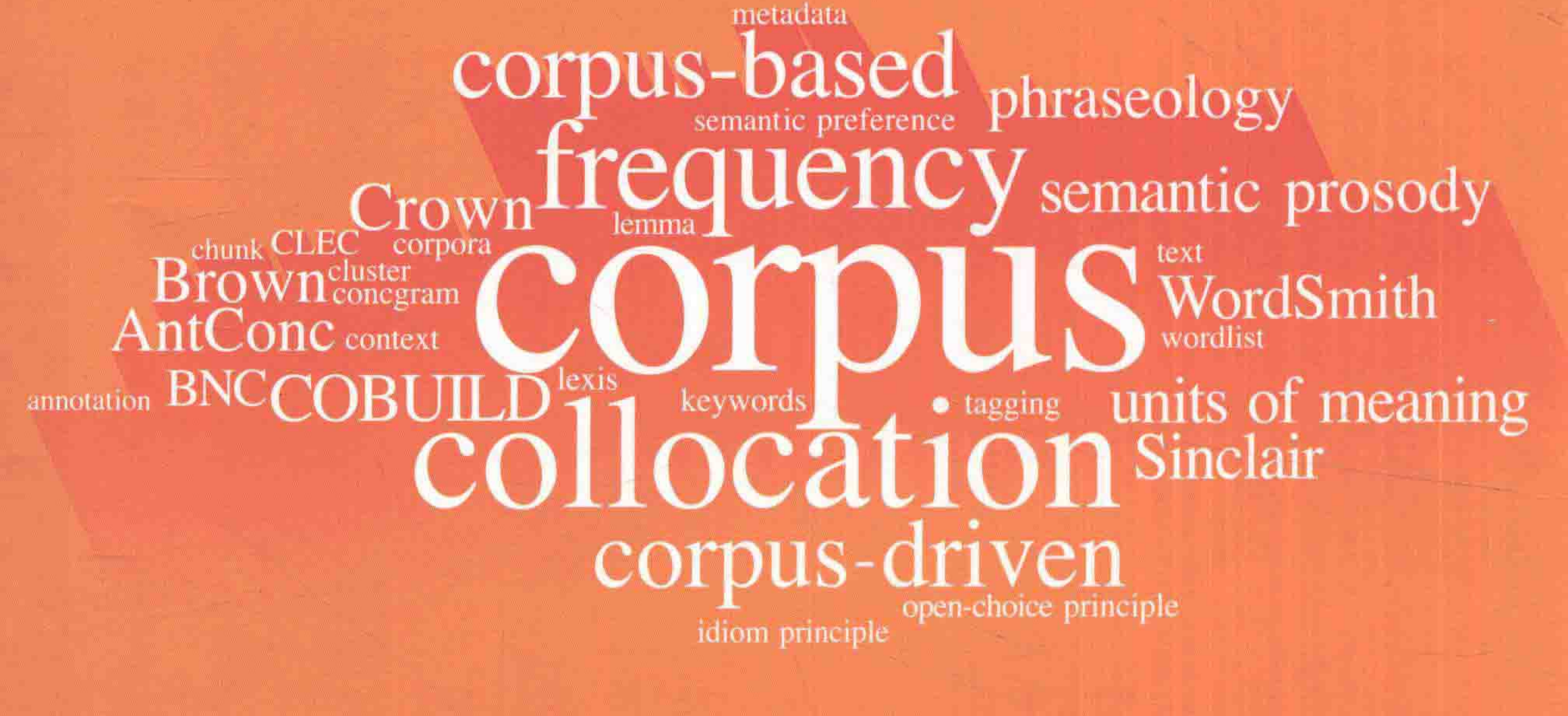
《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

语料库语言学

CORPUS LINGUISTICS

2
Vol. 2 No.2
第2卷 第2期
2015

北京外国语大学中国外语教育研究中心
梁茂成 许家金 主编



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

图书在版编目 (CIP) 数据

语料库语言学．2015．2 ：汉、英 / 梁茂成，许家金主编．— 北京 ：外语教学与研究出版社，2016.2
ISBN 978-7-5135-7178-4

Ⅰ．①语… Ⅱ．①梁… ②许… Ⅲ．①语料库－语言学－汉、英 Ⅳ．①H0

中国版本图书馆 CIP 数据核字 (2016) 第 040599 号

出 版 人　蔡剑峰
责任编辑　毕　争　解碧琰
封面设计　外研社设计部
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　中国农业出版社印刷厂
开　　本　787×1092　1/16
印　　张　7.5
版　　次　2016 年 3 月第 1 版　2016 年 3 月第 1 次印刷
书　　号　ISBN 978-7-5135-7178-4
定　　价　12.00 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所　刘旭东律师
　　中咨律师事务所　殷　斌律师
物料号：271780001

语料库语言学
（半年刊）

Corpus Linguistics
(Biannual)

主　管：中华人民共和国教育部
主　办：北京外国语大学
承　办：中国外语教育研究中心
出　版：外语教学与研究出版社

Administered by the Ministry of Education of China
Directed by Beijing Foreign Studies University
Edited at the National Research Centre for Foreign Language Education
Published by Foreign Language Teaching and Research Press

电　　话：（010）88816828
电子邮箱：bfsucrg@sina.com
投稿网址：http://ylyy.chinajournal.net.cn

本刊地址：北京市西三环北路19号北京外国语大学中国外语教育研究中心《语料库语言学》编辑部（100089）

《语料库语言学》
2015年　第2卷　第2期

目　录

学者聚焦

同题共议

研究论文

研制开发

书刊评介

会讯动态

CORPUS LINGUISTICS

Volume 2, Number 2, 2015

Table of Contents

肖忠华语料库语言学答客问

浙江大学 肖忠华

编者按

《语料库语言学》创刊号有幸登载了桂诗春先生的个人学术访谈。桂先生定稿时自拟题名《语料库语言学答客问》，本刊欣然从之。本期所刊肖忠华教授访谈，仍沿用《语料库语言学答客问》，并缀以受访者姓名，以示区分。据此，期刊数据库收录，读者文献查询时，可免于混淆。

肖忠华教授（1966–2016）是国际知名的语料库研究学者，是华人语言学研究学者的杰出代表。他师从英国兰卡斯特大学Tony McEnery教授，2002年获得语料库语言学博士学位。他的研究领域涉及基于语料库的英汉对比与翻译研究、汉语研究、英语研究、时体理论、语言教育及二语习得等。肖教授著述量多质优，尤其在基于语料库的英汉对比与翻译研究以及汉语研究方面的成果突出。很多论著为相关领域必读必引之作。2016年1月2日，肖教授不幸因病去世。

肖教授生前于病榻之上完成我刊书面访谈，深谈国内外语料库研究进展和个人学术历程，我刊同仁由衷感佩。谨以此文纪念并深切缅怀肖忠华教授。

1. 您最早是什么时候开始接触语料库的？您能描述一下当时国际国内语料库研究开展的情况吗？

我最初接触“语料库”的概念，是在20世纪80年代中期读大学本科的时候。我对英语语法比较感兴趣，所以喜欢研究夸克等人编写的《当代英语语法》和《英语语法大全》，发现这些原版著作对英语语法的描述及其例句和张道真《实用英语语法》等当时国内流行的英语语法之间一个很大的差别就在于，夸克语法更接近真实的语言。当时，我并不知道语料库这个名称，只是了解到夸克语法是以夸克等人建立的“英语用法调查”（Survey of English Usage，SEU）数据库中所收集的英国人实际使用英语的素材为基础的。

真正开始接触“语料库语言学”这个术语，是在1999年联系到英国攻读博士学位的时候。由于一直对英语语法感兴趣，就联系了当时在兰卡斯特大学任教的夸克语法作者之一的Geoffrey Leech教授。由于Leech当时已从讲座教授退休改为研究教授，不再接收新的博士生，所以他把我推荐给了Tony McEnery教授（当时其职称为Reader in Multilingual Corpus Linguistics）。这是我第一次听说“语料库

语言学”这个名称，了解到语料库语言学是用计算机来分析人们实际使用的真实语言，不仅采用传统语言学中的定性分析方法，而且采用数理统计方法对语言的使用作定量分析。由于我本科和研究生读的都是英语和语言学专业，对语言学和数理统计相结合的研究感到十分新奇，而且我对计算机一直很感兴趣，所以就同意从英语语法转为语料库语言学方向。当时，上海教育出版社刚引进出版了《牛津应用语言学丛书》一套28册，其中包括John Sinclair的《语料库、索引与搭配》（*Corpus, Concordance, Collocation*），这是我读到的第一本专门研究语料库语言学的著作。

当我在2000年初到英国兰卡斯特大学开始博士研究时，我对语料库语言学的了解差不多是零起点，第一年只好开始恶补语料库语言学、统计学、计算机编程三大块的知识。当时，该领域除了McEnery & Wilson（1996，2001）的《语料库语言学》等少数专著外，大多数语料库研究基本都是以论文集的形式出版的，这是因为20世纪八九十年代还很少有期刊接受和发表语料库方面的论文。当时，采用语料库的研究方法尚未像十多年后的今天那样普遍为人们接受而显得理所当然，还可以听到各种反对声音（如Widdowson 2000；Newmeyer 2003）。积极倡导语料库语言学的学者（如Sinclair和Leech）对语料库的建库原则和分析方法存在意见分歧。

虽然多语种语料库已于20世纪90年代中后期开始得到了发展（如英语-挪威语平行语料库），但在新世纪初，当人们提到语料库语言学时，基本上是指英语语料库语言学，这是因为在统一码（Unicode）应用于文字编码之前，安装与统一码兼容的Windows 2000之前操作系统的计算机只能处理ASCII编码的语言，除非支持特定的字符集。当时国际上应用最广泛的语料库是英国国家语料库（BNC）和由ICAME发行的包括Brown、LOB、Frown、FLOB在内的语料库光盘。语料库检索与分析软件包括基于DOS的Longman Mini Concordancer与WordSmith 3.0版。由于当时语料库分析工具相当简陋，所以学习语料库语言学基本上都需要学习编程才能满足自己的研究需要。我最初学的编程语言是Perl（当时还没有现在很流行的编程语言Python和R），该语言的正则表达式功能强大，而且非常适合语料库建库和分析。随着学界对语料库语言学兴趣的升温，兰卡斯特大学发起了每两年举办一次的“国际语料库语言学大会”，第一届于2001年召开，即CL2001，到2015年已是第八届了。

在国内，虽然上海交通大学杨惠中教授的团队于20世纪80年代早期就已开始研制科技英语语料库（JDEST），随后石油大学广州分院的祝启波也建了石油英语语料库（GPEC），但即使是在语言学界，了解语料库语言学的人也非常少。记得当时国内有人问我在英国读什么专业，我说是Corpus Linguistics，人家还以为跟尸体有关而感到很恶心。值得一提的是，台湾“中研院”黄居仁、陈克健团队于20世纪90年代中期就成功研制了第一个带词性标注的现代汉语平衡语料库，并在网上对公众开放。

2. 语料库研究的哪些特点最吸引您?

语料库语言学借助自然科学的实证研究方法，利用计算机软件对大规模真实语言数据进行分析，不仅包括传统的定性分析，而且还采用数理统计方法对语言进行定量分析。需要特别指出的是，语料库语言学不像转换生成语法等传统语言研究那么依赖于研究者的语言直觉，而是主要依靠真实语料的实证数据，但同时又不排斥语言直觉，两者有机结合。

语言学研究中常用的数据来源有两类，即真实语料和研究者的语言直觉。语言分析当然离不开语言直觉。例如，语言直觉可用来造句（不管是正确还是错误的例句）用于语言分析，也可用来判断某一表达方式是否可接受或合乎语法。研究者在需要时可立即利用直觉通过内省来编造更纯的例句，这是因为语言直觉随手可得，而且编造的例句不像人们在真实语境中使用语言那样受语言外部因素干扰。从某种意义上甚至可以说，语言直觉在语言学研究中是必不可缺的，因为对语言现象的分类通常涉及基于直觉的判断，而这种分类在构建语言理论时不可避免。然而，正如Seuren（1998：260-262）所述，语言直觉必须谨慎使用。

首先，语言直觉可能会受到个人的地域方言或社会方言影响（Krishnamurthy 2000a：172）。结果就是，一句话对某个人来说不合语法或不可接受，而对另一个人来说却完全正确。因此，我们常可发现在语言学文献中，对某些例句的可接受性争论不休。其次，研究者编造例句来支持或驳斥某一论点时，同时在有意识地监控自己的语言产出。因此，即使其语言直觉是正确的，编造出来的例句也不能代表典型用法。第三，基于语言直觉通过内省得到的语言数据脱离语境，因为它存在于内省者头脑中而非真实语境中，而要判断一句话是否合乎语法或可以接受，语境至关重要。有了合适的语境，即使是脱离语境时显得不合语法或不可接受的语句也有可能会变得合乎语法或可以接受，而人们的想象力十分丰富，即使是最不可思议的话语，也可以想象出可能的语境（Krishnamurthy 2000b：32-33）。第四，基于语言直觉的研究结果很难验证，因为研究者是在头脑中通过内省来造句，无法直接观察。第五，过分依赖直觉会使研究者对语言使用的现实视而不见（Meyer & Nelson 2006）。例如，由于罕用词或不常见的用法具有心理上的突显性（Sinclair 1997：33；Krishnamurthy 2000a：170-171），人们更倾向于注意到不常见的语言现象而又对普通现象熟视无睹。最后，在语言学的某些研究领域中（如语言变异研究、历时语言学、语言习得等等），研究者无法可靠地使用个人的语言直觉，而必须依赖于语料库数据（Meyer 2002；Léon 2005：36）。

通过内省得到的语言数据基于研究者个人的语言直觉，而语料库数据则截然不同，它汇集了许多语言使用者的语言直觉。语料库中的书面语或口语语料样本源自于真实语境中使用的自然语言。由于人们在真实语境中使用语言也是基于自己的语言直觉，可以说语料库也是基于语言直觉的，但它比内省式的语言数据更

加自然，因为它是用于实际的交际目的而不像后者那样是编造出来用于语言分析的。与研究者个人通过内省得到的语言数据相比，语料库数据一般能反映出更多语言使用者的语言直觉。语料库方法还能很容易地提供语言现象的频数，而这很难利用语言直觉可靠地预测（McEnery & Wilson 2001：15）。正因为如此，语料库能使研究者克服自身语言直觉中的偏颇，并使之能够辨别哪些是具有统计意义的典型语言现象，哪些是随机现象。总之，语料库不仅能提供业已验证的、带有语境的定量数据，而且有助于识别语言直觉无法觉察的用法差异（Francis, Hunston & Manning 1996；Kennedy 1998：272）。此外，语料库方法还在过去30年间拓展或突出了语言学中一些无法只通过语言直觉来研究的新领域（如语体变异研究）。

语料库研究的这些特点使之有别于传统的语言研究，并更能取得可靠的研究结果。正如Leech早在20世纪90年代初指出的那样，“50年代的语料库语言学家拒绝语言直觉，而60年代的普通语言学家拒绝语料库数据。两者均未获取近年来许多成功的语料库分析所涉及的数据覆盖面和所取得的精辟见解”（Leech 1991：14）。正因为具备这些优势，语料库方法不仅成为语言学领域的标准研究工具，而且已开始逐渐成为基于文本的人文社科领域中重要的研究工具[1]。

3. 有没有哪（个）些学者或某（个）些论著在语料库研究方面对您影响较大？如有的话，您能说说影响主要体现在什么方面吗？

我最初的语言学研究兴趣是英语语法和语义学。正式接触语料库并系统研究语料库语言学，是2000年初到兰卡斯特大学攻读博士学位才开始的，在此之前对语料库研究知之甚少。因此可以说，在语料库研究方面对我影响最大的是以Leech和McEnery为代表的兰卡斯特语料库语言学传统。

一般认为，在语料库语言学内部有两个不同的取向，即“基于语料库”和“语料库驱动”，或称“语料库作为方法”和“语料库作为理论”（McEnery & Hardie 2012），分别以Leech为首的兰卡斯特团队和以Sinclair为首的伯明翰团队为代表。两者在语料库的性质（即语料库语言学是方法还是理论、对待语言直觉和语料库前理论的态度）、语料库建库（如语料库的平衡性与代表性、语料采用全文还是抽样、语料库标注）、语料库分析（如基于语料库或语料库驱动、推断统计在语料分析中的作用）等方面都存在意见分歧（McEnery, Xiao & Tono 2006；McEnery & Hardie 2012）。当然，两大派别之间的对立存在着人为夸大的因素（Xiao 2009a：993）。再者，随着时间的推移，继承Sinclair和Leech语料库研究传统的两派语料库语言学家之间目前已有较大程度的融合，双方取长补短。

除了兰卡斯特传统，Biber（1988）的多维度分析法对我的语料库研究也有较大的影响。多维度分析法最初用于分析英语口语和书面语之间的语体差异，但在

过去近30年中发展迅速并得到了广泛运用。我在这方面的研究主要集中在3个方面，即世界英语、科技论文摘要、翻译共性（Xiao & McEnery 2005；Xiao 2009b；Cao & Xiao 2013；Hu, Xiao & Hardie forthcoming）。

4. 您如何评价中国语料库研究在过去若干年的发展以及目前的现状？

目前布朗语料库被公认为第一个电子英语语料库，Quirk等人在伦敦大学学院于1959年开始建立的“英语用法调查”也被称为现代语料库语言学研究的鼻祖[2]。然而，由于汉语具有汉字众多的特点，尽管当时还没有语料库这个名称，但汉语研究早就具有采用真实语料来确定常用字词的传统。例如，我国第一个现代意义上的汉语字频统计，即黎锦熙的《国语基本语词的统计研究》，早在1922年就已发表。教育家陈鹤琴及九名弟子花了3年时间收集并分析了6类“语体文”语料共计形符554,498字，类符4,261字，并对频数为5,000、3,000、2,000和1,000以上的频段进行统计，发现这些频段的字数分别为10、19、38和100以上，其结果于1922年发表在《新教育》第5卷第5期，其修订本由商务印书馆于1928年重新出版为《语体文应用字汇》。黎锦熙和陈鹤琴的汉语字频研究无疑为我国基于语料库的词汇研究开了先河。

随着语料库语言学在英美等国逐渐兴起，以及计算机中文信息处理技术的改善，语料库研究也从20世纪80年代开始在我国得以开展，并在过去近20年中得到了迅猛的发展。我国的语料库研究主要集中在以下3个方面：汉语语料库与中文信息处理、学习者语料库与汉语中介语语料库、汉英双语平行语料库。第一类汉语语料库大多是由计算机专业研究者所建的专门用途语料库，缺乏平衡性，主要服务于中文信息处理而非语言学研究。第二类语言教学用语料库研究主要由高校外语教师和对外汉语教师承担，其中学习者语料库主要是专业和非专业英语学习者语料库，收集的语料大多为历年英语等级考试材料，而汉语中介语语料库主要包括日、韩、泰国等亚洲国家在华留学生的作文和口语材料。第三类双语平行语料库建设主要与过去10年左右我国开展语料库翻译学研究密切相关。

语料库语言学在中国的迅速发展，主要得益于政府与学术机构的大力支持以及高校等学术组织对语料库研究方法的推广普及。例如，近10年来，由国家社科基金资助，包括重大课题在内的批准项目每年都有差不多20个，出版社与语言学专业期刊也越来越愿意发表语料库研究成果。近年来国内许多高校都为语言学专业研究生开设了语料库语言学课程，北京外国语大学中国外语教育研究中心和上海交通大学也为高校教师和研究生等开设了多期语料库语言学研修班。另外值得一提的是，由中外学者的民间力量自发组织开发并维护的www.corpus4u.org网站，自建站10年来为语料库研究在我国的推广和发展起到了十分重要的作用。

虽然我国的语料库研究在新世纪得到了长足的发展，但目前还存在不少问题。

首先是学科之间沟通合作不足。语料库语言学涉及语言学、计算机、数理统计等多个学科的专业知识，学科之间的合作不仅能拓宽研究思路、提高研究质量，而且对当今大数据时代的研究来说发挥着越来越重要的作用。而在我国，研究语料库的两个研究群体，即研究汉语语料库和中文信息处理的计算机领域和主要研究外语语料库的外语教学与研究领域（包括涉及汉语的语言对比与翻译研究），由于其研究目标不同，两者之间很少有相互的研究合作。在2011年5月由香港教育学院主办的“汉语语料库及语料库语言学”圆桌会议上，国内的与会者大多是中文信息处理和汉语研究方面的专家。当我提到“中国语料库语言学研究会”，几乎没有人知道或承认这个语料库协会，说这是外语教师的一个组织吧。其实，研究语料库的语言学家与计算机专家之间的合作对双方都有利。一方面，语言学家的参与能使语料库更具有代表性，而另一方面，计算机专家的投入能使语料处理效率更高、语料加工也更具深度。在这方面，兰卡斯特大学的UCREL和CASS语料库研究中心的工作开展得卓有成效。UCREL研究中心的研究人员包括语言学系和计算机系对语料库研究感兴趣的老师，双方相互合作取长补短，承担了包括英国国家语料库（BNC）在内的不少大型研究项目。由“英国经济社会研究理事会”（ESRC）投资430万英镑成立的CASS语料库研究中心更是以语料库为共同研究平台，聚集了语言学、计算机、心理学、医学、历史学、社会学、政治和财经等众多学科的专家，从多学科角度对各种社会问题进行研究。这种学科之间的紧密合作值得我国语料库研究者借鉴。

其次，重复投资、资源利用率不高。虽然国内每年都有许多语料库建设项目得到国家或省部级的资助，但建成的语料库大多仅供内部使用，有些项目建而不研，有的建成后束之高阁。其结果是语料库资源利用率不高，从而引起重复投资和浪费。当然，有些语料库是由于包括大量全文引起版权问题而使得对外开放资源受到限制，但此类版权问题从项目一开始，进行语料库设计时即应加以考虑。其实，只要语料库设计合理，并与版权方充分沟通，这些问题是可以解决的。例如，美国的语言数据协会（LDC）、欧洲语言资源协会（ELRA）和牛津文本档案库（OTA）都发布了大量的语料库资源，其版权问题都得到了妥善解决。要提高语料库资源的共享度，我建议有关部门出台规定，凡是得到国家和省部级资助的纵向课题产生的语料库都必须在结题后一定时间内（如6个月的保护期后，以便项目组享有数据的优先使用权）将资源向公众开放。英国研究理事会的数据政策规定，所有资助项目产生的数据资源必须在项目结束后公开[3]。我国可以借鉴这一做法。

再次，从国内出版和发表的研究成果来看，绝大多数语料库质量不高，语料分析也缺乏深度和系统性；发表的论文翻译引介国外研究的多，而实证研究少。语料库研究质量不高与我国语言学界流行的“一窝蜂上”这一通病有关。从最初

的转换生成语法到系统功能语言学，再到现在的语料库语言学，都存在这个问题。从www.corpus4u.org网站上的提问和讨论来看，国内有不少早期职业研究者，对语料库一知半解，甚至缺乏最基本的语料库知识和分析技能，都在用语料库方法作研究写论文。其实，语料库只是研究方法的一种，而且这种方法不是万能的。有些研究问题用其他方法来研究效率更高。只有弄清楚语料库能用来做什么，不能做什么，如何针对特定的研究问题建立或选择合适的语料库，使用什么工具，以及特定软件的哪些功能，采用哪些统计分析手段，如何将语料库证据和包括语言直觉和其他学科知识在内的资源结合起来，才能够产出高质量的语料库研究。

最后，我国的语料库研究基本上都在国内的中文期刊上发表，而很少有论文发表在高档次的国际期刊上，缺少与国际学术界的互动与交流，以至于国际学术界对中国的语料库研究知之甚少。其实，我国的语料库研究在某些方面（如汉语语料库的加工，涉及汉语的双语平行语料库研究）还是处于国际领先地位的[4]。各高校和科研单位应改革并完善业绩评定与奖励机制，鼓励作者走出去在国际上出版和发表自己的研究成果，让世界听到来自中国的声音，了解我国的研究现状。近年来，我国的学者在这方面已开始取得一些进展（如Tsou & Kwong 2015；Xiao & Hu 2015；Xiao & Wei 2014；Zou, Hoey & Smith 2015；Hu & Kim forthcoming）。

5. 您对中国语料库研究今后发展有什么样的建议和希望?

从上述对我国语料库研究现状的讨论可以看出，今后的发展应该考虑以下几个方面。首先是要加强学科间的研究合作，发展跨学科研究。这种合作有利于语料库研究的深入开展，同时也是基于大数据的研究所必需的。第二，加强纵向项目数据管理，实现数据共享。一个好的语料库通常是可反复利用的资源，而且可以满足多种研究目的，但创建一个好的语料库常常既费时又耗资。根据不同的研究目的实现数据无偿或有偿共享，有利于节省研究时间和资金的投入。第三，加强研究梯队建设，提高研究质量。老一代成熟的研究人员要发挥传帮带的作用，有计划地培养早期职业研究人才，避免一窝蜂上的局面，建立语料库研究梯队，形成我国语料库研究的后劲以利于长期发展。最后，我国的语料库研究要立足国内，并走向世界。中文是世界上使用人数最多的语言，用中文发表研究成果本来无可厚非，但英语作为国际通用的科技和出版语言有利于世界各地的学者进行交流。实际上，有许多非英语国家的作者都是直接用英语发表论文的。我们应鼓励作者把国内包括语料库研究在内的顶级科研成果发表在高档次的国际期刊上；同时把国内发表的优秀论文全文译介到国际上以便交流。在译介我国优秀论文方面，中国知网已成立国际出版中心（http://tp.cnki.net），旨在通过组织高水平的编辑和翻译人员，精选优秀学术期刊中的论文进行汉译英翻译并在线同步出版，以全面

提高国际同行对我国社科领域最新研究成果的了解和认同，进一步提升中国优秀学术成果的海外影响力。

6. 您能谈谈中国语料库研究在国际语料库研究学界应如何自我定位吗？

我国语料库研究在国际上的自我定位，应该遵循“扬我所长、以研促用”的原则。前者是要充分利用自身的优势，后者是要提高研究的实用价值。具体地说，首先是研究我们的母语汉语。到目前为止，基于语料库的汉语研究基本上以现代汉语书面语为主。今后的研究可以更加注重以下几个方面。一是在平衡语料库的基础上更系统地研究现代汉语口语，并对口笔语语体进行比较。二是研究过去20年来随互联网与通讯技术发展而新出现的语体（如社交媒体）。这些新语体具有自身的语言特点，但现有的汉语平衡语料库基本上都没有包含在内。三是研制包含汉语发展各主要阶段的历时语料库。汉字是世界上最古老的文字之一，创建能反映汉语发展史的历时平衡语料库，不仅对我国古籍研究大有裨益，而且也能为自古以来中外语言接触和文化交流的研究提供研究素材和实证依据。四是创建汉语方言语料库。我国具有丰富的语言资源，各地方言多达230多种，对语言接触和语言类型学研究具有十分重要的意义；而对于那些濒危方言，建立语料库则更能起到保护和保存作用。五是开发新的适合汉语并针对汉语特点的语料分析方法和工具。

其次是研制包括可比语料库和平行语料库在内的多语种语料库，开展中外语言对比与翻译研究。涉及像英语、汉语这样大跨度语言之间的语言对比和翻译（包括口译）研究对于语言学理论具有重要意义，而针对主要外语语种和非通用语种的此类研究对外语教学具有指导意义。

第三，开发教学用语料库资源，开展基于语料库的二语习得研究。教学用语料库是指我国各类学生学习外语的学习者语料库和外国人学习汉语的汉语中介语语料库。学习者语料库是语料库语言学中一个比较成熟的研究领域。我国在过去10年中已建成不少此类语料库，但还存在一些问题。比如，现有学习者英语语料库包含的基本上都是各类英语等级考试材料，而现有汉语中介语语料库基本上都只包括韩国、日本、泰国等亚洲国家留学生的语料。目前教学用语料库研究存在的另一个问题是建而不研。语料库建完了项目也就算结束了，而没有对语料进行深入系统的分析，将研究成果用来指导、促进实际的教学工作。教学用语料库研究今后在语料平衡性（包括语料类型和来源等）和研用结合方面尚有待改进。

第四，开展基于多语种平行语料库和可比语料库研究，开发机助翻译、翻译记忆库、多语种术语库等应用产品，并提高机器翻译和自动文摘等应用系统的可靠性和有效性。

最后是利用语料库技术，针对网络诈骗欺凌等社会问题，开展司法语言学研究。网络欺凌在脸书（Facebook）和推特（Twitter）等国外社交网站屡见不鲜，国内的网络诈骗也同样层出不穷、防不胜防。开展此类研究对于防范这类社会问题具有十分重要的社会意义。

总之，“扬我所长”主要是指这前两类研究，而“以研促用”主要指后三类研究。

7. 您如何评价您个人对语料库研究发展的贡献?

贡献可能谈不上，不过在过去10多年中，自我感觉还是在基于语料库的语言研究方面脚踏实地、认认真真地做了一些令自己满意的研究。

我的主要研究领域是语言对比与翻译研究，特别是语料库翻译学和基于语料库的英汉对比研究（如Xiao 2010a）。我出版了国际上第一本基于语料库的英汉对比研究专著（Xiao & McEnery 2010）。我于2006年在*Applied Linguistics*上发表的论文（Xiao & McEnery 2006）从语言对比角度探讨了英汉语中的搭配和语义韵，也具有较大的影响。由本人发起两年一届的“基于语料库的语言对比与翻译（UCCTS）”国际研讨会颇受欢迎，到2014年为止已在中国、英国和比利时成功举办4届。在语料库翻译学方面，我近年来的研究从英汉翻译和翻译体汉语的视角重新审视了以往主要局限于英语及其相近语言的翻译共性假设，对英汉翻译中翻译体汉语的系统研究（Xiao 2010b, 2011, 2015；Xiao & Dai 2014；Xiao & Hu 2015；戴光荣、肖忠华 2011；肖忠华、戴光荣 2010；肖忠华 2012）对于描写翻译学和翻译共性研究具有至关重要的意义。

我的另一个重要研究领域是汉语语料库语言学。我于2004年出版的*Aspect in Mandarin Chinese*（Xiao & McEnery 2004）是世界上第一本在真实语料基础上系统阐述汉语时体系统的专著，其学术价值得到了众多书评的认可。我在过去10多年来所建的一系列汉语语料库和平行语料库基本上全部向学术界免费公开（如LCMC、ZCTC、UCLA2、Babel）[5]，在国际上广为应用。

在语料库分析方法创新方面，我提出的多维分析框架对Biber（1988）的模型进行了扩展，在原有语法分析的基础上增加了语义分析和类联接分析，并将多维分析模型首次应用于世界英语比较和科技论文摘要的对比分析（Xiao 2009b; Cao & Xiao 2013），最新的研究又将多维分析引入了翻译共性研究领域（Hu, Xiao & Hardie forthcoming）。

在语料库语言学教学方面，由本人主笔合著的*Corpus-based Language Studies*（McEnery, Xiao & Tono 2006）是目前最流行的语料库语言学教材，被美国教育部指定为应用语言学必读参考书，并为世界各地70多个研究生课程和本科生课程所采用。我还参与了慕课课程*Corpus Linguistics: Method, Analysis, Interpretation*的教

学，主讲多语种语料库及其应用，该课程由兰卡斯特大学和Futurelearn推出[6]，前两期学员人数已超过6,000人。过去10年左右我投入较多时间和精力参与建设和管理的www.corpus4u.org网站产生了较大的影响，为语料库研究在我国的推广普及发挥了重要作用。

最后，通过学术兼职为国际语料库研究领域服务。本人多年来兼任*International Journal of Corpus Linguistics*、*Corpora*、*Chinese Language and Discourse*、*Languages in Contrast*等8种学术期刊的编委和近30家期刊和出版社的审稿人，以及英国社会经济研究理事会（ESRC）、英国艺术与人文研究理事会（AHRC）、美国国家科学基金会（NSF）、加拿大社会科学与人文研究理事会（SSHRC）、葡萄牙科学技术基金会（FCT）、中国香港研究资助局（RGC）等多个国家和地区研究基金的项目评审专家。此类学术兼职不仅使自己清楚地了解国际语料库研究的前沿动态，而且能提高国际学术界发表论文的质量。

8. 在您看来，从事语料库研究应具备哪些方面的学科素质？您对从事语料库研究的年轻学子有什么样的忠告？

语料库是语言研究中一种十分有用的工具和资源。虽然我们在前文已讨论过使用语料库方法的种种优势，但跟所有工具一样，语料库不是万能的。首先，一个语料库不可能包括一种语言的所有语句，抽样就不可避免，因而语料库涉及代表性的问题。目前还没有可靠的科学手段来保证语料库的代表性。用Leech（1991：27）的话来说，语料库的代表性仍然是一种“信仰行为”。换言之，当一个语料库的规模和覆盖面达到一定程度时，人们对其代表性的信心就会增加。其次，需要用更复杂、更严格的统计方法来分析语料库数据。在语料库研究中，定量分析与定性分析同等重要。目前语料库研究中许多常用统计方法假设数据呈正态分布，而在语言运用中正态分布并不普遍。因此，我支持Gries（2006）所提出的“更严格的语料库语言学”这一观点。第三，语料库不能提供反面证据。一个语料库不管多么大、多么平衡，除非它代表高度专门化的语言，都不可能穷尽一种语言中的所有语句，因为语言本身就是无穷尽的。因此，语料库不能告诉我们语言中哪些现象可能，哪些不可能。比如，如果你没有在语料库中找到某个结构，也不能说该结构在语言中不存在[7]；同样，也不能说在语料库中能找到的结构就一定合乎语法或可以接受，因为语料库数据属于语言使用数据（performance data）而有可能包含语误。最后，虽然语料库方法可以帮助我们观察到一些非常有趣的语言现象，却无法解释观察结果，而必须依赖于包括语言直觉在内的其他方法和资源来提供解释（Xiao 2009a）。尽管语料库方法存在这些问题，但由于其具备显而易见的优势，仍然越来越被语言研究者接受。其实，不同的工具具有不同的用途，关键是选对工具。比如，望远镜和显微镜都是十分有用的工具，我们不能指责显微镜无

法用来观察远处的东西，而望远镜无法用来观察细微的东西。同样，我们不能指望用语料库来研究它不擅长回答的研究问题，那些问题仍然需要用其他方法来研究（Hunston 2002）。因此，取得语料库研究成功的第一步，就是要根据语料库研究方法的特点，确定哪些研究问题可以用语料库来研究而哪些不能，并且学会如何将语料库方法和其他研究方法有机结合起来，融会贯通，充分利用各种资源，使语料库研究既具描述性，又具解释性。

由于语料库仅仅提供一种研究方法和资源，从事语料库研究时必须确定自己的研究主体。语料库方法可用来研究语言学和基于文本的人文社科领域中一系列的问题（McEnery, Xiao & Tono 2006；McEnery & Hardie 2012）。因此，针对特定的研究目的和研究问题创建或选用合适的语料库非常重要。

就语料库分析而言，基本的统计知识和量化分析技术十分重要，因为语料库研究中定量分析和定性分析同等重要，而要使量化分析具有一定的深度，就不能仅仅局限于比较频数和百分比等描写统计方法，而应该采用更复杂、更严格的推断统计方法，甚至是各种多变量分析方法。

熟练运用语料检索和量化分析工具在语料库研究中也很重要。要做到熟练，就必须勤学多练。现有的语料库分析工具（如AntConc、WordSmith、CQPweb等）功能都很强大，大多数语料库研究者已不再需要学习计算机编程。当然，如果你学习一门脚本语言（如Perl、Python），那就不仅会大大提高建库或语料分析的效率，而且还能进行一些常规软件无法进行的分析。当然，编程的学习曲线很陡峭，需要花一定的时间，但一旦学会，就会终身受益。

鉴于语料库语言学的研究本体是人们在真实语境中实际使用的语言[8]，从事语料库研究就首先要求研究者对语言使用具有敏感性。这种敏感性基于语言直觉，是通过长期使用语言和扩大知识面而积累起来的。因此，语料库研究的初学者应该避免急功近利、一蹴即就的心态，脚踏实地把基本功打扎实，以便获得语料库研究必备的学科素质。

注释

1. 参见兰卡斯特大学CASS语料库研究中心（http://cass.lancs.ac.uk）近年来在这方面取得的重大成就。

2. “英语用法调查”以卡片形式收集了1955-1985年30年间的语料，其口语部分后来转化为电子化的“伦敦-伦德语料库”（London-Lund Corpus）。

3. 参见英国研究理事会的数据政策（http://www.rcuk.ac.uk/research/datapolicy/）。

4. 例如，由上海交通大学出版社出版，王克非和胡开宝主编的《语料库翻译学文库》是目前世界上第一个、也是唯一一个语料库翻译学丛书系列，现已出版5

本高质量的专著（胡开宝2011、王克非2012、肖忠华2012、戴光荣2013、黄立波2014）。

5. 汉语语料库研究可见http://www.fass.lancs.ac.uk/projects/corpus/Chinese。

6. 语料库语言学MOOC见http://www.futurelearn.com/courses/corpus-linguistics。

7. 虽然语料库不能提供反面证据，但正如Stefanowitsch（2006）所述，完全有可能通过分析语料库来区分“显著缺失”和“偶然缺失”的语言现象。

8.“文本”在这里是广义的文本，包括口语和多媒体语料。

参考文献

Biber, D. 1988. *Variation across Speech and Writing* [M]. Cambridge: CUP.

Cao, Y. & R. Xiao. 2013. A multidimensional contrastive study of English abstracts by native and nonnative writers [J]. *Corpora* 8(2): 209-234.

Francis, G., S. Hunston & E. Manning. 1996. *Collins COBUILD Grammar Patterns 1: Verbs* [M]. London: HarperCollins.

Gries, S. 2006. Some proposals towards more rigorous corpus linguistics [J]. *Zeitschrift für Anglistik und Amerikanistik* 54(2): 191-202.

Hu, K. & K. Kim (eds.). Forthcoming. *Corpus-based Translation and Interpreting Studies in the Chinese Context* [C]. Basingstoke: Palgrave Macmillan.

Hu, X., R. Xiao & A. Hardie. Forthcoming. How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis [J]. *Corpus Linguistics and Linguistic Theory*.

Hunston, S. 2002. *Corpora in Applied Linguistics* [M]. Cambridge: CUP.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics* [M]. London: Longman.

Krishnamurthy, R. 2000a. Size matters: Creating dictionaries from the world's largest corpus [A]. In *Proceedings of KOTESOL 2000 – Casting the Net: Diversity in Language Learning* [C]. Taegu, South Korea. 169-180.

Krishnamurthy, R. 2000b. Collocation: From silly ass to lexical sets [A]. In C. Heffer, H. Sauntson & G. Fox (eds.). *Words in Context: A Tribute to John Sinclair on His Retirement* [C]. Birmingham: University of Birmingham. 31-47.

Léon, J. 2005. Claimed and unclaimed sources of corpus linguistics [J]. *Henry Sweet Society Bulletin* 44: 36-50.

Leech, G. 1991. The state of the art in corpus linguistics [A]. In K. Aijmer & B. Altenberg (eds.). *English Corpus Linguistics* [C]. London: Longman. 8-29.

McEnery, T. & A. Wilson. 1996. *Corpus Linguistics* [M]. Edinburgh: Edinburgh University Press.

McEnery, T. & A. Wilson. 2001. *Corpus Linguistics (2nd Edition)* [M]. Edinburgh: Edinburgh University Press.

McEnery, T., R. Xiao & Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource*