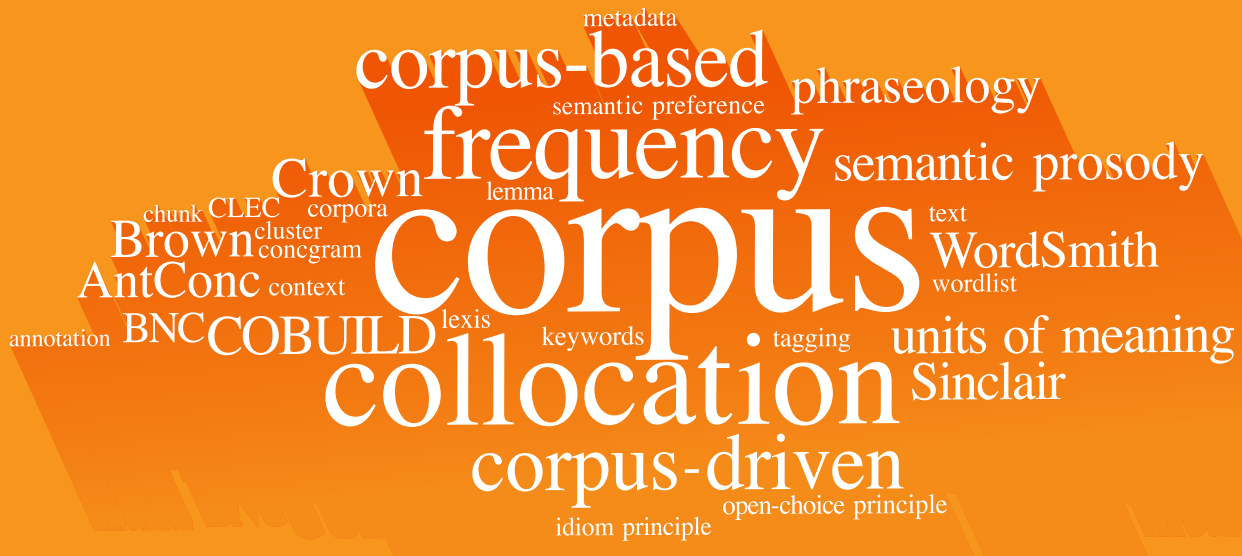
《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

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



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

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

主　编：梁茂成、许家金
编　校：徐秀玲、华　雨

Editors: Liang Maocheng and Xu Jiajin
Proofreaders: Xu Xiuling and Hua Yu

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

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，肖教授不幸因病去逝。

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*

Book [M]. London: Routledge.

McEnery, T. & A. Hardie. 2012. *Corpus Linguistics: Method, Theory, Practice* [M]. Cambridge: CUP.

Meyer, C. 2002. *English Corpus Linguistics: An Introduction* [M]. Cambridge: CUP.

Meyer, C. & G. Nelson. 2006. Data collection [A]. In B. Aarts & A. McMahon (eds.). *The Handbook of English Linguistics* [C]. Oxford: Blackwell. 93-113.

Newmeyer, F. 2003. Grammar is grammar and usage is usage [J]. *Language* 79(4): 682-707.

Seuren, P. 1998. *Western Linguistics: A Historical Introduction* [M]. Oxford: Blackwell.

Sinclair, J. 1997. Corpus evidence in language description [A]. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds.). *Teaching and Language Corpora* [C]. London: Longman. 27-39.

Sinclair, J. 1999. *Corpus, Concordance, Collocation* [M]. 上海：上海外语教育出版社。

Stefanowitsch, A. 2006. Negative evidence and the raw frequency fallacy [J]. *Corpus Linguistics and Linguistic Theory* 2(1): 61-77.

Tsou, B. & O. Kwong (eds.). 2015. *Linguistic Corpus and Corpus Linguistics in the Chinese Context* [C]. Hong Kong: The Chinese University Press.

Widdowson, H. 2000. The limitations of linguistics applied [J]. *Applied Linguistics* 21(1): 3-25.

Xiao, R. 2009a. Theory-driven corpus research [A]. In A. Lüdeling & M. Kyto (eds.). *Corpus Linguistics: An International Handbook (Volume 2)* [C]. Berlin: Mouton de Gruyter. 987-1007.

Xiao, R. 2009b. Multidimensional analysis and the study of world Englishes [J]. *World Englishes* 28(4): 421-450.

Xiao, R. (ed.). 2010a. *Using Corpora in Contrastive and Translation Studies* [C]. Newcastle: Cambridge Scholars Publishing.

Xiao, R. 2010b. How different is translated Chinese from native Chinese? [J]. *International Journal of Corpus Linguistics* 15(1): 5-35.

Xiao, R. 2011. Word clusters and reformulation markers in Chinese and English: Implications for translation universal hypotheses [J]. *Languages in Contrast* 11(2): 145-171.

Xiao, R. 2015. Source language interference in English-to-Chinese translation [A]. In J. Romero-Trillo (ed.). *Yearbook of Corpus Linguistics and Pragmatics* [C]. Berlin: Springer. 139-162.

Xiao, R. & G. Dai. 2014. Lexical and grammatical properties of translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective [J]. *Corpus Linguistics and Linguistics Theory* 10(1): 11-55.

Xiao, R. & T. McEnery. 2004. *Aspect in Mandarin Chinese: A Corpus-based Study* [M]. Amsterdam: John Benjamins.

Xiao, R. & T. McEnery. 2005. Two approaches to genre analysis: Three genres in modern American English [J]. *Journal of English Linguistics* 33(1): 62-82.

Xiao, R. & T. McEnery. 2006. Collocation, semantic prosody and near synonymy: A cross-linguistic perspective [J]. *Applied Linguistics* 27(1): 103-129.

Xiao, R. & T. McEnery. 2010. *Corpus-based Contrastive Studies of English and Chinese* [M]. London: Routledge.

Xiao, R. & N. Wei (eds.). 2014. *Translation and Contrastive Linguistic Studies at the Interface of English and Chinese* (Special Issue of *Corpus Linguistics and Linguistic Theory* Volume 10 Issue 1) [C]. Berlin: De Gruyter.

Xiao, R. & X. Hu. 2015. *Corpus-based Studies of Translational Chinese in English-Chinese Translation* [M]. Berlin: Springer.

Zou, B., M. Hoey & S. Smith (eds.). 2015. *Corpus Linguistics in Chinese Contexts* [C]. Basingstoke: Palgrave Macmillan.

戴光荣，2013，《译文源语透过效应研究》[M]。上海：上海交通大学出版社。

戴光荣、肖忠华，2011，译文中“源语透过效应”研究：基于语料库的英译汉被动句研究[J],《翻译季刊》(4)：85-108。

胡开宝，2011，《语料库翻译学概论》[M]。上海：上海交通大学出版社。

黄立波，2014，《基于语料库的翻译文体研究》[M]。上海：上海交通大学出版社。

王克非，2012，《语料库翻译学探索》[M]。上海：上海交通大学出版社。

肖忠华，2012，《英汉翻译中的汉语译文语料库研究》[M]。上海：上海交通大学出版社。

肖忠华、戴光荣，2010，寻求“第三语码”——基于汉语译文语料库的翻译共性研究[J],《外语教学与研究》(1)：53-61。

梁茂成谈语料库语言学与计算机技术

北京外国语大学 梁茂成

1. 您觉得哪些计算机技术与语料库语言学研究密切相关?

计算机技术与语料库语言学密切相关，表现在语料库语言学高度依赖计算机技术。没有计算机技术，语料库语言学就不可能发展起来，研究者就只能像Alexander Cruden（1699-1770）那样，或像蔡庭干（1861-1935）那样，靠人工将《圣经》和《道德经》制作成逐字索引，人们所能观察到的数据必然十分有限。早期靠人工制作词语索引的学者，其意识超前，不辞劳苦的精神值得称道，更重要的是这种做法反映了这些学者对经验主义认识论的推崇。

经验主义倾向是大多数人与生俱来的，也是语言研究中最不可动摇的方法。远的不说，20世纪上半叶，描写主义（descriptivism）语言学在美国就曾风靡一时，Franz Boas（1858-1942）等一批研究者深入到美洲印第安人中，对印第安诸语言进行了广泛的调查，采集了大量十分珍贵的语言数据。后来，以Charles Fries（1887-1967）为代表的语言学家传承了这种以田野调查为主要数据采集手段的方法，以经验主义为哲学基础的语言学研究得以大行其道，成了语言学研究中的主要方法。这种以观察数据为依据的方法主导着语言学研究，一直到Chomsky的普遍语法提出之后才有所收敛。在一些研究者（Fries 2010）看来，Fries的语言观和语言学研究方法与当今的语料库语言学别无二致，Geoffrey Sampson和Diana McCarthy更是从Fries所著的*The Structure of English*（1952）中节选了第三章，将其作为开篇之作收录到他们编著的*Corpus Linguistics: Readings in a Widening Discipline*（Sampson & McCarthy 2004）一书中。Leech（1992：105）甚至认为，美国学界50年代的结构主义语言学就是语料库语言学。由于当时这是唯一的语言学研究方法，人们自然不会采用“语料库语言学”这个名称。

我们认为，以上学者的观点的确可以表明，语言学研究中的经验主义方法由来已久，但并不能表明语料库语言学在前计算机时代就已经存在。语料库语言学不同于以往的田野调查方法，两者间至少存在以下差异：

1）前计算机时代语言学研究中的观察数据量远远小于当今语料库语言学研究中的数据。在基于田野调查的实证语言学研究中，研究者受到当时技术手段的制约，所能得到的数据量十分有限。比如，弗里斯在研究中使用的数据是约25万词

的电话录音。现如今，数据收集变得如此方便，以至于25万词的数据很难被称作为语料库。数据量的大小是前计算机时代语言学研究和当今语言学研究之间的重要区别。我们无意否认弗里斯的研究价值，但数据量的大小势必会对研究结果产生很大影响，凭借少量数据对语言现象进行概括很可能是以偏概全甚至是徒劳的，这一点在后来的语料库语言学研究中被反复证明。

或许在有些人看来，数据量的差异并无大碍，不会对研究结果有质的影响。然而，Sinclair（1991：100）的经典名句“若是同时观察很多语例的话，语言看上去会迥然不同”（The language looks rather different when you look at a lot of it at once）所强调的正是数据量的差异，当我们同时观察大量语言使用实例时，所得到的结果常常与我们预期的大不相同。通过计算机软件，语料库中多种语言现象一览无余，十分有利于研究者形成更为全面的概括，这自然有别于前计算机时代管中窥豹式的做法。很显然，Sinclair希望说明的道理是，数据量的差异是语料库语言学得以存在的基础，也是语料库语言学得以超越其他研究的前提所在。

2）前计算机时代的数据分析方法远远落后于当今语料库语言学研究中的数据分析方法。在没有计算机的年代，数据分析需要投入大量的人力，统计结果也常常难免出错。计算机可以对大量数据进行最为客观的分析，其效率之高是前人难以想象的。无论是Cruden的《圣经》逐词索引，还是蔡庭干的《道德经》逐字索引，都是花费巨大的人力、物力方才完成的，而且人工制作索引很难保证不出差错。为了提高数据分析的效率和准确性，早期语言学家曾作过不懈的努力。早在计算机刚刚问世不久，人们就尝试用计算机来分析语言。1963年，Lamb & Gould（1963）就出版了*Concordances from Computers*一书，书中系统介绍了他们设计的词语索引软件，该软件在IBM 709/90/94计算机上运行，可以极大地方便语言研究，并提高分析的准确性。再如，据Lavid（2007：10）、Renouf（1984：23）和Renouf（2007）记载，早在上世纪80年代初期，Sinclair就组织伯明翰大学的一班人马大张旗鼓地建设Birmingham Corpus，投入巨资购买了当时只有极少数研究机构才可能拥有的大型计算机（mainframe computers）（大约有公用电话亭那么大），并花费70,000英镑购买了当时最先进的扫描仪。每逢创建词表这样的大型工程，整个伯明翰大学不得不停电为其让道。在对120万词的语料库进行检索时，研究者甚至不得不将语料库分成6份分别处理，最后再把结果合并起来。Sinclair是最早使用计算机分析语料库的研究者之一，这一点反映了Sinclair对计算机重要性的认识，也正是因为有了计算机，Sinclair才得以建成COBUILD语料库，开创了语言研究的全新视野。

笔者认为，语料库语言学的发展基于大量的数据，而大量数据的分析又离不开计算机技术，因而计算机技术对语料库语言学发展的重要性是不言而喻的。概

括地说，与语料库语言学密切相关的计算机技术包括硬件技术和软件技术。硬件性能的提高和软件技术的发展，都是语料库语言学研究进一步深化的重要基础。硬件技术为语料库语言学的发展提供了可能，也使得我们对经验主义的追求在数据规模上得以超越前人，不再停留在费时、低效的田野调查水平上，这正是语料库语言学的区别性特征所在。而计算机软件技术，特别是以计算语言学研究为基础的自然语言处理技术，将为语料库语言学的发展提供强大的技术支撑，可以极大地丰富语言分析的维度和层面，远远胜过肉眼观察。未来的大数据分析技术更为语料库语言学的发展提供了无限的遐想。

诚然，我们应该牢记，计算机技术始终处于辅助和服务地位，语言研究才是我们真正的目的所在，在处理这一对关系时切不可本末倒置，一味地追求计算机技术而忽略对语言本体的关注。

2. 您如何看待语料库语言学与计算机技术之间的关系？

具体地讲，计算机技术在以下几个主要方面可以为语料库语言学提供服务：

1）语料的收集和整理。当今的计算机网络技术为文本的收集提供了极大的便利，人们设计了各种网络爬虫，可以快速从网络上采集到大量文本。WaC（Web as Corpus）技术的开发和应用更使我们能够对网络上采集来的文本加以定制。计算机扫描识别技术（即OCR技术）使我们能够把纸质版的各类书籍文档转成电子文本。语音识别技术的应用有利于大规模口语语料库的建设。从文本的整理看，利用文本整理软件，可以去除文本中各类噪音，从而保证文本加工的顺利完成。

2）语料的加工。计算语言学研究的不断深入使各类词性标注软件（part-of-speech tagger）、句法剖析软件（parser）等工具成为可能，而且这些软件的准确率不断提高，有效保证了语料库语言学研究的效率和信度。随着自然语言处理技术的发展，近年来人们甚至开发了语义标注（semantic annotation）软件、情感分析（sentiment analysis）工具等，极大地方便了语言研究。

3）语料库的分析。语料库分析技术不仅包括较为传统的索引行分析、词表分析和主题词分析，同时还有近些年来开展起来的多维度分析（multidimensional analysis）、多因素分析（multifactorial analysis）、聚类分析（cluster analysis）等。这些分析方法无不依赖计算机技术。索引行分析已经由原来的词语检索逐渐发展到框架（frame）检索、构式提取（如Stefan Gries的collostruction）、类联接分析（如许家金、熊文新 2009）等，词表分析也由单词列表扩展到多词列表，主题词分析已经扩展到主题词串分析和词性码串分析，多维度分析和多因素分析则是依赖对文本的深度加工、标注和复杂的统计技术，甚至融入了文本分类和机器学习

技术，这些都离不开计算机技术。近几年来，人们还将多维度方法用于网络文本的分析中，对计算机技术的需求越来越大。自动语义分析、情感分析更是在计算语言学最新研究成果的基础上发展起来的。

3. 您认为计算机技术在语料库发展过程中有过什么重要影响？

计算机技术在语料库语言学发展过程中起到了至关重要的作用。

首先，计算机技术催生了语料库语言学，使得语料库语言学得以从无到有。我们之所以认为前计算机时代的语言研究并非语料库语言学，是因为当时的田野调查数据不仅规模较小、取样不够科学，数据处理方法也相对原始。Sinclair等研究者借助计算机技术，创造了语料库语言学学科。在Sinclair（1991：1）看来，语料库语言学是一种崭新的语言观，这种语言观与（计算机）技术紧密相关（a new view of language and the technology associated with it）。计算机技术的介入，使我们可以同时观察到大量的语言事实，发现仅凭直觉无法预期的语言使用规律。因此，笔者认为，经验主义哲学是语料库语言学产生的哲学基础，而计算机技术则是语料库语言学产生的技术基础，两者缺一不可。语料库语言学是经验主义语言观与计算机技术结合的产物。没有计算机技术，语料库语言学就失去了可操作性。

计算机技术还是语料库语言学发展的推动力量。80年代之后，随着大规模集成电路这一硬件技术的突破，大型计算机很快被个人计算机取代，计算机迅速得到了普及，而且运算能力大大提高。到了90年代，计算机技术在语料采集、语料加工和语料分析中得到普遍应用，使语料库语言学学科得以快速前行。由此可见，计算机技术不仅促成了语料库语言学的产生，还极大地推动了语料库语言学的加速发展。纵观语料库语言学发展的简短历史不难发现，在计算机技术特别是自然语言处理技术得到快速发展后，语料库语言学得到了几乎同步的发展。笔者在Google Ngram Viewer中分别输入personal computers, corpus linguistics和world wide web，所得到的结果如图1所示：

从图1中可以直观地看到，个人计算机在70年代前后问世。语料库语言学几乎同时问世，并在短时间内得到了突飞猛进的发展。到了90年代，互联网开始逐渐进入大众生活，大量语言资源实现了网络化，更有力地推动了语料库语言学的发展。语料库语言学高度依赖计算机技术，而互联网的普及更使语料库语言学进入大数据时代，孕育着一系列新的变化。我们可以毫不夸张地说，计算机技术是语料库语言学得以产生的前提，也是语料库语言学得以发展的推动力。没有计算机技术，就没有语料库语言学。

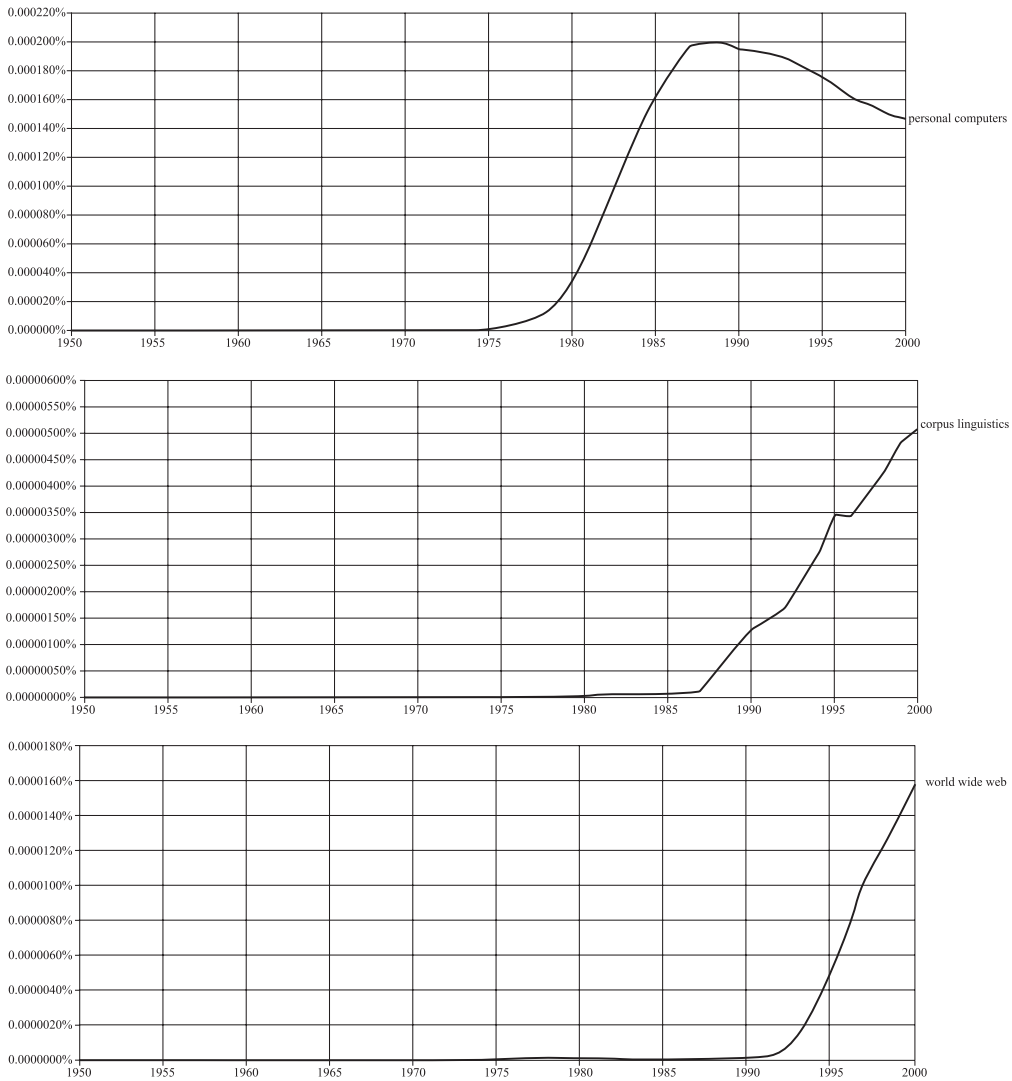


图1. 语料库语言学与计算机技术的同步变化

4. 请您结合自己的研究实践，谈谈计算机技术在语料库建设加工、语言学分析以及研究成果应用方面的作用。

在语料库建设与加工和语言学分析方面，所能涉及的最重要的计算机技术莫过于文本的标注。本人有幸参与了由文秋芳教授主持的“中国大学生英语口笔语语料库”（SWECCL）的建设。该项目于2002年开始，在语料库建设过程中我们曾得到桂诗春教授、何安平教授和英国伯明翰大学Susan Hunston教授的指导，并有机会向卫乃兴教授、李文中教授和濮建忠教授求教。

SWECCL包括口语和笔语两部分。在SWECCL建设初期，文秋芳教授领着我

们进入了一间仓库，里面存放着成千上万盒录音磁带，记录着历年全国英语专业考试的学生口语录音，我们的任务是要从这些磁带中抽样，并将抽样后得到的样本逐一转写成文本，配以电子化语音文件，建成语料库。我们采用分层抽样和系统抽样相结合的方法，得到了几千盒磁带。在大规模转写开始之前，我们先组织几位同学对几十盒磁带进行了试转写，以发现转写过程中可能出现的问题。之后，我们邀请以上几位教授就转写方案和标注方案进行了研讨。几位教授都有丰富的语料库建设经验，研讨过程中给出了十分有益的建议，使我们少走了很多弯路。其中，给我留下最深印象的是标注问题。根据我们原先的计划，准备先组织人力对所有的录音磁带进行转写和校对，然后对得到的所有文本进行细致的错误标注（error-tagging）。我们甚至设计了一套完整的标注方案和错误分类体系，开发了计算机程序以方便标注（也就是在这一过程中我开始学习计算机编程）。然而，在研讨会上，桂诗春教授结合自己创建“中国学生英语语料库”（CLEC）的经验，指出错误标注不仅是一个耗时费力的过程，而且不同的标注者对错误的认定很难取得一致。Hunston教授更是不赞成错误标注，认为我们应该保持文本的原样，其他几位教授也提出了相同或相似的观点。现在回顾起来看，幸亏当时我们征求了几位专家的意见，否则我们会陷入错误标注的泥潭之中。我现在的观点是，对于口语语料库建设而言，语料库建设者的任务是将口语转写成文本，同时需要把文本与语音文件对齐，以方便使用者检索。至于标注问题，特别是错误标注问题，应该留给研究者自己去完成。毕竟，由于研究目的不同，不同研究者对错误的认识和分类也会大相径庭，语料库建设者不可能设计出一个可以满足不同研究目的的标注方案。总之，对文本的标注要十分慎重，还需要充分考虑研究的目的。

尽管我们当时并没有进行大规模的人工标注，但就在对几十盒磁带的试标注过程中，我们学会了标注的基本方法。在后来的研究中，我常常需要对文本中的某些语言特征进行标注，虽然此类工作十分辛苦，但每当完成一定量的标注任务，总会有一种说不出的满足感。基于对标注过程的深刻理解，我在后来的研究中设计了两款标注工具。其中一款叫Text Annotation Tool（TAT），可以由研究者自行设计简单的标注方案或复杂的层级标注方案，极大地方便了标注过程。我一直认为，研究人员最了解自己的需求，设计的工具在适用性方面远远胜过由计算机专业人员设计的工具。

本人设计的另外一款标注工具叫KWIC-based Annotation Tool（KAT），即基于索引行的标注工具。研究者先对文本进行检索，找出自己感兴趣的词语或结构，然后加载自行设计的标注体系，直接在索引行中对节点词进行标注，这对语言研究者十分有用。比如，认知语言学认为，一词多义是词语的常态。在汉语中，介词“中”除了常见的空间域语义外，还具有丰富的隐喻意义，如“在语言学中”、“在孤独中”、“在他们中”，其中的“中”分别表示“领域”、“心理状态”、“范围”

等语义域。我们可以根据这些语义域，设计一个标注体系，然后在语料库中检索“在……中”，并按照不同语义场对这一构式进行统计和分析（见下图）。

图2. KWIC–based Annotation Tool

不仅如此，该软件还可以根据检索词语的语境相似性，对所有的索引行进行自动识别和标注，以方便语言研究。

基于自身的经验，笔者认为，在语料库技术的开发过程中，应该广泛征求研究者的需求，决不可闭门造车。

5. 您觉得目前计算机技术在应对英语、汉语和双语语料库建设和研究方面的重点和难点有哪些?

大规模英语语料库建设早在上世纪60年代就开始了，发展到今天，不仅规模上大大领先其他语种，而且其加工深度也为其他语言所不及，在语料库建设理念上也具有引领作用，相关技术的开发很快扩展到其他语言。从语料库的规模看，网络技术的利用使得数亿词级的语料库接连问世。此前几十年里，伯明翰大学与柯林斯出版公司合作建成的“英语文库”一直是无可争议的最大的英语语料库，但现在比“英语文库”更大的语料库并不少见，如Mark Davis主持建成的当代美国英语语料库（COCA）等一系列语料库都具有相当的规模。网络语料库（Web as Corpus）技术的开发更使语料库的规模以几何倍数增长。WaCKy和SpiderLing等工具的推出和WaC研讨会的召开很快普及了这种技术。基于此，Sketch Engine研究团队开发了十几个语种的大型语料库，并进行了加工和标注，发布到Sketch

Engine平台上，供教师、学生、研究者、翻译人员等使用。他们把这些语料库称为xxTenTen Corpora，其中的xx代表语言（如frTenTen Corpus是法语语料库，zhTenTen Corpus是汉语语料库），而TenTen指语料库规模达到10^{10}词级。基于以上现状，我们认为，语料库语言学已步入大数据时代，正孕育着一系列变化。

至此，语料库的规模已经不再是重点，更不是难点。有了大型语料库，接下来的问题当然就是语料库的加工和分析。

尽管Sinclair的干净文本原则被许多学者所推崇，但由于从生语料库中所能获取到的有价值信息十分有限，语料库的自动标注成为学界关注的重要问题，各国自然语言处理研究工作者尝试各种方法，努力提高词性标注的准确率，并在此基础上开发句法剖析系统、语义标注系统等，为语言教学、语言学研究甚至智能生活提供多种服务。然而，自从自然语言处理领域的主流方法由基于规则的方法过渡到基于统计的方法之后，标注的准确率虽然有了明显提高，但似乎已经达到了瓶颈阶段，很难取得更大的突破。由于词性标注是大规模语料库自动分析的重要基础，也是句法剖析和语义标注的前提，同时还与短语提取有着不可分割的关系，预计对不同语言进行词性标注这一项基础研究将成为计算机技术辅助语料库建设和语言研究中的一项重点工作，其目的自然是不断地提高标注的准确率。笔者一直认为，将基于规则的方法和基于统计的方法相结合，可以有效提高标注的准确率。同时，笔者十分认同李文中（2012）对标注的看法，对语料库仅进行有限标注，特别是要慎用汉语句法剖析等尚不十分成熟的技术。

双语语料库建设除了单语语料库建设中涉及的问题之外，还面临对齐（alignment）问题。虽然一些自动对齐工具已经取得了不错的效果，但对自动对齐结果的校对仍需投入大量人力。如何改进WaC技术，从互联网上自动获取双语文本，也将是双语语料库建设中的重点问题之一，计算机技术在其中责无旁贷。

在对大型语料库进行语言学分析方面，计算机技术的有效应用将面临重大难题。在笔者看来，如今的语料库虽然规模庞大，但大数据的基本特性之一——多样性，也在语料库中暴露无遗，这给语言学分析带来了极大的困难。如何按照语种、来源、语体等属性对庞杂的网络文本进行自动分类将是一大挑战。这一问题的解决需要自然语言处理领域的专家和语言学家的共同努力。除此之外，数据量大了，结果自然也就更复杂了。如何以概括形式将文本的各种特征科学地呈现出来，也将是我们面临的重大难题。或许，可视化技术在其中会起到重要作用。

6. 您最期待语料库分析技术在哪些方面有所突破？

作为一名语言研究者，笔者关注的自然还是如何对大型语料库进行有效的语

言学分析。比如，Patrick Hanks对大型语料库进行分析，从中归纳和提取了英语动词的主要型式，比如将Birmingham beat Coventry City.一句中动词beat的用法进行概括，抽象出诸如[[Human1 | Human Group1 = Competitor (Winner)]] *beat* [[Human2 | Human Group2 = Competitor (Loser)]]这样的型式（Hanks 2013：38）。如果这项工作中能够结合机器学习技术，可能会大大提高工作效率，也十分有利于语言研究的深入和辞书的编纂。

笔者对短语学理论深信不疑，但如何界定短语的边界，如何自动识别短语，如何对短语及其临近的各类词语进行范畴化（抽取各种类联接），以使得语料库语言学的研究结果超越词语层面而具有一定的范畴意义，这是笔者十分希望计算机技术能够解决的问题。很显然，这在更大程度上是一个语言学问题。只有语言学领域有了确定的标准和可操作的方案，计算机技术才能有所作为。

7. 您能给语料库研究初学者在计算机技术的学习方面提供一些建议吗？

在笔者看来，计算机技术十分重要，任何语料库语言学研究者都需要对计算机技术有一些基本了解。当然，这并不是说计算机技术对于语料库研究者来说是最重要的。恰恰相反，语料库研究者首先是语言学研究者，需要对语言问题具有高度的敏感性，否则就成了文本工匠。在处理文本时，我们注重的是文本的物理属性，而在分析文本时，我们关注的则是文本的意义和文本中的语言学现象。同时精通语言学和计算机技术是不太现实的，可能也是没有必要的。笔者反对工具至上，提倡语言学至上。

然而，建设一支语料库语言学团队则有所不同。笔者一直主张团队成员之间应该具有一定的共同性和互补性。共同性使得成员之间便于交流，但共同性太大了，就成了雷同。同样，互补性要求各成员各有所长，这样才会更容易产生新的思想。在语料库语言学团队中，人人都应该对计算机技术有所了解，但其中一部分人应该更加精通操作甚至能够编写计算机程序，团队中同时也应该有一些人更熟知语言理论、善于思辨，这样的合作才会更有意义。

8. 在大数据时代，语料库分析方法可能会发生哪些变化？

大数据时代的语料库分析方法可能会发生一些重要变化，我这里着重说三点。

首先，语料库的存储方式和检索方式会不同从前，这一点已经有所显现。此前，语料库是以光盘版或单机版形式存储的，可以拷贝，一般通过单机版软件工具检索。这样做的好处是我们可以对语料库进行多种个性化的、开放式的操作，但只适用于小型语料库。然而，语料库规模的扩大不仅占用大量的计算机磁盘空

间，单机版软件时常还会无法加载大型语料库，甚至会出现系统崩溃的情况。在大数据时代，语料库将存储于云端，只要有网络，语料库就无处不在，因此我们不再需要拷贝语料库，也不再需要单机版的软件。当然，如何对云端语料库进行个性化的再加工和检索将成为一个新的问题。

大数据时代的语料库分析方法也必将发生变化。比如，随着语料库规模的扩大，检索到的索引行可能会成千上万，仍靠人工解读很难完成，而对索引行进行抽样势必造成数据浪费和遗漏。在这种情形之下，我们或许可以对索引行进行自动聚类，并将分析结果以直观的图形方式呈现出来，点击图形中的特定区域，可以激活或调取相关联的文本或语境。近年兴起的数据科学（Data Science）将在大型语料库分析中起到至关重要的作用。

大数据分析善于发现相关关系（correlation），但并不揭示因果关系（causality）。比如，我们可能会发现某种语言特征与另外一些语言特征具有共现（co-occur）关系，还会发现某些语言特征与另外一些语言特征之间则存在共变（co-vary）关系。大数据分析并不能告诉我们为何会有这些共现关系和共变关系。就如同Rayson, Leech & Hodges（1997）发现男性话语中更多使用定冠词the一样，至于男性为什么比女性更多使用定冠词，这一点不太容易解释。在大数据时代，数据的解读具有很大的挑战性。

参考文献

Fries, P. 2010. Charles C. Fries, linguistics and corpus linguistics [J]. *ICAME Journal* 34: 89-119.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations* [M]. Cambridge, MA.: The MIT Press.

Lamb, S. & L. Gould. 1963. *Concordances from Computers* [M]. Berkeley, CA.: Mechanolinguistics Project, University of California.

Lavid, J. 2007. To the memory of John Sinclair, Professor of Modern English Language [J]. *Estudios Ingleses de la Universidad Complutense* 15: 9-12.

Leech, G. 1992. Corpora and theories of linguistic performance [A]. In Jan Svartvik, (ed.) *Directions in Corpus Linguistics*. [C]. Berlin: Mouton de Gruyter. 105-122.

Rayson, P., G. Leech & M. Hodges. 1997. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus [J]. *International Journal of Corpus Linguistics* 2(1): 133-152.

Renouf, A. 1984. A new specialized corpus: EFL materials [J]. *ICAME News* 8: 22-23.

Renouf, A. 2007. Corpus development 25 years on: From super-corpus to cyber-corpus [A]. In R. Facchinetti. (ed.). *Corpus Linguistics 25 Years On* [C]. Amsterdam: Rodopi. 27-49.

Sampson, G. & D. McCarthy. 2004. *Corpus Linguistics: Readings in a Widening Discipline* [C]. London: Continuum.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

蔡廷干，2014，《老解老》摘登 [J]，《语料库语言学》（2）：81-90

李文中，2012，语料库标记与标注：以中国英语语料库为例,《外语教学与研究》(3)：336-345。

许家金、熊文新，2009，基于学习者英语语料的类联接研究：概念、方法及例析[J],《外语电化教学》(3)：18-23。

通讯地址：100089　北京市北京外国语大学中国外语教育研究中心

邢富坤谈语料库语言学与计算机技术

解放军外国语学院 邢富坤

1. 您觉得哪些计算机技术与语料库语言学研究密切相关?

计算机技术可以分为3个层面来看待，分别是（1）工具;（2）方法;（3）思想。不同层面的计算机技术都与语料库语言学研究有联系。

1）工具层面。计算机是用于表示、存储和处理数据的工具。语料库语言学研究的基本素材是机器可读的语言数据，因此在工具层面，语料库语言学必然与计算机技术之间有着密切联系。首先是数据编码技术。语料库语言学的研究对象是语言文字，利用计算机对语言文字进行研究，首先要使语言文字在计算机内部能够得到统一表示，理想目标是人类使用的每一个语言符号在计算机内部都有一个唯一的编码与之对应，并且该编码在不同国家、不同语言、不同操作系统以至不同计算机之间都具有一致性，使得同一编码能够被不同计算机解释为同一语言符号。其次是存储技术。存储技术的发展表现为存储能力不断提升，存储成本不断降低，这为语料库的建设与管理提供了重要支持，尤其是大容量的存储设备，甚至是分布式存储技术的出现，为构建和应用超大规模语料库提供了基本前提。第三是数据检索技术。数据检索技术提供了发现与挖掘语言内在特征与模式的技术条件，研究者可以根据研究需求设定检索条件，利用计算机检索并统计真实语言中的语言实例，从而针对实例开展语言研究工作。第四是数据呈现技术。该技术是将语料库的数据分析结果以恰当的方式传递给研究者，使得研究者能够透过数据更全面地观察和分析语言特征与模式。语言数据呈现技术既包括以数据形式呈现语料基本特征，如型符数、类符数等，以索引行的形式呈现检索结果，也包括利用数据可视化技术，例如词云、词网、频谱图等形式对语言现象的分布以及词语等语言单元之间的关系等予以呈现。

2）方法层面。在方法层面，与语料库语言学研究相关的计算机技术主要表现为文本表征、序列标注、分类聚类等技术。文本表征有多种方法，主要包括字符串表征、词袋表征、词向量表征、语境向量表征等类型，其中字符串表征最为简单，表示能力也最弱，语境向量表征则较为复杂，表示能力也更强。文本表征方法的主要目的是服务语言计算，但也能够为语料库研究提供多种观察语言的视角，辅助研究者更好地归纳总结语言的规律特征。序列标注技术是为语言符号序列标

注属性信息的技术，词性标注是典型的序列标注，此外还包括命名实体标注、基本短语标注、句法结构标注、语义角色标注等，甚至中文分词目前采用的主流方法也是基于序列标注方法。序列标注能够为文本增加额外信息，这些信息并没有显性地表现在语言符号层面，借助序列标注将隐含的信息给予显性表达，能够更好地支持语言的研究工作。目前序列标注主要借鉴了机器学习的思想与方法，基于大规模的文本数据学习标注模型，实现较高准确率的标注效果。分类聚类技术以文本表征技术为基础，利用分类聚类方法实现对文本的分类或聚类，将特征相似的文本归并在一起，特征差异较大的文本分属不同的类别。分类聚类技术能够帮助研究者从语言特征的整体以及关联关系的视角观察语言，在不同类别下考察语言特征。

3）思想层面。形式化思想是计算机技术的核心思想。形式化思想要求以有限符号和确定步骤的形式将研究对象与过程给予表示，在给定输入的前提下，计算机能够经过确定的有限步骤处理，给出输出结果。形式化思想与语料库研究联系紧密，一方面是因为语料库研究的工具是计算机，计算机处理问题时必然要求研究问题与处理过程能够形式化；另一方面是因为形式化的思想对于充分利用计算方法，定量研究语言现象具有基础性作用。实际上，语料库本身就是形式化思想的一种体现，是利用了人类实际语言中具有代表性的一小部分样本代表了不可能完全获取到的人类语言的全部，这种以有限样本代替无限总体的抽样思想使得本来漫无边际的人类语言可以使用定量方法开展研究。在此基础上，当语料库达到一定规模后，某些层面的语言现象就会呈现出统计规律性，可以利用统计方法对语言现象进行深度挖掘，这是更具体的形式化思想。例如，当语料库达到一定规模后，词语的使用规律就会呈现出统计性特征，利用统计方法可以将一些特有的搭配和使用模式抽取出来，基于这些数据可以更加深入地研究语言的特征规律。当然在语料库基础上提出的N元统计语言模型、基于互信息的搭配获取方法、基于向量空间的语义计算方法等具体语言表示与计算方法，是形式化思想在语言研究与处理中更为具体的体现。

2. 您如何看待语料库语言学与计算机技术之间的关系?

语料库语言学与计算机技术之间的关系表现在3个方面：

1）语料库语言学为计算机技术提供了应用场景，计算机技术是语料库语言学研究的基本工具。技术要为应用服务，根据不同的应用需求和特点，会产生与发展相应的计算机技术。例如针对生物医学的需求，会研制开发出存储、表示和挖掘生物基因模式的计算机技术；针对金融服务的需求，会研制开发出预测证券市场波动变化的计算机技术。语料库语言学研究为计算机技术提供了一个新的应用场景，针对语言学研究的需求，研制开发针对语言分析的计算机技术。语料库语

言学研究的需求具体表现在大规模语料库的构建与管理、多样化语言特征的检索与统计、语言模式的识别与发现、语言特征的演化与比较、语言意义的形式化表示与计算等方面。这些特定需求依靠已有的通用计算机技术难以满足，必须针对语言特点，在语言学研究基础之上，开展相应计算机技术的研制与开发。

2）计算机技术为语料库语言学研究提供了新的动力。语言学研究需要动力，传统动力来源于人的需求，人对语言理解与使用的需求推动着语言学研究的发展。随着计算机的出现与普及，人类更多地依靠计算机处理语言，并利用自然语言与计算机进行交互。由于计算机的机械特质，其与人在语言学习与语言能力方面有着本质差别，因此计算机对语言研究提出了新的需求，主要表现在：计算机不仅需要简单且概括的语言规律和语法规则，同时需要更小颗粒度的语言知识与特征；计算机不仅需要典型的个案式语言分析，更需要在真实语言中具有广泛覆盖度的语言知识；计算机需要将只可“意会”的语言意义转变为可“言传”的具有形式化特征的意义形式等。以上需求对语料库研究提出了新的要求，需要在语言自身规律特点、语言形式与意义等方面开展深入的工作。

3）计算机技术为语料库语言学研究提供了新的检验评价途径。科学研究需要检验评价，在检验评价的基础上才能查找不足，不断前进。语言学研究同样需要检验评价，以往对语言学研究成果的检验评价大多依靠专家评判或是小规模的验证测试，可重复性与可比较性都难以得到保证。计算机技术以应用为目标，构建系统规范的评价体系，实现对研究结果的客观检验，从而不断改善和提高研究水平。例如，语音识别、信息检索、机器翻译等领域都拥有自己的评价体系，在统一评价机制下，这些领域都得到了快速发展（Palmer & Strassel 2007）。语料库语言学研究以真实语言为研究对象，研究结果需要接受真实语言的检验。计算机技术提供了一种新的检验途径，可以将语言研究成果应用到特定计算机技术之中，如语音识别、信息检索、机器翻译、文本分类等技术中，通过检验计算机技术的性能指标，达到对语言研究成果进行评价的目的。以文本分类为例，选择并确定文本分类的特征是语言学研究者需要回答的问题，一般计算机研究者会直接以字或词为特征单位进行文本分类，但字或词是否是最好的特征单元，是否还有一种能够更好地代表文本特征的语言单位，这些需要语言研究者进行研究。研究结果的有效性可以通过文本分类性能进行评价。以计算机技术应用为评价途径，能够更客观地评价语言研究成果，增强语言研究与语言应用之间的相互支撑。

3. 您认为计算机技术在语料库发展过程中有过什么重要影响？

计算机技术对语料库发展的重要影响主要体现在关键技术对于语料库建设与使用上的影响，概括为以下几项：

1）编码技术的影响。编码技术是将人类使用的符号转化为计算机内部的编码，从而使得计算机能够对语言符号进行存储与计算。编码技术的发展受制于计算机自身的编码表示能力。最初计算机的编码能力只有8位，也就是只能编制出256个不同的代码，这大大限制了计算机对语言符号的处理。随着计算机处理能力的提升，编码能力也不断提升，目前的主流计算机都具有32位编码能力，不少计算机已经可以有64位编码能力，编码能力的提升为计算机表示人类语言符号提供了基础保证。此外，编码还受到不同组织机构之间编码不统一的影响。随着统一码（Unicode）编码体系在国际上的普遍应用，编码也趋于一致，这就使得在不同平台、不同语言环境下可以一致性地存储与处理不同语言符号，为多语言语料库的建设与应用提供了重要保障。

2）索引技术（indexing）的影响。索引技术的发展对大规模语料库的高效检索与使用具有重要影响。索引的基本结构是词项与词项所在的位置。根据研究需求不同，可以将词项定义为字、词、短语等语言单位，也可以是作者、语体、年代等关于文本自身的信息。索引技术需要解决的问题是索引构建、索引更新、多层级索引、索引压缩等技术。高效率的索引具有占用空间小、索引结构优、更新速度快等特点，是语料库应用的基础条件。

3）互联网技术的影响。互联网使得语料库构建有了源头活水，电子文本难以获得不再成为构建语料库的瓶颈。由于互联网已成为人类信息交流的重要媒介，不仅传统媒介，如图书、报纸、期刊等媒体都将各自的信息内容通过互联网传播，同时还出现了一批网络特有媒体，例如论坛、邮件、博客、微博等，这些媒体每天都由普通民众生成和传递大量信息，信息内容多样，语言特色鲜明，为语料库的构建提供了重要素材来源。同时，互联网技术也为语料库应用平台的开发设计提供了新的渠道。很多语料库的使用都基于浏览器-服务器的模式开发，用户不需要在本地机器上存储语言数据，也不需要安装专门的语料库应用软件，就可以通过浏览器访问语料库所在的服务器，使用服务器提供的各类检索功能，不仅减轻了用户存储负担，也避免了诸如版权等问题的困扰。此外，利用互联网的搜索引擎进行语言检索也成为一种语料库研究的形式。

4）机器学习技术的影响。机器学习技术的基本思想是利用已有的经验数据，通过一定的学习算法，得到一个与经验数据拟合度较高且泛化能力较强的模型，利用该模型对未知数据进行计算分析（Pustejovsky & Stubbs 2012）。机器学习以经验数据为基础，这一点与语料库语言学如出一辙，不过机器学习更多的是利用计算方法对经验数据进行分析总结，形成可计算的模型，而语料库语言学更强调在机器辅助之下，对语言的内部规律进行深度研究。但无论如何，机器学习技术的出现，使得机器对于语言的处理能力变得更强，也为语言研究者提供了更多的观察与分析手段。目前，在机器学习框架下，语言模型得到了很大程度的优化，

从以往的N元语言模型发展到基于词向量表示的分布式语言模型。模型的优化最直接的体现是对语义计算的更好支持，在分布式语言模型的支持下，机器可以进行语义的代数计算，例如vector (“Madrid”) - vector (“Spain”) + vector (“France”)得到的结果是vector (“Paris”)，再例如vector (“King”) - vector (“Man”) + vector (“Woman”)得到的结果是vector（“Queen”），这种计算方法是以往其他语言模型难以比拟的（Mikolov *et al*. 2013）。基于机器学习技术得到的计算结果，结合语言自身的属性特征，语料库研究者可以更加深入地开展语言研究工作。

5）可视化技术的影响。可视化技术利用了人类对图形敏感的特点，以图形形式展现数据内容，从而辅助研究者把握数据的整体特点，更好地作出推断总结。可视化技术的发展对语料库建设与应用都有重要影响，主要表现在语料库检索结果、语料内部模式的直观展示以及研究者与应用系统之间的交互等方面。

4. 请您结合自己的研究实践，谈谈计算机技术在语料库建设加工、语言学分析以及研究成果应用方面的作用。

1）计算机技术在语料收集中的作用。语料收集有多种渠道，最直接的就是将已有的电子文本收录到语料库，间接的则需要文本的识别转换，例如通过OCR识别将图片类的文本转换成电子文本，或是在语音识别技术辅助下将语音素材转录为电子文本。随着网络的发展，从互联网直接获取网页文本成为构建语料库的重要素材来源，利用网络爬虫技术可以提高网络文本的获取效率，并实现网络文本内容的动态更新，为构建开放的动态语料库提供重要支持。

2）计算机技术在语料加工标注中的作用。语料加工标注主要包括语言素材的格式加工、文本元信息标注、语言属性标注等内容。在格式加工方面，主要利用了文本内容的解析技术，例如通过正则表达式对网页文本进行解析，提取其中关键部分内容，去除格式标记等内容，也可以通过分析文本的内部结构，构造相应的数据解析结构，实现对文本内容的解析与提取。文本元信息主要包括文本的来源、时间、作者、语体等关于文本自身的属性信息。语言属性则包括词语切分、词性、命名实体、基本短语等属性信息，语言属性信息的标注需要根据语料库的构建目的进行有选择性的标注。在属性标注方面，目前采用比较广泛的是XML格式语言，根据标注属性的需要，统一XML标注语言的格式，在统一格式框架下进行标注，在该框架下的标注结果具有很好的兼容性，能被不同应用程序解析和使用（邢富坤 2015）。

3）计算机技术在语言学分析中的作用。在语言学分析方面，计算机技术的首要任务是充当研究者的助手，辅助研究者开展语言研究。最典型的辅助功能是按照研究者的需求实现在大规模语料中对特定语言现象的查找，并对查找结果进行

直观展示。研究者基于查询结果，观察和分析语言现象，总结语言特征规律。其次，在计算机技术辅助下，能够实现对不同语料的对比分析，通过对比发现不同语料的独特语言特征，例如通过对不同语体语料中词语使用进行对比分析，查找出隶属于不同语体的特有词汇，形成具有某种领域特征的关键词汇，进而辅助相关研究与教学工作。第三，利用计算机技术，可以实现对历时语料的分析，给出语言与时间之间的关系，辅助语言演化研究。利用计算机技术，在一定程度上也能够主动发现语言中蕴含的某些模式特征，最典型的就是通过一些统计方法，例如互信息、t检验、卡方检验等方法，自动提取语言中的典型搭配，计算得到的结果对于提高词典编纂效率以及对搭配的深入研究都有重要意义（邢富坤 2012）。随着语料规模的不断扩大，机器学习方法在文本特征挖掘方面的应用越来越广泛，它能帮助研究者更全面深入地考察语言特征。

4）计算机技术在研究成果应用中的作用。从研究成果的应用角度看，计算机技术有两方面作用，一方面是通过构建人机交互的良好界面系统，实现语言研究成果的实用化，使得用户可以比较方便地获取和使用已有的研究成果。例如，基于语料库的词典编纂工作就可以利用计算机技术开发出的词典辅助编纂平台，平台集成了语料库研究中有关检索、搭配、聚类等研究成果，方便编纂者对词条进行检索、统计、排序以及聚类分析等工作，提高词典编纂效率。另一方面是将语言研究成果集成到已有的应用系统之中。例如，将搭配研究成果集成到信息检索系统之中，利用搭配信息优化查询与文档之间的相似度计算效果。由于信息检索本身具有一套较为成熟的评价体系和评测数据，因此可以通过评测实验，比较不同语言研究成果加入后的系统性能指标，从而判断语言研究成果的有效性。

5. 您觉得目前计算机技术在应对英语、汉语和双语语料库建设和研究方面的重点和难点有哪些?

1）语料库的规模、结构与加工问题。随着互联网发展，电子文本的获取越来越容易。但语料库建设的规模该如何确定，是否越大越好；如何定量地评价语料库效益，能否以某种评价函数的形式给出语料库效益的计算方法，指导语料库建设；如何确定语料库的内容结构，是否需要按确定比例对不同文类的语料数量进行规定；如何利用计算机技术辅助开展语料加工工作，在不同加工者之间建立协同机制，提高加工的一致性，这些问题无论对于单语还是多语语料库建设都是需要面对的问题（邢富坤 2013）。在应对这些问题时，需要考虑语料库应用目标、构建成本、计算能力等因素，综合给出应对办法。对于双语语料库建设而言，具有翻译关系的平行语料获取难度更大，在扩大语料规模的同时，更要有效地评价语料质量，有效地滤除翻译质量低劣甚至是机器翻译的文本进入语料库。对于构建双语或多语可比语料库而言，语料的获取难度要更低，但需要面对如何确定可

比关系以及如何有效分析利用可比语料的问题。解决以上问题不仅需要计算机技术，更需要计算机技术与语料库研究很好地融合，从语言自身规律与计算机能力两方面共同给出解决办法。

2）语言形式与语言意义的对应问题。在语言问题上，计算机面临的是符号形式与语义内涵不对等的难题。计算机技术擅长对数据进行匹配与计数，这些工作都是在符号层面进行。但由于语言中同一符号会对应多种语义，有些词的不同语义之间具有一定关联，例如“包裹、命题、发明”等，这些词的不同语义之间具有行为与行为结果的关联，而有些词的不同语义之间没有任何关联，例如“制服、分别、把手、马上”等。对于形式相同而语义不同的词语进行匹配与计数时，就出现了形式与意义的不对等问题，如果不顾语义而只求形式上的相同，则会造成匹配与计数结果与实际目标之间的偏差。此外，形式与语义的不对等问题也会对机器学习方法造成影响，不同语义的同一个形式在特征层面具有本质性差异，应该作为不同特征对待，如果将其混同为同一个特征，会给机器学习带来噪音，影响机器学习效果。因此，在利用计算机技术开展语料库研究的过程中需要重视形式与语义的关系问题，寻找可行的办法对该问题给予一定程度的解决。

3）语言属性与语言结构的研究与使用问题。语言属性是指在语言符号层没有显性表现，而是蕴含在语言内部、具有规律性的特征。语法类别属性（词性）是传统语言学研究中被广泛使用的语言属性，借助词性可以将具体的语言实例划分为不同的语言类别，从而使得语言研究结果具有一定的泛化能力，词性也可以辅助分析语句结构。当然语言的属性不局限于词性，语料库语言学视角下的搭配、语义倾向、语义韵等都可以作为语言的属性，如果让计算机使用此类属性，需要对此类属性有较为明确的定义，并针对属性的标注有规范的操作规格和流程。语言结构是在语言形式层之上的一种隐含的特殊语言属性，这类属性并不是在单独的语言单位上，而是语言单位之间的彼此关系。研究者对语言结构有不同的认识，有研究者将语言结构当作层级树状结构，也有研究者将语言结构当作线性结构，但无论何种结构都需要将其外化，并形成具有较大覆盖度的语言实例。语言属性与结构是语言研究的关键问题，计算机技术在面对这一问题时需要作两方面工作，一是在人工标注基础上，实现语言属性与结构的自动标注，为语言的深层挖掘与研究提供基本素材，这方面工作面临的困难是语言属性与结构的研究还有待深入，自动标注方法与标注模型的研究仍需不断提升。另一难题是基于已经标注语言属性与结构信息的语料进行有效检索与分析，从大量标注数据中发现规律性的语言使用模式，进而提高语言研究成果在语言教学、词典编纂、信息处理等领域的应用水平。

6. 您最期待语料库分析技术在哪些方面有所突破?

语料库分析技术的突破依赖于语言研究与计算机技术的融合与发展，其中以语言研究为突破关键。现有的语料库分析技术主要包括频数统计、词语索引、搭配、词丛、主题词等分析技术，分析对象主要是词，分析方法以频数统计和词语检索为主。语料库分析技术的目的是辅助研究者对语言进行观察分析，更好地总结归纳语言规律，支持相关语言应用。语言分析技术的辅助功能主要表现在两个方面：一是为语言研究者提供相关语言分析数据，使研究者能更全面地观察语言；二是为语言研究发现的相关规律提供检验，验证语言规律的有效性。

从提供语言分析数据的角度看，目前语料库分析技术主要集中在符号层，将语言作为一种数据符号，利用统计、检索等方法进行分析。期待下一步能够从符号层进入到语言属性层与语言结构层，能够支持语言属性、语言结构的统计与检索，在复杂多样的语言形式之上，找到具有更强概括性的语言模式特征，在不同语言形式之间建立起联系，更好地发现语言内含的规律性特征。实现该突破的核心是对语言属性、语言结构的研究，研究成果需要具有较强的形式化特征，且能够在较大规模的实际语言数据中得到实现与验证。

从检验语言规律的角度看，对语料库分析技术的更大期待是构建一套语言研究成果的检验评价机制与相关评价数据集。语料库语言学对于语言研究的重要贡献在于提出了一种从真实语言数据出发对语言进行研究的思想、方法和工具，并取得了一系列的语言发现。面向真实语言的研究发现应具有“可操作性”与“高覆盖性”（宋柔 2013），因此需要将已经取得的研究发现，放回到真实的语言数据上进行检验，检查相关语言发现在真实语言上的吻合程度，查找例外并加以完善。

构建评价机制与评价数据集的目的在于为不同的研究者提供统一的评价平台，从而对不同研究成果进行客观评价，减少不必要的争论，推动整个研究领域的滚动发展。实现这一突破的难度更大，因为语言研究绝大多数都是对语言规律的探索性研究，大部分成果是概念性、个案性的，且未最终定型，而构建评价数据集的前提是对研究问题已经有了较成熟的研究基础，形成了较完备的评价标准与评价实例。解决这一问题不能期待一步到位，也不能期待先构建一个完备的评价数据集，而是需要研究者在研究过程中边探索，边总结，边检验，边完善，需要多轮反复。在这个过程中，语料库分析技术需要承担的任务是管理已有的研究数据，将已有研究数据与最新研究数据进行对比分析，为研究者提供对比分析结果，更好地辅助研究者开展相关评价。

7. 您能给语料库研究初学者在计算机技术的学习方面提供一些建议吗?

计算机是语料库研究的辅助工具，工具的基本特征是技术门槛尽可能低，操

作使用尽可能便捷。对于语料库研究初学者而言，应尽可能降低技术对语言研究的影响与限制，将研究重心与精力放在语料库研究的基本方法与研究问题上，针对研究问题，使用已有的语料库分析软件有目的地开展研究工作。对于某些语言研究问题，可能现有分析软件难以满足研究需求，对于这类问题，首先是考虑是否有必要开发程序，如果手工能在可接受的时间内完成处理工作，则不必专门开发程序。如果处理数据量大，且以后需要重复进行类似工作，那么可以考虑专门开发程序加以实现。在程序开发之前，应对研究问题进行认真梳理，按照计算机处理的流程给出具体处理步骤，最好能够给出形式化的流程描述，为程序设计提供基本依据，一定要避免边写程序边设计。初学者如果有一定的数理基础，可以学习一门程序开发语言，但不是必需。通过学习和使用程序设计语言，可以提高形式化思维能力，培养形式化思维习惯，同时也能够通过程序自主实现一些特定的语料库分析功能，更好地辅助开展研究工作。在选择程序设计语言时，主要考虑的因素是自己身边是否有人在使用并能够教授这门语言，如果身边有一位对自己所学语言非常熟悉、经验丰富的使用者，那么可以大大提高语言的学习效率。

8. 在大数据时代，语料库分析方法可能会发生哪些变化?

大数据是相对于传统数据而言的。人类产生并可供使用的数据规模较过去有了很大幅度的提高。语言数据也是如此。语言数据规模扩大至少体现在两方面，一是语言数据的量大了，二是与语言数据相关的信息多了。语言数据量的扩大对于语料库分析方法影响不大，因为自语料库产生之初，就面对语言总体无限的难题，采用以有限语言样本代表无限总体的处理办法，到了大数据时代依然如此，不过有所改变的是语言素材的来源更广，话语形式更丰富，语言的动态特征也更强。语料库规模扩大，需要在存储、计算性能上有新的发展，借鉴分布式存储与计算的模式，有效地对大规模语言数据进行存储、管理与使用。

相对于语言数据规模的扩大，语言数据相关信息的增多对于语料库分析方法的影响更大。语言相关信息包括语言使用者的信息，例如微博中发表的内容都与博主关联；相关信息还包括话语的时间、位置等信息，尤其是随着移动智能设备和移动互联网的普及，在移动设备上产生的话语信息都带有了时间、位置等信息，这类信息与语言内容信息进行有效整合，对于更全面地把握语言特征具有重要价值。如果能够通过合理渠道获得语言相关信息，那么语料库分析方法也要随之发生变化。例如，语料库的索引就不再只是对语言符号进行索引，而是要加入语言相关信息索引，使用户在语料库检索时，不仅能够得到查询词的语言内容索引行，同时也能够根据语言相关信息对内容索引行给予更全面的描述，将言内与言外信息有效融合。

相对于变化而言，大数据时代也需要关注语言研究不变之处，加强对语言自身形式特点的研究，以语言研究的成果支持大数据的深度处理与分析。目前在大数据研究领域，针对语言数据的处理方法与声音、图像等数据的处理方法基本相同，语言学知识的使用非常有限，研究重点在于大规模数据的存储与计算方法上。但从语言的本质上看，语言具有不同于语音、图像等符号的特点，提高语言的处理效果，必须建立在语言自身规律全面深入研究的基础之上，就如同计算生物学的发展建立在生物学自身研究基础之上，计算本身无法替代研究对象的自身规律特点研究。在大数据时代，应该利用好语言数据与计算机工具，深入研究语言自身问题，例如语言的基本单元确定问题、语言的属性与结构问题、语言形式与意义之间的对应问题等。通过语言自身的规律特点研究，提高计算机分析与处理语言的能力，让计算机技术更好地服务语言研究与应用。

综上，本文认为计算机技术与语料库研究之间有着密切关系，彼此影响，互相促进，共同发展。计算机作为技术工具必然为语料库研究服务，其基本角色是研究助手，而不是研究的门槛或阻碍。计算机技术有严格的形式化要求，这也对语料库研究提出了挑战，语料库语言学的研究需要面向真实语言，研究成果要尽可能形式化，并且研究成果要接受真实语言的检验，通过检验评价来指导语料库语言学的发展。在大数据时代，语料库语言学有着新的发展机遇，需要新的变化，但同时也需要更加清醒地认识语言学自身的使命与任务，守住语言研究的主线，以语言自身研究的成果支持大数据时代的语言处理与应用。

参考文献

Mikolov, T., W. Yih & G. Zweig. 2013. Linguistic regularities in continuous space word representations [A]. In *Proceedings of NAACL-HLT* [C]. 746-751.

Palmer, M. & S. Strassel. 2007. Historical development and future directions in data resource development [OL]. http://www.itl.nist.gov/iaui/894.02/minds.html (accessed 12/20/2014).

Pustejovsky, J. & A. Stubbs. 2012. *Natural Language Annotation for Machine Learning* [M]. Beijing: The O'Reilly Press.

宋　柔，2013，汉语篇章广义话题结构的流水模型[J],《中国语文》(6)：483-493。

邢富坤，2012，多词单位的描写识别与词典编纂[J],《当代语言学》14 (4)：407-417。

邢富坤，2013，中文分词中未登录词分布规律及处理方法研究[J],《解放军外国语学院学报》36 (5)：27-32。

邢富坤，2015，面向语言处理的语料库标准：回顾与反思[J],《解放军外国语学院学报》38 (3)：8-13。

通讯地址：471003　河南省洛阳市解放军外国语学院语言工程系

Zipf定律及Zipf语言经济论剖析

洛阳师范学院 丁 政

提要：Zipf认为，因言者与听者的行为均受最小努力原则支配，双方立场上的省力是一对矛盾，基于言者经济的单一化力量与基于听者经济的多样化力量交锋，造就了语流中词语的一种规则的分布格局，即被后人称为Zipf定律的序号频率分布律（rank-frequency distribution）。然而，根据已达成共识的相关数学研究，Zipf定律另有原理。Zipf的语言经济论因此失去了一项其赖以成立的证据。不仅如此，语言学理性审视下，Zipf语言经济论实则远非无懈可击。立足语言学视角，本文欲解决两个问题：其一，以力求直观的方式对Zipf定律一探究竟，明确该定律并非Zipf所谓两种力量或言者听者双边经济矛盾的产物；其二，为Zipf语言经济论提出去伪存真的评价。

关键词：Zipf定律、最小努力原则、语言经济论

1. 引言

为现代语言研究打开新局面的趋势之一是具有人文传统的语言学与数学、统计学相结合。语言经济思想的奠基者之一，美国语言学家George Kingsley Zipf就是这种研究范式的先驱者之一。Zipf最为世界所熟知的研究成果莫过于Zipf定律与最小努力原则。Zipf定律是最早提出的计量语言定律之一，究其来龙去脉，在前人成果的基础上，Zipf以大量数据对该定律加以验证，将其纳入语言经济研究以及人类生态学范畴。最小努力原则是Zipf毕生治学历程的结晶，这一历程始于他对人类语言用词经济效应的钻研，以其对用词经济之心理本质的哲思为积淀，最终归于人类行为的基本规律。Zipf（1949）在著作《人类行为与最小努力原则：人类生态学引论》中提出最小努力原则。这部著作可谓Zipf毕生研究成果集大成之作，是人类生态学理论的重要组成部分。该书构建了以最小努力原则为纲的语言经济论，并且他先前在用词经济性研究中取得的重要成果均被重新诠释，Zipf定律就是其中之一，该定律原名为“词语序号频率分布（rank-frequency distribution of words；以下简称序频分布律）”（同上：25），本是Zipf用于论证其语言经济理论的一项实证依据。出于语言之混沌，自然文本词语的序号频率分布却呈现一种耐人寻味的规则格局，加之序频分布律在1949年著作的后半部分被Zipf推广应用于社区规模、城市人口等领域，Zipf的实证定律引起了科学界的浓厚兴趣与广泛关注，并被后人称为Zipf定律、Zipf分布。

2. Zipf语言经济理论之精髓

2.1 最小努力原则与语言经济机制假说

在《人类行为与最小努力原则：人类生态学引论》一书开篇，Zipf提出了最小努力原则，指出人类行为普遍受这一基本原则支配。最小努力之实质包括两个层面：其一，最小努力是“最小工作量的一种变体”（Zipf 1949：1）；其二，最小努力是一种平均量。也就是说，人的行为不可能时时处处都做到最小努力，但行为路径之多个步骤、解决系列问题之多个过程的平均工作量趋于最低。此外，以斧锯刨凿与木匠活、战时转产军车的民用汽车厂等例子作比，Zipf（同上：8）为人类行为打了一个比方，或者说建构了一个用以表述人类行为的模型，即工具与任务关系：在最小努力原则支配下，工具与任务之间相互选择、相互依存；一方面，工具适应任务才能有效降低工作量；另一方面，执行何种任务能够达成最小努力又取决于已经掌握的工具。

提出最小努力原则后，Zipf（同上：19）明示：他对人类行为生态的研究首先关注人的语言行为；对语言行为的探讨始于将人类言语视为一系列工具的组合，具体而言，将词汇看作一系列工具的组合。基于这个工具任务关系，Zipf（同上：20-21）提出如下概述的假说。其一，语言中存在一种经济机制、一种潜势，以或多或少尽量俭省的方式将词语与意义结合起来。其二，从言者立场看，使用一个工具完成所有的任务最经济，也就是仅使用一个词语就能表达所有的意义；在听者立场上，因为要在特定语境中解读言者使用的词语，所以最省力的方式是每个词语仅表达一个意思。其三，Zipf构想了“单一化（unification）”与“多样化（diversification）”两种力量，分别基于言者立场与听者立场，单一化力量趋于将词汇的数量减少至一词，而多样化力量则趋于将每词表达的意义缩减到一个，两种力量的交锋决定了语流中存在多少词语以及词语承载多少意义。

2.2 序频分布律与意义分布律

对于其语言经济论中的假说，Zipf进行了如下概述的论证：假设存在两种矛盾经济、两种相对的力量，或许会造就某种“词汇平衡（vocabulary balance）”。“词汇平衡”为何物？ Zipf（同上：22）如是说：“我们显然尚不知道在假想的两种力量之间是否真的存在这样一个状态”。可见所谓“词汇平衡”，除字面意思外并无其他内涵。在这种典型的Zipf式论述逻辑中，他提出了“词汇平衡的实证依据”（同上：24），即被后人称为Zipf定律的序频分布律。基于小说《尤利西斯》的词频表，Zipf指出：将构成文本的词语按频率降序排列，词语在词表中的序号r与其相对频率f的乘积约等于一个常数C，即$r \times f = C$。在双对数（log-log）坐标系中，序号频率关系图像呈45度角下倾直线形状，其低频部分呈阶梯状（同上：

24），简图示例如图1：

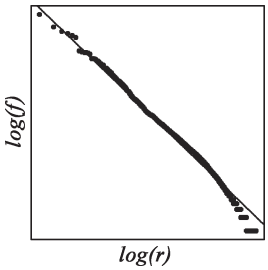


图1. Brown语料库序频分布图像

Zipf将45度角下倾直线形状的序频分布看作所谓词汇平衡以及两种力量之说的证据。至于该证据是否充分，Zipf的解释相当单薄：序频分布图像的作图方式是将数据点标于图中，绘图前没有关于图像是否规则的预期，若以线段将数据点连接起来，这些线段的角度随机，但绘图后发现这些随机线段的倾斜方向如此规则，所以词频分布势必受到某种原则的支配（同上：27）。后有学者就序频率分布图像的倾角补充了如下解释：

“如果一种语言只有一个单词，它的出现率会是100%。相反，如果每个单词都只有一个意义，那么，一个语篇的不同单词数会跟总词数一样，而且各个单词的出现次数都会是1。如果用坐标表示，前者是一条竖线，后者是一条横线。把它们合在一起，正好构成一个90度直角。现在的45度斜线，恰恰是前两种情况的中和、妥协。既然前两种情况分别只考虑了说话人利益，或听话人利益，那么，中和前两种情况的第三种情况就既考虑了说话人利益，又考虑了听话人利益，就是‘单一化力量’和‘多样化力量’之间平衡、妥协的结果”（Poosala 1997，转引自姜望琪 2005：90）。

假想的一纵一横两直线何以至于在客观数据之对数坐标系图像中构成45度角下倾直线？以近乎栩栩如生的意象思维臆测语言问题的道理，有违最基本的科学严谨，以上这段解释非但不能支持Zipf的理论，反而更令人生疑。从两种力量到45度角斜线，这个由前后相扣的假说、不甚严密的论证构成的理论体系实则捉襟见肘，下文“Zipf语言经济论献疑”一节将在理性思辨的基础上批判其中破绽；至于Zipf定律规则的图像奥秘何在，下文“Zipf定律原理剖析”一节将有透彻探讨。

Zipf对两种力量之说的论证并未止于序号频率分布，序频分布律还有一个称为“意义分布律（law of meaning distribution）”的姊妹篇：既然词频分布规则，那么意义分布很可能也是规则的；两种力量的交锋之下，不存在能够表达一切意义的词语，但势必有一些词语能够表达多种意义；设文本中最高频词语的频次为F_1，该词语表达意义的数量为m_1，则$m_1 \times f_1 = F_1$，其中f_1是以该词语表达各意义的平均频率。因单一化与多样化力量的平衡，m_1与f_1趋于相等，由此，$m_r = \sqrt{F_r}$，下标r指的是词语在词频降序排列中的序号；在横轴为r、纵轴为m_r的双对数坐标系中，预期得到一条斜率为-0.5的直线（Zipf 1949：28）。对于这个假设，Zipf的实证检验采用Thorndike（1932）汇编的《教师20,000词词书》作为资料，该词书以500为单位分组，并按照词频降序排列。Zipf记录了Thorndike词表中每组词的词典义项数量，按组求得词均意义数量并绘制图像。据Zipf对绘图结果的总结，得到了

一组线性排列的数据点，图像线性斜率接近-0.5（Zipf 1949：30）。

Zipf的意义分布律与统计检验缺乏说服力。其一，Zipf调查的是*r*~*m*关系，意义分布律公式陈述的是*F*~*m*关系，两者显然不能等同。其二，即便不深究前述这个破绽，必须充分论证*r*~*m*关系的线性斜率确实趋近-0.5才能为两种力量之说提供令人信服的证据。所以，基于一个不甚严密的模型，用一次性的统计调查验证一个玄虚有加的假说，Zipf的意义分布定律假说经不起推敲。不过，Zipf的实证调查并非乏善可陈，推究意义分布律的前因后果，可溯及Zipf（1945：251）曾提出的一个称为“频率意义律（meaning frequency law）”的用词经济法则：频用词语趋于具有更丰富的多义性。不难看出，意义分布律是频率意义律在以最小努力原则为纲的语言经济论之下的再诠释。Zipf的统计调查不足以支持意义分布律，却是频率意义律与用词经济效应的确凿证据。

限于本文主旨，对Zipf著作的详解止于观研Zipf语言经济论的精髓部分以及Zipf独特的论述模型：与其他人类行为一致，语言行为中的经济性无外乎最小努力；人类言语行为被解读为工具（词语）与任务（意义）的关系；语言经济之最小努力机制是言者与听者的经济对立、单一化与多样化两种力量的交锋与平衡；序频分布律、意义分布律既是这种机制的现象，也是其证据。然而，正如上文已经提出的若干评论，Zipf的语言经济论由双边经济论、两种力量论等假说构成，而Zipf对其假说的论证远非无懈可击。因此，Zipf语言经济论遗留了一些值得商榷的问题：如何评价Zipf的假说？ Zipf的实证定律能否作为这些假说的确凿支持？

3. Zipf定律原理剖析

3.1 “于语言而论，Zipf定律甚浅”

为解答Zipf定律是否即所谓词汇平衡、是否是两种力量之确凿证据的问题，不免涉及关于Zipf定律的数学研究。统计语言学家Herdan（1966：33）指出：“语言学家认为Zipf发现了一个数学定律，而数学家认为Zipf发现了一个语言学法则”。经由此话，Herdan表达了一种主张：在Zipf定律这个问题上，语言学与数学两个视角应互通有无，否则语言学家有可能对Zipf定律的数学本质不求甚解，而数学家为Zipf定律建立的解析模型未必符合语言实情。

Zipf本人关于Zipf定律原理的看法，即最小努力原则造就了Zipf定律，并没有得到广泛认可。数学家Mandelbrot，心理学家、语言学者Miller早已通过数学研究为该定律道破玄机，Miller贡献的启示更加斐然。Mandelbrot率先提出了Zipf定律的数学论证以及数学推广，即Zipf-Mandelbrot定律。Mandelbrot的研究显然受到最小努力原则的启发，使用了一个抽象晦涩的“单位信息最小平均成本”模型

（转引自Miller 1957：313）。Miller（同上：311）沿用前者的数学模型，为其破除玄虚，提出Mandelbrot模型的一种直观解释，即“猴文本（monkey text）”。“猴文本”可解读为：猴子任意敲击打字机键盘，产生随机字母组合，随机出现的空格将字母组合切分为形式词符，形式词符堆砌产生形式文本。基于这一模型，Miller成功完成数学推导并得出结论：“简单无奇的数学过程便能够产生Zipf定律，无需为最小努力、最小成本之类原则建模”（同上：313）。猴文本研究后Miller曾多次发表旨在为Zipf定律去伪存真的看法，在为Zipf著作之再版撰写的导读中如此评价Zipf的研究：

“面对Zipf定律的数学秩序性，选择无非有二：或以人类心理的某种共同属性解读定律，或视其为某种概率规则的必然结果。Zipf的选择是制造假说，试图以最小努力原则去解释用词行为中在似是而非的单一化与多样化之间的某种平衡。其他人则大多寻求基于概率原理的解释。三十多年的研究后，这个问题已经明朗，后者才是正确的。视消息源为一个随机过程，Zipf曲线所描述的无非是该过程的必然结果”（Miller 1965：vi）。

Miller & Chomsky（1963：463）也曾指出：“Zipf定律并不能说明存在某种能够塑造人类语言交流行为的普遍心理作用”。Mandelbrot（1982：346）后来也承认最小成本的概念并无必要，并提出一句言简意赅的评论：“于语言而论，Zipf定律甚浅”。该如何理解Mandelbrot的这句话？人的身高是典型的正态分布总体，但不能从身高的生物学基础为正态分布钟形曲线找原因。同理，Zipf定律确系语言之特征，但并非语言之特有，至于Zipf曲线何以呈现规则形态，并非一个语言学问题，而是一个数学问题。

3.2 Zipf定律成因之直观模型

本文无意深入Zipf定律的数理分析，仅希望通过一种较直观的方式阐明“于语言而论，Zipf定律甚浅”的道理，不妨从猴文本说起。猴文本是一个用于数学论证的模型，与自然文本有很大的区别，如下是一个模拟猴文本并分析其词语序频分布的R语言程序以及该程序输出的图像实例：

```
#程序1
alph="_ABCDEFGHI"
#代码清晰起见，以下划线表示空格
#以连字符表示字母间无空格
alph=unlist(strsplit(alph,split=NULL))
txt=paste(sample(alph, 50000, replace = TRUE),collapse="-")
ws=unlist(strsplit(txt,"_"))
```

```
ft=table(ws[which(ws!="-")])
ft=data.frame(ft)
f=sort(as.numeric(ft[[2]]),TRUE)
r=seq(1:length(f))
plot(log(r),log(f))
```

程序1的原理与步骤正如本文上节对猴文本模型的解释，凭经验斟酌，设置了一个短小的字母表，生成尽可能接近自然文本的猴文本序频分布。尽管有所干预，程序1生成的序频分布（如图2）与自然文本的Zipf分布（如图1）相去甚远，但也显现出整体格局的相似，这说明形式上的自然文本与猴文本本质相通，所以不妨以后者为启发，构思一个更接近自然语言的模型。

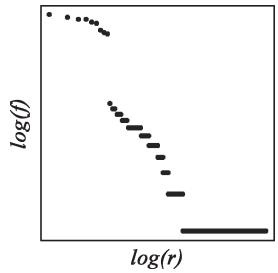


图2. 猴文本序频分布

猴文本的形式词符是随机字母组合，短者概率高，长者概率低，合乎自然语言词符参差不齐的使用概率；猴文本可看作一个概率随机模型，自然语言由词符构成语句的形式过程理论上也是如此。所谓概率随机，直观起见，以投掷骰子作比。将一枚骰子看作一个消息源，其符号表含6个词语，概率均为1/6，连续掷骰子的过程就是以此系统产生文本的随机过程。由自然语言词符构造语句的过程与连续掷骰子相似，具有很高的随机性。而自然语言词符的概率分布参差不齐，封闭类词语的使用概率普遍高于开放类词语，后者之列也有较常用、较不常用的概率之别。基于上述原理，设计了如下R语言程序，输出模拟序频分布图像：

```
#程序2
nw=50000;st=50000
#词表长度与形式文本长度
mean=1;stdv=exp(1)
#对数正态分布参数（凭经验斟酌）
ws=as.character(seq(1:nw))
rn=rlnorm(nw,mean,stdv)
p=rn/sum(rn)
txt=sample(ws,st,TRUE,p)
f=sort(table(txt),TRUE)
f=as.numeric(f)
r=seq(1:length(f))
```

plot(log(r),log(f))

程序2细节如下：（1）假设以数量为50,000的词符集作为消息源，生成长度为50,000的随机文本；（2）对于消息源的概率分布，采纳Herdan（1960：42）的理论："词频分布总体受对数正态分布律支配"，假设消息源服从对数正态分布。凭经验斟酌其参数，选择1为均值，以自然常数*e*为标准差；（3）从这个消息源连续随机采样50,000次，相当于在一个概率随机过程中生成如此长度的随机文本，对随机文本进行频率分析后绘制序频关系的log-log图像。以下是程序2运行结果的一个实例：

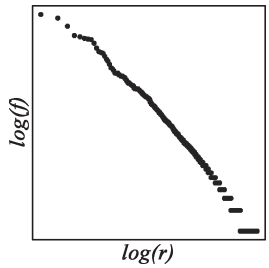


图3. 模拟自然语言形式文本序频分布

由图3可见，程序2模拟自然文本Zipf分布的效果显然优于程序1，在目测之下与自然文本Zipf曲线（如图1）相差无几，足以印证Miller所谓"简单无奇的数学过程就能够产生Zipf分布"。程序2的原理无外乎从对数正态分布总体随机采样，与猴文本模型一样简单无奇，且更符合自然语言形式上的语句构造过程。

3.3 关于Zipf定律系数的研究

根据Joos、Mandelbrot等学者的研究（详见冯志伟 1983），Zipf序频分布律原公式缺乏普遍性，有失精确。实际应用中的Zipf定律公式是Zipf原公式的修正与推广，可写作$r^{B}f=C$，Zipf原公式是当B=1时修正公式的特例。修正公式中B、C是两个系数，并非两个常数。因为目测下形似斜线，Zipf分布被看作log-log坐标系中的线性关系，由修正公式可推得：$log(f)=log(C)-B(log(r))$。该线性关系的截距是$log(C)$，其斜率是$-B$。不过，Zipf分布实质上比双对数线性关系更复杂，尚无明确的线性拟合截距取值标准，C取值不定，而B的测定一般采用最大似然估计（MLE）方法。以下程序中的B.MLE是基于MLE且专用于计算Zipf定律系数B的R语言函数：

```
#程序3
B.MLE<-function(f)
#f是自然数词频数据的矢量
{library(stats4)
LLH=function(B){
r=seq(1:length(f))
L=B*log(r)+log(sum(r^-B))
sum(f*L)}
fit=mle(LLH,start=list(B=-1))
B=as.numeric(fit@coef)
```

```
return(B)}
f = scan("data.txt")
#数据文件格式是空格分隔的自然数词频
B.MLE(f)
```

采用MLE方法对多个英文语料库的Zipf定律系数实施精确测定，结果见下表：

表1. 11个英文语料库的Zipf定律系数*B*及相关数据

语料库	类符	形符	TTR	*B*
BNC法律	35,190	2,245,003	0.016	1.049
BAWE	106,566	6,667,353	0.016	1.037
BNC口语	22,429	917,333	0.024	1.049
BNC笔语	47,879	1,006,395	0.048	0.997
BROWN	49,805	1,022,553	0.049	1.006
CLOB	50,072	1,014,668	0.049	1.000
CROWN	49,103	1,019,036	0.048	0.993
FLOB	51,729	1,017,410	0.051	1.004
FROWN	52,162	1,021,014	0.051	0.995
商务英语(自编)	10,300	175,372	0.059	0.977
读者文摘(自编)	28,023	603,176	0.046	0.999

基于11个英文语料库测得的一系列*B*数值的均值为1.01，极差为0.072。这一调查结果说明英文文本的Zipf定律系数的确近似于1，或者说用45度角下倾直线拟合英文语料库Zipf分布图像的误差不大。

一些学者曾指出，Zipf定律系数*B*因语言而异小幅变化，与词汇丰富度（lexical richness）或类形比有关（Gelbukh & Sidorov 2001）。所谓类形比，是类符数与形符数的比率（type-token ratio，TTR），是词汇丰富度的一个简单指标。基于表1所列数据，实施相关性分析，结果如下：

表2. 基于表1数据的相关分析

	类符	形符	TTR	*B*
类符	1			
形符	0.847	1		
TTR	-0.365	-0.708	1	
B	0.237	0.539	-0.952	1

由相关分析结果可见，B与TTR间存在高度相关（r=-0.952，p<0.01），Zipf定律系数受TTR影响。既然如此，Zipf定律图像的线性倾角服从某个常数角度的趋势不存在。确凿证据再次表明序频分布并非Zipf所谓两种力量造就的词汇平衡。

综合本节中的实证研究，既然可以通过简单数学模型生成Zipf分布，其奥秘无外乎Miller（1965：vi）所谓“视语言的消息源为一个随机过程，Zipf曲线无非是该过程的必然结果”，并非最小努力的产物；既然Zipf分布的线性斜率与文本的类形比高度相关，其直线形状与线性倾角与所谓词汇平衡无关。基于上述结论足以断定Zipf定律与Zipf语言经济论的关系：后者不能解释前者，且前者并非后者的有效证据。既然如此，Zipf的语言经济论因缺乏证据而成为无本之木，产生了另一个有待探讨的问题：该如何评价Zipf的语言经济论。

4. Zipf语言经济论献疑

4.1 两种力量论之无端

Zipf的语言经济理论由统辖于最小努力原则之下的一系列假说构成，而Zipf对这些假说的论证又不乏破绽。那么，Zipf的理论是否中肯？是否能够揭示语言经济的奥秘？缜密的理性审视之下，Zipf的理论其实存在诸多破绽，其中最为致命的就是两种力量论。造就所谓词汇平衡的两种力量是否存在？按Zipf的假说，言者希望用尽可能少的工具完成尽可能多的任务，也就是希望词语的多义性强，在这个方向上存在趋于将多种意义集于一词的单一化力量；听者希望解读词汇的工作量尽可能小，也就是希望词语的多义性弱，在这个方向上存在趋向于一词仅表达一义的多样化力量。不过，这种理论捉襟见肘，与语言的规约性相悖。语言是一个规约体系，词语能表达何种、多少意义约定俗成，言者与听者势必遵循同一套规约，而自行裁定词语之意义、能指之所指的自由度很小，所以根本不可能大刀阔斧地实现两种力量。既然两种力量不可能产生实质效应，何谈平衡，何至于左右词频分布格局？由此可见，两种力量之论不仅缺乏证据，且因违背了语言的基本属性而站不住脚。

既然涉及语言的规约性，不妨从这个角度浅谈语言经济性的道理。规约性可谓语言经济性的基本要件。在这一点上语言与其他人造的或自然的通信体系一致。若言者与听者不能构成一个以共同知识为规约的通信体系，那么两者之间或根本无法沟通，或必须为传达消息付出许多额外工作。此外，规约机制为降低语言交流的开销提供了一种必要的可能性，即最少仅用一个符号、一项语言手段就可以传达一个意义实体，以至于一方面语言系统势必动用大量符号为动态变化的、潜在数量无限的主客观实体赋予称谓，而另一方面，因为人脑有处理语境制约的能力，一词多义也广泛存在，两方面均体现了语言之最小努力原则。可见，语言因

规约性而省力，且为了省力，语言必然是一个规约体系。

4.2 双边经济之玄虚

两种力量无端，与两种力量相纠结的双边经济也并非顺理成章。Zipf认为言者、听者经济立场上的单边最小努力以两种力量的方式交锋，却忽视了一个不言而喻的道理：只有交流成功才有省力可言，否则必然导致额外开销，所以交流成功是达成最小努力的前提。诚然，单边最小努力的矛盾是否存在是一个真伪难断的问题，但强调这种玄虚的矛盾没有实际意义，即使有某种对立，也只能是行为方式的对立：以实现成功交流为纲，言者付出表述的工作，听者付出解读的工作，达成最小努力的方式既相反且相成。对Zipf语言经济思想有所沿革的两种著名理论，即Martinet的经济原则与Horn的R、Q原则，均明示或隐含了这个道理。

在提出R、Q原则的论文中，Horn（1984：11）开篇即引用了Zipf两种力量之说，随后又援引了Martinet的经济原则："为了理解语言如何变化、为何变化，语言学者须谨记两个永恒相悖的因素：其一，成功交流的要求、言者对成功传达消息的要求；其二，以达成交际目标为前提，将言者生理、心理能量付出降至最低的最小努力原则"（转引自Horn 1984：11）。

对Zipf与Martinet的理论，Horn如此概括："正如Zipf所谓'两种对立的经济极端矛盾'，Zipf、Martinet等认为语言的变化正是在这个矛盾的熔炉中炼成的"（Horn 1984：11）。而后，Horn指出："这两种矛盾的力量及其交互作用就是Grice会话原则以及由其所衍生的语用推理机制之主要根源"（同上：12）。Horn将Zipf的双边经济之说与Martinet的经济原则混为一谈，这分明是一个误区。Martinet将最小努力原则纳入了经济原则，也强调这是一个二元矛盾，但成功交流与最小努力的矛盾和Zipf所谓两种经济的交锋显然有本质区别，且经济原则的精髓所在就是强调成功交流是最小努力之前提。

尽管Horn将Zipf与Martinet的原则无差别概括为"两种矛盾的力量及其交互作用"，他在此基础上建构的理论并没有破绽。如下是Horn归纳的Q、R原则框架（同上：13）：

1）Q原则（基于听者）：充分话语量；说得尽可能多（以R为前提）；下限原则，诱发上限会话含义。

2）R原则（基于言者）：必要话语量；只说必须说的（以Q为前提）；上限原则，诱发下限会话含义。

强调R、Q互为前提就是对成功交流是最小努力之前提这一原则的尊重。所以尽管Horn沿用了Zipf的双边经济，却没有延续其玄虚。此外，"'基于言者'与'基于听者'相对，'下限'与'上限'相对……，这是一个工整的对称"（Carston

2005：305），R、Q两原则体现了言者与听者的方式相对与经济统一。Horn（2006：2）后来更新了他的理论，用“语用之阴阳互动”来描述其理论中的两种宏观对立的原则。既相反且相成，“阴阳互动”的提法显然比片面强调交锋更加合理。

5. 小结

Zipf是一位风格独特的语言学者，他视语言为一个生物、心理、社会过程，以统计、数学分析加哲学探讨的方式研究语言，进而将从语言研究中得出的哲理加以推广。Zipf的理论与计量研究中之所以不乏破绽，一方面是因为在Zipf的时代以数论理的技术条件简陋。另一方面，Zipf其人有一些值得计量、实证语言研究领域学者力戒的弱点：Zipf的数学能力有限、统计学知识不足，对数据之性质不求甚解却常在臆测理论的方向上渐行渐远（Wyllys 1981：47）。在以去伪存真的态度审视Zipf语言经济理论的同时，有两点深刻的感受：瑕不掩瑜，尽管其理论不无瑕疵，Zipf开创的理论视野博大，时至今日尚未被充分探索且仍大有可为；以Zipf为先驱的研究范式，即计量调查加推究哲理，在今天的语言研究中仍有方兴未艾之势与广阔的前景。

参考文献

Carston, R. 2005. Relevance Theory, Grice and the neo-Griceans [J]. *Intercultural Pragmatics* 2(3): 303-319.

Gelbukh, A. & G. Sidorov. 2001. Zipf and Heaps Laws coefficients depend on Language [A]. In A. Gelbukh (ed.). *Proceedings of Confenence on Intelligent Text Processing and Computational Linguistics* [C]. Berlin: Springer-Verlag. 332-335.

Herdan, G. 1960. *Type-Token Mathematics* [M]. The Hague: Mouton.

Herdan, G. 1966. *The Advanced Theory of Language as Choice and Chance* [M]. Berlin: Springer.

Horn, L. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature [A]. In D. Schiffrin (ed.). *Meaning, Form, and Use in Context: Linguistic Applications* [C]. Washington: Georgetown University Press. 11-42.

Horn, L. 2006. Speaker and hearer in Neo-Gricean pragmatics [J]. *Journal of Foreign Languages* (4): 2-26.

Mandelbrot, B. 1953. An informational theory of the statistical structure of languages [A]. In W. Jackson (ed.). *Communication Theory* [C]. Woburn, MA.: Butterworth. 486-502.

Mandelbrot, B. 1982. *The Fractal Geometry of Nature* [M]. San Francisco: Freeman.

Miller, G. 1957. Some effects of intermittent silence [J]. *American Journal of Psychology* 70(2): 311-314.

Miller, G. & N. Chomsky. 1963. Finitary models of language users [A]. In R. Luce, R. Bush & E. Galanter (eds.). *Handbook of Mathematical Psychology* [C]. New York: John Wiley. 419-492.

Miller, G. 1965. Introduction [A]. In G. Zipf (ed.). *The Psycho-biology of Language: An Introduction to Dynamic Philology* [C]. MA.: The MIT Press.

Thorndike, E. 1932. *A Teacher's Word Book of the Twenty Thousand Words* [M]. Teachers College, Columbia University.

Wyllys, R. 1981. Empirical and theoretical bases of Zipf's Law [J]. *Library Trends* 30(1): 53-64.

Zipf, G. 1945. The meaning-frequency relationship of words [J]. *The Journal of General Psychology* 33(2): 251-256.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* [M]. Cambridge, MA.: Addison-Wesley.

冯志伟，1983，齐普夫定律的来龙去脉 [J],《情报科学》(2)：37-41。

姜望琪，2005，Zipf与省力原则 [J],《同济大学学报》(2)：87-95。

通讯地址：471022　河南省洛阳市洛阳师范学院外国语学院

汉语时间词“年”、“月”、“天”的搭配行为研究*

华南师范大学 方清明

提要：本文利用语料库方法考察时间词“年、月、天”的搭配行为。它们与整数、半数、约数、位数的搭配颇为复杂，需要条分缕析。传统研究仅以若干句法标准考察特定高频词语是否具有同质性的做法，有待进一步探讨和检验。“年”的搭配力远远高于“月、天”，“年”至少有9个义项是“月、天”所没有的，其超高频率和用法的多样性充分说明“年”具有突出的例外性和特殊性。“月”的特殊之处在于无标记序数用法与需要前加量词“个”才能表达基数用法。“有一天、第二天”等含有“天”的多词单位具有语用语篇功能。本文认为多个相关的高频词之间，往往只具有部分的同质性，很难有完全的同质性可言。

关键词：时间词、年、月、天、日、语料库、搭配

1. 引言

汉语时间词“年、月、日”已经得到学界比较深入的研究。《现代汉语语法讲话》（丁声树等 1961）指出“年、天”为准量词，“月”为普通名词。陆俭明（1987）首次较为系统地对比了三者的差异，认为“年、日”是时量词，“月”则是名词。

陆丙甫、屈正林（2005）则持不同意见，他们关注“年、月、日”的同类性，认为作为序量词的“年、日”和“月”之间并没有本质的差别。以往的语法分析把“月”看作是“年、月、日”这个连续统中的例外；而根据他们的描写，在“世纪”到“秒”这个连续统中，例外的其实是“年”。陆和屈还进一步从认知角度解释了“大单位容易兼有名词形式，小单位容易兼有基量词形式”的倾向性。

邓思颖（2012）对陆丙甫、屈正林（2005）的观点提出了质疑，如：1）为什么“年”是例外？如果“年”是例外，为什么仍然说“年、月、日”之间“并没有本质的差别”？ 2）邓认为“毫秒”和粤语的“字（五分钟）”有名词的用法，

* 本文是国家社科基金青年项目“基于语料库与AntConc3.2.4w技术的汉语抽象名词搭配研究”（14CYY033）的阶段性成果，并得到教育部人文社科青年基金项目（13YJC740018）、暨南大学华文教育研究院2013创新平台研究重大项目（CXPTZD201315）和华南师范大学（2012）青年教师科研培育基金项目的资助。文责自负。

因此否认了“单位大小与词类之间存有必然的认知关系”的说法。

邓思颖明确支持陆俭明（1987）的观点，并进一步论证汉语时间词的不对称性，如例（1-4）。邓先生认为“年、周、天、分钟、秒”应该分析为量词（基量词），而“月、星期、钟头”则属于名词。量词和名词的差异是语法上的差异，跟认知无关。另外，张霄军（2010）、裴雨来（2011）等也有相关论述。

（1）年年、*月月、天天（重叠测试）

（2）一年年、*一月月、一天天（重叠测试）

（3）一年一年、*一月一月、一天一天（重叠测试）

（4）多少年、*多少月、多少天（疑问代词测试）

综上，上述成果无疑加深了人们对“年、月、天/日”性质的认识；但是客观来说，前贤研究多囿于若干句法标准来探讨时间词的同质性问题，其局限性是显而易见的。第一，上述研究都尚未考虑频率问题，但事实上，时间词的高频搭配对语法规律的呈现有着重要影响，非常值得关注。第二，尚未从搭配角度全面揭示“年、月、天/日”用法的丰富性。第三，尚未对含“年、月、天”多词单位的语用和语篇功能进行考察。

“众所周知，运用语料库对语言现象进行定量定性的实证研究，已经成为近年来语言学发展的一个重要趋势。汉语学界真正把大规模语料库和软件技术相结合进行相关研究的案例并不多见”（方清明 2014）。鉴于学界目前缺乏对时间词搭配现象的系统研究，本文试图跳出句法词类视角，利用“国家语委现代汉语标注语料库”（以下简称“国家语委语料库”）和AntConc3.2.4w[1]技术相结合的方法对时间词“年、月、天”[2]进行分析。

2. 数量与“年、月、天”的搭配

数量与“年、月、天”组配时，呈现出非常复杂的关系。下面从整数、半数、位数、约数四个方面进行描写。

2.1 整数与“年、月、天”的搭配

2.1.1 基数、序数与“年、月、天”的搭配

整数可分为基数和序数两大类。其中：

基数、序数与“年”搭配有三种表达方式，如例（5-7）。

（5）一年、两年、三年、100年（基数用法）

（6）第一年、第二年、第100年（有标记“第”的序数用法）

（7）1949年、2000年、2014年（纪年时，为序数用法）

基数、序数与“月”搭配有两种表达方式，如例（8-10）。

（8）一个月、两个月、三个月（基数用法）

（9）第一个月、第二个月、第三个月（有标记“第”的序数用法）

（10）一月、二月、三月、十二月（无标记序数表达，范围限于“1-12”）

基数、序数与“天”搭配有两种表达方式，如例（11-12）。

（11）一天、两天、三天（基数用法）

（12）第一天、第二天、第三天（有标记“第”的序数用法）

以上用例至少能说明以下三点：第一，进行基数表达时，“年、天”与“月”不同，“月”需要前加量词“个”。第二，进行序数表达时，“年、天”与“月”也不同，“年、天”需要加标记“第”，而“月”可加可不加，无标记用法范围限于前加数词“1-12”。第三，“年”表示纪年时，“年”也是序数表达。

2.1.2　整数越大，越倾向于与“年”搭配

“年”与整数搭配具有较高的适配性。从理论上说，从一到无穷大的整数都可以与“年”搭配，如1年、2年、100年、10,000年、1亿年。

“月”与整数搭配则较为受限，一般以“1-12”为常规，13以上就有条件，因为“13-23个月”可以说成“一年多”，“25-35个月”可以说成“两年多”。理论上来说，只要超出12个月，都可以用“年”来折算。“天”与数量搭配也有此类限制，如“32-59天”可以说成“一个多月”，“400天”可以说成“一年多”。从折算的优选性来看，“年”是最具优势的时间单位。其原因大致是：第一，人类思维、自然现象和社会现象，其本质是模糊的，不是精确的。从计量时间方面来说，需要模糊计量的情况往往比较多。例如可以说“我在广州生活了八年”，一般不会说“我在广州生活了96个月”或者“我在广州生活了2,920天”。第二，作为时间方面的大单位和基本单位，当数量越大，折合为“年”的可能性也就越大，因为这是最便捷的。这就像人们到银行取一万元钱，人们得到100张100元的可能性是最大的，除特别要求外，银行一般不会给1,000张10元的。

2.2　半数与“年、月、天”的搭配

“半”与“年、月、天”的搭配，丁声树等（1961：177-178）、陆俭明（1987）已经有所论及。不过我们发现，随着数量的增大，“年、月、天”与“半”搭配的可接受度越来越低，如例（13-16）。

（13）半年、一年半、两年半……?十一年半、?十二年半、*一百年半[3]

（14）半个月、一个半月、两个半月……?十一个半月、?十二个半月、*一百个半月

（15）半天、一个半天、两个半天……?十个半天、*一百个半天

（16）半天、一天半、两天半……?十一天半、?十二天半、*一百天半

上述用例表明，随着数量的增大，“半”的使用越来越受限，如“一百年半、一百个月半、一百天半、一百个半天”在现实语言里为超低频现象（方清明 2012a），可接受度非常低。这也与时间表达的模糊性有关，例如大数量加“年”多为模糊量，而“半年”却是精确量，因此二者搭配时较难被接受。另外，前贤多指出“一天半、两天半、三天半”这类用法，但是现实语言里还有“一个半天、两个半天、三个半天”这样的表达，例如“我这个学期上课比较轻松，每周只要上两个半天就可以了”，这个例子里的“半天”整体上具有名词性。

2.3 位数与“年、月、天”的搭配

朱德熙（1982：45）把“一、二、三……九”称之为系数词，把“百、千、万、亿”称之为位数词，“十”既可以是系数词也可以是位数词，而系位构造是由系数词和位数词两部分组成的复合数词。“年、月、天”与系位构造搭配时，所受到的限制不尽相同。例如：

（17）十年、十个月、十天　　（17’）十年、十个月、十天

（18）百年、*百个月、百日　　（18’）一百年、?一百个月、一百天

（19）千年、*千个月、千日　　（19’）一千年、?一千个月、?一千天

（20）万年、*万个月、??万日　　（20’）一万年、?一万个月、?一万天

（21）亿年、*亿个月、???亿天　　（21’）一亿年、?一亿个月、?一亿天

例（17-21）为位数词与“年、月、天/日”的搭配，例（17’-21’）为系位构造与“年、月、天/日”的搭配。上述用例说明，“年”可以与位数词直接搭配，也可以与系位构造搭配，其搭配力最强；“天/日”搭配力居中；“月”不能与位数词直接搭配。统计发现，“一千日、一万日、一亿天”、“一亿个月”等都没有用例出现，可见这类搭配可接受度非常低。即系位构造数量越大，也越不适合与“月、天”搭配。

2.4 约数与“年、月、天”的搭配

2.4.1 约数“多”与“年、月、天”的搭配

丁声树等（1961：177-178）详细地指出，“比三年、三天多点儿，‘多’字加在‘年’或‘天’的后面。多于三个月，‘多’字加在‘月’前面”。我们观察语料发现，情况远比这复杂，请看例（22-23）。“十”是个分水岭，当数字小于十，“多”位于“年、天”之后，当数字大于十，特别是出现“百、千、万、亿”时，

“多”位于“年、天”之前。当数字正好是“十”时，“多”与“年、天”搭配可前可后，但意思有所不同，如“十多年”是大于十年小于二十年的意思，而“十年多”是大于十年小于十一年的意思。

（22）三年多、十年多、*二十年多、*三十年多、*一百年多

（22’）*三多年、十多年、二十多年、三十多年、一百多年

（23）一天多、十天多、*二十天多、*三十天多、*一百天多

（23’）*一多天、十多天、二十多天、三十多天、一百多天

另外，我们通过检索“国家语委语料库”发现“多年”二字单位为1,576例，而“年多”二字单位为315例。“多天”二字单位为110例，而“天多”二字单位为13例，这从整体上说明，“多”位于“年、天”之前的频率更高。“多日”已经固化成词。

“一百多年”能说，而“一百年多”不能说，这是为什么？可能的原因是，第一，从语义上来看，“一百多年”是指101年到199年这样的约量时间；而“一百年多”里的“多”的语义无法落实，如果“多”是“1-12个月”的意思，那么这个相对的小量与表示大量的“一百年”极不匹配。如果是“1-99年”的意思，则与“一百多”语义重复。因此“一百年多”在语义上不具有合法性。第二，从韵律来看，“一百多年”属于“2+2”自然音步模式，而“一百年多”则是“2+1+1”音步模式，这类模式造成韵律不和谐，因此很难被接受。

2.4.2　其他约数与“年、月、天”的搭配

“X年来”是非常高频的多词单位[4]，共计出现1,464例。“X年来”的实例里有“一年来、两年来”等具体表达，但是出现较多的是“十多年来、多年来”等约量时间表达，这类表达方式以“年”的搭配能力最强，请看例（23-26）。

（23）十多年来、多年来、近年来、几年来、近百年来

（24）*十多个月来、*多月来、*近月来、*几月来、*近百月来

（25）十多天来、*多天来、*近天来、几天来、*近百天来

（26）十多日来、多日来、近日来、几日来、*近百日来

胡玲（2007）指出“X以来”以表示长时居多，在一个很长的时间段中，过于精确的起点显得没有必要。“X以来”具有方向性，由过去时间指向说话人所在的时间，这种方向性具有一种动态效果。拿“几年”与“几年来”相比，“几年”是静态的，是时间流程上任意截取的一个时间段；“几年来”是时间流程上相对于定的一个时间段，起点是几年前，终点是说话时，有一个动态的过程。据统计，虽然“十多日来、多日来、近日来、几日来”有用例出现，但是频率都不高，倒是另外一个多词单位“连日来”比较高频。

3. 从频率和搭配看“年、月、天”的共性与个性

“频率”可分为实例频率（token frequency）和类型频率（type frequency）（Bauer 2001；王洪君、富丽 2005；Thompson 2007），一般来说，实例频率要远远高于类型频率。

3.1 “年、月、天”在类型频率搭配方面的共性

三者都表时间，其左搭配以数量成分最为高频，这是“年、月、天”之间的共性所在。下面罗列三者都能搭配的类型以展示它们的共性，如例（27）所示。

（27）次年、次月、次日 | 当年、当月、当天 | 几年、几月、几天 | 每年、每月、每天 | 哪年、哪月、哪天 | 那年、那月、那天 | 年报、月报、日报 | 年产、月产、日产 | 年历、月历、日历 | 年薪、月薪、日薪 | 全年、全月、全天 | 同年、同月、同日 | 元年、元月、元日 | 这年、这月、这日 | 整年、整月、整天 | 逐年、逐月、逐日 | 年收入、月收入、日收入

如果有人把“年、月、天”三者看作同类词的话，那么这种语言直觉应该与例（27）里三者在类型搭配方面所体现出来的共性密切相关。需要指出的是，尽管例（27）类型频率一样，但是实例频率依然有所差异，如“每年<1,113>、每月<237>、每天<6,289>”，人们日常生活中使用“每天”的概率要高很多。

3.2 关于“年”搭配的个性分析

陆丙甫、屈正林（2005）认为在从“世纪”到“秒”这个连续统中，例外的是“年”；不再按传统观点把“月”看作“年、月、日”这个连续统中的例外。邓思颖（2012）以“年、天”与“月”的一系列句法不对称性否定了“年”的例外性。本文与上述分析视角都有所不同，我们认为从频率和搭配角度来看，“年”无疑具有例外性或者特殊性。

“年”的频率远远高于“月”和“天”。依据“国家语委语料库”的统计，它们的实例频率分别为“年<22,029>”、“月<6,336>”、“天<6,401>”。“年”的超高频暗示该词含有“月”和“天”所不具备的额外功能。“年”的高频组合是揭示其用法的重要线索。

第一，“年”具有纪年作用，表达某个年份，如“2014年”、“2020年”等。这一点，前文已述。

第二，“年”可以表示一生中按年龄划分的不同阶段，如“早年、童年、幼年、少年、青少年、青年、壮年、中青年、中年、中老年、老年、晚年”等。

第三，“年”可以表达年龄或岁数的大小，如“年少、年幼、年轻、年青、年老、年迈、年纪大、年龄小”等。

第四，“年”可以对人生某个十年进行称谓，如“而立之年、不惑之年、耄耋之年”等。

第五，“年”与春节、年货有关，如“过大年、过年、过新年、年初一、年除夕、年糕、年关、年画、年货、年节、年三十、年味、年夜饭、年意、新年”等。

第六，“年”与十二生肖有关，如“鼠年、牛年……狗年、猪年”。其中特别值得注意的是，“猴年马月”已经固化，它属于“似非而是”的表达。按常理应该表达为“猴年马年”，但是由于避免同音，加之“X年X月”的构式压制作用，人们最终采用了“猴年马月”这样的表达。如果看其内部构成成分的话，“马月”似乎并不合法，它不是对“月”的某种划分或命名。“形态词”是指构词成分本身不是词（Packard 2001：11-12；方清明 2015a: 39）。“马月”的性质类似于形态词。

第七，作为时间单位，“年、月”本身可以作为词素对“年”或者“月”进行内部划分，并且方式多种多样，如例（28）。“天”往往较为单一，似乎只有“早晨、上午、中午、下午、黎明、傍晚”等划分，但这些都不含有“天”这个词素。

（28）年头、月头、*天头 | 年初、月初、*天初 | 年中、月中、*天中
年底、月底、*天底 | 年末、月末、*天末 | 年尾、月尾、*天尾
年终、月终、*天终

第八，用于与“学校”有关的语域，表示“年级”，如“一年级、二年级、低年级、高年级、学年”等。

第九，“年”所反映的其他事情，如“编年、丰收年、更年期、荒年、纪年、年成、年代、年会、年检、年鉴、年景、年轮、年限、年岁、年华”等等。以上关于“年”的用法可以形成表1：

表1. 关于“年”的9个义项与用法

项目	年	月	日
1. 与纪年有关	+	-	-
2. 与划分年龄的阶段有关	+	-	-
3. 与年龄、岁数大小有关	+	-	-
4. 称谓人生某个十年	+	-	-
5. 与春节、年货有关	+	-	-
6. 与十二生肖有关	+	-	-
7. 内部划分方式多种多样	+	+	-
8. 与学校的年级有关	+	-	-
9.“年”所反映的事情	+	-	-

表1里，“年”的9个义项基本都是“月、天”所不具备的。“年”具有义项丰富、搭配能力强大的个性，因此完全可以认为“年”具有例外性和特殊性。

3.3 “年度”词汇化的个案分析

如前文所述，“年”参与“构造已有词语的组合能力”（王洪君、富丽 2005）非常强大。不仅如此，以“年”为成分的单位还具有固化成词的潜能，下面以“年度”为例进行探讨。

吕叔湘（1999：604）指出“一度”表示一次或一阵，常和“一年”连用，组成“一年一度”修饰名词，作定语。我们检索“国家语委语料库”，发现共有“年度”218例，其中绝大部分用例被标注为量词，但这并不符合语言事实。从词性上来看，“年度”应属于名词，主要充当定语，如“年度人物、年度计划、年度建设、年度预算”等，也可以充当中心语受其他成分修饰，如“本年度、上年度、下年度、该年度、按年度、各年度”等。

更值得注意的是，“年度”是在当代固化成词的。检索“北京大学现代汉语语料库”发现，新中国成立前的语料基本不见该词，新中国成立到改革开放期间有零星使用，在改革开放以后，在如当代报刊、网络语料、人民日报等语料中却出现了19,748例。高频效应导致“年度”越来越固化，最后完成词汇化。

“年度”来自“一年一度、两年一度”等表达的缩略。缩略的动因主要有以下几个方面：第一，“一度”具有黏着性质，“‘一度’修饰名词作定语在现代汉语中是很受限制的，一般只能出现在与前面的时段词语对举的句子中”（于立昌、吴福祥 2011）。第二，界面差异，“一年一度”多用于句法层面的描述，如“一年一度的体检又开始了”，而“年度体检”则属于粘合结构或概念化的名名复合词层面。第三，韵律促动，如“2013年一年一度杰出人物颁奖典礼现在开始”里划线部分显得较为繁杂，不经济，因此有“2013年年度”的潜在表达，但是由于汉语的“同音噬词”规则，最终产生“2013年度”的说法。“2013年度”类表达产生之后，由于“年”与“度”句法位置相邻且高频连用，二者语义互相浸染，其内部结构发生了重新分析，结构边界得到了重新调整，最终“年度”凝固成词。“2013年度”从读音上来切分应该是[2013[年度]]，属于“2+2+2”或“4+2”音步模式。如果是[[2013年]度]，则属于“[4+1]+1”模式，这种音步模式不合法，非常别扭。音节韵律方面的要求直接导致了“年度”的固化与词汇化。

3.4 关于“月、天”搭配的个性分析

3.4.1 关于“月”构成的多词单位

通过语料库考察发现，“月”最大的特点是：1）“1-12”与“月”直接组合表示序数；2）表示基数时，“月”之前要加“个”。关于这两点，前文已述。另外，

有少数以“月”为单位所反映的事情，如“正月、满月、月子、月经、个把月、半月谈、蜜月、岁月”等等。要注意的是，新近的语料显示，“月度”也有成词趋势，但是本文所检索的语料库里未发现用例。

在口语中，一些带有指称性的表达[5]，如“这月、上月、下月、那月、哪月”里，有的以不出现“个”为常。依据“国家语委语料库”的统计，如“上月<48>”、“上个月<24>”；“下月<27>”、“下个月<10>”；“哪月<3>”、“哪个月<1>”。举具体用例来说，我们感觉“下月十五号”比“下个月十五号”要简洁经济。另外“三月半、七月半、八月半、九月半”等表示节日的短语，“半”可以位于“月”之后。以月份对特殊事件进行命名，如“十月革命、五月巡游”等。

3.4.2　由“天”构成的多词单位

“有一天”多词单位非常高频，相关数据如“有一年<33>”、“有一个月<9>”、“有一天<453>”。

“有一天”常用作句首前位成分（句首之前的成分）（方清明 2012b），其本身不能独立成句，具有篇章方面的启下性。其主要任务不在于担负线性句法结构中的基本角色，而在于具有语篇开启与触发功能。“有一天”主要为篇章叙述提供一个模糊的参考时间，其后经常出现停顿标记，如例（29-30）。“有一天”后面的逗号不仅起停顿作用，而且创立了话语边界。逗号可以凸显句首前位成分，起到标示语篇手段的作用。

（29）有一天，张学良从南京回到西安，一下飞机就直达军训团去点名，发现不少人回了家。

（30）有一天，张明山看见街上耍猴戏，他很爱那活猴，同时他发觉真猴子不该是白颜色的，为了更像真的，他当晚觑着父亲入睡时就把泥猴子都涂上灰黄色。

“有一天”的时间意义往往比较弱，主要是充当说话者回忆过去某个事件时的宽泛标记，具有明显的过去时标记功能。有时候，它可以被替换为“有一次、有一回”等形式，如例（31）里的“有一天”换作“有一次”，其功能基本相当。上述例（29-30）也可如此分析。

（31）有一天做实验，孙韶渝偶然发现一种可能达到上述要求的体系，从而打开了一个新的境界。

我们通过双库检索，与其他两个参照项“第一天”和“第三天”相比，“第二天”频率异常偏高。具体如表2：

表2. “第二天”的异常高频表现

项目	第一天	第二天	第三天
国家语委语料库	106	674	96
北京大学现代汉语语料库	3,942	15,361	2,009

从逻辑方面来看，序数词“第一、第二、第三、第四……”之间的地位应该是平等的。从人类认知主观性角度来看，语言认知赋予“第一、第二、第三”这三个序数更多的机会和重要性，其中“第一”应该具有更高的认知凸显性。方清明（2014）对“世界第一<97>、世界第二<29>、世界第三<21>”等极性意义多词单位的考察就是一个证明。但是在“第一天、第二天、第三天”这个序列中，最高频的不是“第一天”，而是“第二天”，数据见表2。

“第二天”的异常高频该如何解释呢？存在即合理，“第二天”看似违背逻辑和认知规律，其实恰好暗示“第二天”有着“第一天”或“第三天”所不具备的额外功能。“高频复现的多词单位不一定显示很强的意义关系，却极有可能会导致该词在功能上的扩张”（Mahlberg 2005）。观察语料索引行发现，“第二天”可以表示时间顺序，如例（32）；但很多时候含有“第二天”的结构主要是充当连接词功能，起到语篇连贯作用，而非单纯的时间序数用法，例（33）里“第二天”与“今天”相呼应，即“第二天”是指“今天”之后的一天。例（34-35）里的“第二天”都是用来连接两个先后相关的事件。例（36）里“第二天”的时间意义最弱，前面未出现相应的时间词与之呼应，它表达“后来”的意思，其连接意义已经较为显豁。

（32）其中的一只公鸡正在想：“第一天早晨有米吃，第二天早晨有米吃，……第九十九天早晨有米吃，所以今天，第一百天的早晨，一定有米吃”。

（33）我觉得做艺人有一点不好，就是不管你今天心情多么不好，第二天你该以灿烂的笑容面对大家的，还是得这么做。

（34）我走了整整一夜，第二天清晨才回到我的租住地。

（35）我好不容易才把他扶上出租车，回到家，他就一头栽进卫生间里狂吐起来。第二天他醒来，才把事情告诉我。我只好和声细语地安慰他。

（36）他曾经厚着脸皮约我去一家咖啡馆喝咖啡，我立即拒绝了。他不死心，第二天又来约我吃饭，被我再一次拒绝后，他终于死了心。

再系统观察例（32-36），我们不难发现“第二天”非常符合“联系项居中原则”（刘丹青 2003：68-73），即连词“第二天”用在两个被联系事件之间，起到关联前后事件的作用。“第二天”之前都有边界标记，如句号、逗号等等。这种边界的创

立，凸显了前事件与后事件的相对独立性，它们需要通过“第二天”这类时间标记来粘合和关联。正因为“第二天”符合“联系项居中原则”，经常起到语篇关联的作用，具有类似时间连词的功能，因此其频率比“第一天”高也就不足为奇了。

另外，我们检索“半天<425>”发现，“半天”很少为实指用法，多为说话人主观认为时间长的意思。“等你半天了”不是真的等了6个小时或12个小时，而是等你很久的意思。“半天”的主观长时用法体现了语义晦涩的一面，不具有切分性（张霄军 2010）。其他具体实例还有“半天才到、半天才来、沉吟了半天、大半天、好半天、叫了半天、看了半天、老半天、愣了半天、问了半天、想了半天、找了半天、折腾了半天”等。

4. 余论

基于语料库的定量定性研究能对语言现象作出更为全面的描述与分析，为传统研究提供必要补充和拓展，同时也在“很大程度上弥补了传统研究因语言材料不够充分而多依赖主观自省的不足，从而使传统的直觉经验方法转向基于实验和统计的方法”（潘璠 2012：30）。本文重视真实语料，并借助语料库和软件考察时间词“年、月、天”的搭配现象，发现它们与整数、半数、约数、位数搭配时，情况颇为复杂。传统研究仅以若干条句法标准考察特定词语是否具有同质性的做法，有待深入探讨和进一步检验。同质性分析不利于计算机对词性的加工处理，也不利于教学把握，而基于频率与搭配对“年、月、天”各自之间的异质性加以分析，做到了细化处理，既有利于计算机自动分词，也有利于对外汉语教学处理。

从频率和搭配的视角来看，“年”最为特殊，其搭配力大大高于“月、天”。如整数越大，越倾向于与“年”搭配；“年”既可以与位数词直接搭配，又可以与系位构造搭配。系位构造的数量越大，越不适合与“月、天”搭配。另外，“年”至少有9个义项是“月、天”所没有的。“年”的超高频率性和用法多样性充分说明“年”具有突出的例外性和特殊性。

“月”的特殊之处在于无标记序数用法与前加量词“个”才能表达基数用法，“月”不能与位数词直接搭配。“有一天、第二天”等含有“天”的多词单位具有语用语篇功能。总之，多个相关的高频词之间，往往只具有部分的同质性，很难有完全的同质性可言。

语料库语言学的任务是“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、离散的数据变为精炼的、系统化的知识”（冯志伟 2010）。其他时间名词也可利用本文的方法进行逐一分析。这里不妨再以时间词“最后”为例进行说明。“最后”表示“在时间上或次序上在所有别的之后”（《现代汉语词典》

2005：1823），因此在话语里，“最后”也多出现在其他话语之后。我们在语料库里发现“，+最后”形式达到769例，占总数的26.32%。这一数据说明“最后”的语义与总体的语篇结构具有象似性。例（37）里，“鉴真”是先花11年时间，然后才“最后”成功。

（37）鉴真为了赴日，自公元七四三年春开始，前后六次东渡，历时十一年之久，最后才得到成功。

我们以“最后”为节点词，以“一”为右搭配词，跨距为10R（R为英文Right的缩写）进行检索发现，“最后……一”搭配为486例，占总数的16.63%。“最后……一”搭配是具有浮现特征的半固定短语（semi-fixed phrase）。“最后……一”高频搭配，凸显了“最后”的某种属性，即“最后”不仅“在时间上或次序上在所有别的之后”，而且它在量性特征上属于少量，这一潜在语义特征与“一”的小量性具有语义上的相宜性。如例（38）里，“最后一次机会”凸显的是机会很少，只有一次了。我们进行了跨库验证，在“北京大学现代汉语语料库”里发现“最后一<15,355>”，约占“最后”总用例61,529的25%。

（38）她也明明知道对于她这是最后一次机会，因为明年她就超过高考规定年龄了，那样，几年寒窗苦读的心血将付之东流。

人们习惯说“我给你最后一次机会”，但一般不说“我给你最后两次机会”，更不会说“我给你最后三次机会”。因为“一次机会”才是最后的，“两次机会”里还有先后之分，不是严格意义上的最后。再如，我们经常能够听到“最后，我来说一句”，但是很少听到“最后，我来说两句”，更不会有“最后我来说三句”的表达。上述分析都说明，“最后……一”高频搭配具有“时间少、量性小”的特殊的语义、语用功能。这种用法的倾向性是在大量索引行例句的定量统计基础上，作出的定性分析。

总之，语料库语言学既是一种工具，也是一种范式，必定会“为语言学研究提供了新的途径、带来新的理念、新的方法，这方面的研究也必然使人们加深对语言本质的理解”（杨惠中 2010）。

注释

1. 关于语料库和AntConc3.2.4w软件的介绍请参看方清明（2014，2015b）。

2.“天”和“日”既有相同点，也有不同点。比如“三日就能完成这项工作”也能说成“三天就能完成这项工作”；但有时候又不相同，如“十月一日是国庆节”不能说成“十月一天是国庆节”。另外，“日”和“号”有时是同义的，如“十月一日”和“十月一号”。考虑到行文方便，本文优选“天”，仅在必要时选用“日”。对二者差异方面不作细致探讨，引文照旧。

3. 本文使用“*”星号表示不能说，“?”表示可接受度有问题。语料库语言学不仅关心能不能说，更关心多大概率能说。

4. 多词单位是指两个或多个词语高频复现的序列，详细介绍请参看方清明（2014a）。

5. 陆俭明（2013）论证了“这三个苹果”为有定指称，而非真正的数量表达，可参看。

参考文献

Bauer, L. 2001. *Morphological Productivity* [M]. Cambridge: CUP.

Mahlberg, M. 2005. *English General Nouns: A Corpus Theoretical Approach* [M]. Amsterdam: John Benjamins.

Packard, J. 2001. *The Morphology of Chinese: A Linguistic and Cognitive Approach* [M]. Cambridge: CUP.

Thompson, S. 2007. Three frequency effects in syntax [A]. In J. Bybee (ed.). *Frequency of Use and the Organization of Language* [C]. Oxford: OUP.

邓思颖，2012，再说“年、月、日”[J],《语言教学与研究》(2)：39-43。

丁声树等，1961,《现代汉语语法讲话》[M]。北京：商务印书馆。

方清明，2012a，现代汉语副词连用频率考察 [J],《汉语学报》(3)：87-94。

方清明，2012b，论现代汉语“XP的是，Y”有标格式 [J],《语言教学与研究》(1)：44-51。

方清明，2014，汉语抽象名词的语料库研究 [J],《世界汉语教学》(4)：532-544。

方清明，2015a,《现代汉语名名复合词的认知语义研究》[M]。北京：科学出版社。

方清明，2015b,《现汉》抽象名词语义韵的定量、定性研究——基于语料库和Antcon3.2.4w技术 [J],《辞书研究》(4)：17-23。

冯志伟，2010，双语语料库的建设与用途 [J],《现代外语》(4)：420-421。

胡　玲，2007，“X以来，Y”句的成句条件 [J],《汉语学习》(4)：87-92。

刘丹青，2003,《语序类型学与介词理论》[M]。北京：商务印书馆。

陆丙甫、屈正林，2005，时间表达的语法差异及其认知解释——从“年、月、日”的同类性谈起 [J],《世界汉语教学》(2)：12-21。

陆俭明，1987，说“年、月、日”[J],《世界汉语教学》创刊号：35-36。

陆俭明，2001，现代汉语时量词说略 [A]，载《语言学论丛》(第二十三辑) [C]。北京：商务印书馆。1-35。

陆俭明，2013，有关汉语数量表达的几个问题 [J],《澳门语言学刊》(2)：4-11。

吕叔湘，1999,《现代汉语八百词》[M]。北京：商务印书馆。

裴雨来，2011，再说“年、月、日”的词类 [J],《语文学刊》(4)：54-55。

王洪君、富　丽，2005，试论现代汉语的类词缀 [J],《语言科学》(5)：3-17。

杨惠中，2010，语料库语言学的应用研究与贡献 [J],《现代外语》(4)：421-422。

于立昌、吴福祥，2011，时间副词“一度”的语义演变 [J],《古汉语研究》(4)：27-30。

张霄军，2010，“两个半月”和“两个半天”——面向词法自动分析的涉数时间语素说略[J],《语言教学与研究》(3)：84-90。

中国社会科学院语言研究所词典编辑室，2005,《现代汉语词典》(第5版) [Z]。北京：商务印书馆。

周小兵，1995，谈汉语时间词 [J],《语言教学与研究》(3)：85-93。

朱德熙，1982,《语法讲义》[M]。北京：商务印书馆。

通讯地址：510631　广东省广州市华南师范大学国际文化学院

本质、特征、关系：外壳名词三分法及人际功能研究*

香港大学/哈尔滨师范大学 姜 峰

提要：“外壳名词”（shell nouns）喻指传递和表征命题信息的一类抽象名词（如fact、claim等）。以往研究集中于描述这类名词的语篇衔接作用，系统考察其人际功能（如，立场表达）却较少，其主要原因可能是对外壳名词语义分类模糊、人际功能理解不足。本研究以60篇学术论文为语料，提出“本质、特征和关系”的外壳名词语义分类新方法，并在学科话语共同体视阈内探讨外壳名词的立场建构人际功能。

关键词：外壳名词、语义分类、立场表达、人际功能

1. 引言

“外壳名词”（shell nouns）是由德国认知语言学家Hans-Jörg Schmid在总结前人相关研究（如Halliday & Hasan的“普通名词”、Francis的“回指名词”和“标记名词”等）的基础上提出的概念，于1997年首次进行了论述并在2000年的专著中给予了进一步丰富和补充，其意指“可用于传递和表征命题信息的外壳”的这一类抽象名词，其中“外壳”比喻其装裹和呈递信息的特性（Schmid 2000）。Schmid（2000）概括外壳名词具有“衔接（linking）”、“表征（characterization）”和“临时概念形成（contemporary concept-forming）”的功能。尽管这一概念是在认知语言学范畴内提出的，但它却被更多地应用到功能语篇分析中。然而在功能语篇视角下，这一概念在语义分类及其人际互动功能方面却存在不足，束缚了其在语篇研究领域的深入和发展，因此本文评议了外壳名词及相关同类名词概念在分类与功能方面的不足，并由此提出基于表达功能的语义再分类，同时以6个学科的60篇学术论文为语料，探讨外壳名词的立场劝谏功能。

* 本文系第七批中国外语教育基金项目“学术语篇中立场名词的结构范式与劝谏功能——基于语料库的研究”（ZGWYJYJJ2014A53）及黑龙江省哲学社会科学研究规划项目“基于语料库的学术语篇中话语立场与声音建构的研究”（14C048）的研究成果。衷心感谢香港大学Ken Hyland教授的悉心指导。

2. 外壳名词及相关概念：分类与功能

Schmid在外壳名词概念下关注的名词衔接和表征功能可追溯到Vendler（1967）在对事实与事件的哲学思考中提及的“容器名词（container noun）”。根据Vendler，事实或事件常常被名物化为容器名词，在伴随的句子中展开或施为。例如在John's *speech* was inconsistent.[1]中，演讲事件被名物化为容器名词speech。Halliday & Hasan（1976）将此类名词称之为“普通名词（general noun）”，并认为此类名词在与指示代词（the、this、that或such）作用下起到重要的语篇衔接功能，但并未指明具体的衔接关系与形式。Francis深入研究了此类名词的衔接和指示作用，认为它们不仅可以回指上文还可预示下文，她通过“回指名词（anaphoric noun）”（1986）和“标记名词（label noun）：回指标记（retrospective label）和预示标记（advance label）”（1994）分别研究此类名词的回指与预示衔接功能。Francis指出此类名词的回指作用有三个特点：谈论当前话语、指向前文已知信息、引出下文的新信息（1986：3-4）。这种篇章组织功能不仅存在于回指作用，也体现在下文预示作用中。Francis（1994）基于功能语言学的“概念、人际和篇章”元功能对此类名词的下文预示作用进行了阐述。作为句子的物质或关系等过程的参与者，此类名词添加当前话语的概念信息。

（1）The New York Post, which has been leading the tabloid pack, has added two salacious *details* to this bare outline. It reported that the alleged attack took place on a concrete staircase that runs from the Kennedy house to the beach. More sensationally, the Post claimed on Friday that Ted Kennedy, half naked, was romping round the estate with a second woman while the alleged attack was taking place. This allegation was at best dubious and at worst an outright fabrication.

Francis（1994）认为通过选用特定词汇details而非allegations，此句作者对预示的下文信息予以了评价，体现出标记名词在下文预示中的人际意义。这里标记名词details位于句子述位，构成新信息的焦点（end focus），可见它包含了篇章功能。Flowerdew（2003）对“指示名词（signalling noun）”的衔接作用进行了基于语料库的研究，他发现此类名词的衔接作用不仅发生在此前研究所述的段际、句际之间，也存在于句子内部。如下例：

（2）The *reason* why they're green is that they have chlorophyll.

Ivanič（1991）对此类名词的语篇衔接功能给予了重要补充，她认为“载体名词（carrier noun）”还会衔接文本外信息，这种文本外信息通常是文本作者与读者之间的共享知识。如下例中的food problems：

（3）Again, the associations can call on the resources of the Commonwealth Mycological Institute which maintains a collection of fungi many of which are of interest in research into certain *food problems*.

Schmid在此类名词相关研究的基础上将外壳名词上下文衔接作用归纳为“连结（linking）”功能和“路标（signposting）”功能（2000：339-359）。

尽管Schmid与Halliday、Hasan、Francis等学者都提及外壳名词具有评价功能，但是绝大多数研究仍集中在此类名词的衔接作用上（Flowerdew 2003；Francis 1994；Halliday & Hasan 1976；Ivanič 1991；娄宝翠 2013)。换句话说，对外壳名词的研究过多地重视其篇章功能，较少关注其人际功能。然而从篇章的社会互动视角而言，篇章功能是以读者为中心的作者靠近读者的交互过程，人际功能更大程度上是以作者为中心作者对命题和读者表达立场的互动过程（Thompson 2001；Hyland 2005)。因此对外壳名词立场评价和人际功能的研究不足，会进而忽略此类名词赋予作者表达立场、劝谏互动的机会与能动性，这构成了我们再思考的聚点。

外壳名词立场评价功能与其语义紧密联系。首先，此类名词的语义不具体，需要通过语境实现语义具体化（lexicalization）（Francis 1986；Winter 1982）。

（4）This question reflects the *assumption* that Sadat was the main architect of the policy as well as the one who achieved its realisation.

Francis（1986：31）表示上例中assumption的具体语义通过其后面that从句的内容得以实现。但是我们更要注意到，在语义具体化的另一方面，作者选用特定的词汇对语境语义进行评价、表达立场。例如上句中作者选用assumption而非fact或assertion表明其立场，即认为Sadat was the main architect of the policy as well as the one who achieved its realization是一种不确定的主观信息。而且这种名物化的立场加载“可以使过程、特征和评价变为事物，或者说把本来的动态范畴变为静态范畴”（朱永生 2006：87），由此可见外壳名词有助于作者固化其立场和评价。

此外，为了划分此类名词的语义范围，Francis（1986，1994）、Schmid（2000）和Flowerdew & Forest（2015）曾尝试提出不同的语义分类。Francis（1986）将此类名词分为言语名词、认知名词、文本名词和无主名词；Francis（1994）把分类调整为施为名词、语言活动名词、认知名词和文本名词。而且Francis以谈论当前话语为标准圈定外壳名词，可见她的分类是围绕语言的自反性展开的，没有囊括event、evidence、approach等与施事动作相关的名词。Schmid（2000）将外壳名词分为6种（事实、语言、认知、情态、事件和情境），Flowerdew（2015）类似地提出6种指示名词（行为、情境、事实、观点、言语和情态），但是他们的分类方法存在两方面不足。一是他们将包含态度评价的名词归入带有中性语义的事实类名词，如下例：

（5）The *advantage* is that there is a huge audience that can hear other things you may have to say.

Schmid（2000：126）将该例中advantage归为事实，掩盖了此句作者积极的态度立场。

二是没能有效区分事实类名词与情态类名词。譬如下例中，Schmid（2000：98）认为句中the fact that是事实陈述，然而the fact that也是对信息肯定性的认知情态判断（李战子 2005），也就是说，Schmid和Flowerdew没能有效地界定和区分其分类中“事实”与“情态”类别。

（6）We never expected John that they would just change it like that in the light of the *fact* that they’re going to change local government across the country any way.

可见，外壳名词的语义划分不清会阻碍其在语篇评价意义和人际功能方面的深入。此外评价和立场表达并非随意，而是受到话语共同体期望和接纳的约束，“体现共同的价值观和认知评判”（Thompson & Hunston 2000：6）。因此本文在多学科语料库基础上以立场评价视角对外壳名词的语义范畴进行重新分类，并探讨和回答以下两个问题：

1. 语篇作者通过外壳名词表达什么样的立场和评价？

2. 不同学科作者在话语共同体影响下的立场建构有什么不同？

3. 语料与方法

本研究基于BNC语料库的学术体裁子库中6个学科的60篇学术论文语料库（共约64万词符），学科分别为隶属“软学科”的人类学、社会科学和法学以及“硬学科”的医学、机械工程学和自然科学。在BNCweb检索平台[2]，选取学术体裁语料和学科文本构成列表，而后从该体裁语料中提取期刊学术论文。

本研究将外壳名词立场表达与劝谏互动的人际功能探究限定在“名词+that补足语从句”结构，因为该结构突出作者选用特定的外壳名词以表达其对补足语从句信息的立场判断（Biber *et al.* 1999）。与此同时，从信息处理过程的角度看待该结构，读者先接触外壳名词，然后再读取该名词的具体语义，这样外壳名词建立了作者态度立场的语用预设（程晓堂 2003；姜峰 2015），从而使读者在作者劝谏作用下融入作者的立场与视角。如下例中，作者选取danger而非issue或influence等词突显其对补足语信息Japan’s attempts to assert herself in Korea…持否定的态度立场，同时“名词+that补足语从句”结构在读者认知过程中预设作者的立场，助使读者认同作者的立场与视角，即让读者可能跟随作者的态度立场，也认定Japan’s attempts to assert herself in Korea…是一种“危险”。

（7）Japan had growing interests in mainland Asia, and there was a *danger* that Japan’s attempts to assert herself in Korea would bring her into direct confrontation with the expanding Russian empire.

（人类学）

语料文本经过TreeTagger词性标注后，针对“名词+that补足语从句”结构，运用AntConc 3.4进行正则表达式检索。对满足外壳名词使用条件的索引行进行反复浏览以观察外壳名词表达何种功能意义并给予分类（如表1），此后通过MAXQDA软件根据表1人工标注索引行中外壳名词的类别。

4. 语义再分类与立场劝谏功能

经过反复观察索引行中的外壳名词，发现它们或是标明事物（件）的本质，或是描述事物（件）的特征，或是表达事物（件）之间的关系，详见表1。

表1. 外壳名词的分类

本质	描述	例如
事物	标明元文本的语言	essay, report, paper
事件	事件、过程或事态	change, process, evidence
话语	言语行为或表达	argument, claim, conclusion
认知	认知观点或态度	decision, idea, belief, doubt
特征	描述	例如
品质	赞赏或批评好的或坏的品质	advantage, difficulty, value
方式	事件发生或事物形成的方式	time, method, way, extent
状态	认知、任务或动态情态判定	possibility, choice, ability
关系	描述	例如
因果、相似性等	因果、不同或相关	reason, result, difference

外壳名词标明事物（件）的本质，或是当前语篇的元文本，例如essay、report或paper；或是事件和过程，例如change、process或evidence；或是言语表达，例如argument、claim或conclusion；或是认知态度或观点，例如decision、idea或doubt。当外壳名词评判事物（件）特征时，一方面表达对事物（件）的赞赏或批评，例如advantage、difficulty或value；另一方面，描述事件发生或事物形成的方式或情境，例如time、way或extent；或是在认知、任务或动态情态方面对事态进行判断，例如possibility、choice或ability。此外，有些情况下外壳名词还可表达

事物（件）之间的因果、异同等关系，例如reason、result或difference。

另外，我们需要有效区分事件描述和认知情态评判。Labov（1972：381）认为“评判发生在主观阐述与背景信息或价值观对比的即刻”，因此Labov所述的评判发生的对比性为我们区分事实描述和认知情态评判提供参考。如下例：

（8）In order to analyse this case, it is plausible that we must posit a difference regarding our phenomenal evidence when we perceive red 1 and red 2, despite the **fact** that we cannot distinguish between the two shades of red.

（医学）

上例中despite表明the fact that从句与主句背景信息是一种对比关系，由此the fact是对that从句概念信息的情态评判，而非事件描述。据此，我们可以较为有效地辨别事件描述和情态判断。

基于功能对外壳名词进行语义分类，不仅可以认清作者对事物（件）本质的界定、对事物（件）性质的价值评判，而且能够体现外壳名词赋予作者表达构建立场的机会，以及作者如何通过立场表达与读者劝谏互动，因为作者对外壳名词的选择表达了其对物质信息和读者的立场，这种选择和立场表达并非随意，需要符合读者的认知期望和话语共同体劝谏规约，体现共同的价值观和认知评判。

5. 结果与讨论

本研究共发现1,302例“名词+that补足语从句”结构中的外壳名词，平均每篇学术论文22例。多数情况下外壳名词用以界定事物（件）本质，其中认知外壳名词频率最高，占所有立场表达的34.6%，表明作者更倾向于把概念信息定义为观点和主观认识。在描述事物（件）特征时，外壳名词用以判别确定性和必要性的状态子类频率最高，占所有立场表达的7.5%。作者用外壳名词界定具体化的元文本事物频率最低。详见表2。

如表3所示，外壳名词表达的立场建构因学科不同而相异。总体而言，软学科作者运用外壳名词表达立场的频率显著高于硬学科：前者每万字32.8例；后者每万字10.1例（LL=8.71，$p< 0.001$）。换句话说，几乎89%的外壳名词来自于软科学，体现出该类学科更倾向于评价讨论中的内容以及自己和他人的观点（Hyland 2004）。

表2. 外壳名词总体分布

类别	项目总数	每万词数目	占外壳名词总数比例
本质	970	15.2	74.5
事物	25	0.4	1.9
事件	287	4.5	22.0
话语	207	3.2	15.9
认知	451	7.0	34.6
特征	**250**	**3.9**	**19.2**
品质	71	1.1	5.5
状态	97	1.5	7.5
方式	82	1.3	6.3
关系	**82**	**1.3**	**6.3**
共计	**1,302**	**20.3**	**100.0**

表3. 外壳名词学科间分布每万词（百分比）

类别	人类学	社会科学	法学	医学	机械工程学	自然科学
本质	**17.7 (77.1)**	**14.9 (77.1)**	**18.5 (70.6)**	**10.4 (81.3)**	**6.6 (66.7)**	**7.5 (72.3)**
事物	0.2 (1.0)	0.3 (1.5)	0.6 (2.4)	0.6 (4.7)	0.2 (2.0)	0.3 (3.2)
事件	4.5 (19.6)	4.5 (23.5)	4.9 (18.8)	4.5 (35.2)	3.5 (35.4)	3.9 (37.7)
话语	4.3 (18.8)	2.6 (13.6)	4.5 (17.2)	1.7 (13.3)	0.8 (8.1)	0.9 (8.9)
认知	8.6 (37.6)	7.4 (38.5)	8.4 (32.2)	3.6 (28.1)	2.1 (21.2)	2.3 (22.5)
特征	**4.2 (18.3)**	**3.6 (18.4)**	**5.2 (19.9)**	**2.2 (17.2)**	**2.9 (29.3)**	**2.3 (22.7)**
品质	1.2 (5.4)	0.7 (3.5)	1.9 (7.4)	0.6 (4.7)	0.6 (6.1)	0.3 (3.2)
状态	1.8 (7.7)	1.3 (7.0)	1.8 (6.9)	1.1 (8.6)	1.2 (12.1)	1.1 (10.6)

（待续）

（续表）

类别	人类学	社会科学	法学	医学	机械工程学	自然科学
方式	1.2 (5.2)	1.5 (7.9)	1.5 (5.5)	0.5 (3.9)	1.1 (11.1)	0.9 (8.9)
关系	**1.1** **(4.6)**	**0.9** **(4.5)**	**2.5** **(9.6)**	**0.2** **(1.6)**	**0.4** **(4.0)**	**0.5** **(5.0)**
总计	23.0 (100.0)	19.3 (100.0)	26.2 (100.0)	12.8 (100.0)	9.9 (100.0)	10.3 (100.0)

如表2所示，在事物（件）本质类别中，事件和认知子类频率最高，表3则反映出两个子类在学科之间的分布不同。软学科更多地通过外壳名词建构认知立场，自然科学恰好相反，更多地表达事件立场。可见作者并非运用外壳名词随意表达立场，而是体现各自学科的价值观和认识论，反映不同学科范式是如何界定概念信息和世界知识的。根据Chafe & Nichols（1986），事件和认知分别与实证主义和理性思想紧密关联，反映文理学科不同的思维模式和获知方式。人类学、社会科学和法学等软学科常常依靠认知理解和理论建构，而医学、机械工程学和自然科学等硬学科则更大程度上从科学实验和实证证据中获取知识与信息（Becher & Trowler 2001)。例如：

（9）Both approaches prevented any understanding of the actual processes of local politics, and thus both helped to further the **orthodoxy** that local politics were largely absent in the immediate post-war period.

（法学）

（10）Present-day writers are generally in **agreement** that the position of the unmarried mother under the Poor Law showed remarkably little change before World War II.

（人类学）

（11）Our results thus provide the first direct **evidence** that tamoxifen does not have antioestrogenic effects on bone in postmenopausal women and indicate a possible oestrogenic effect.

（医学）

（12）Here we show, using hydroxy radical/DNase I digestion and differential helical phasing **experiments** that the curvature is directed towards the major groove

and is located in the GGGCCC, but not the CTAGAG segments.

（自然科学）

在描写事物（件）特征时，软学科作者使用外壳名词明显多于自然科学，前者每万字4.6例，后者每万字2.5例（LL=10.01，p<0.001）。这再次反映出学科之间在建立知识和劝谏措辞方面存在的差异。人文和社会学科作者多是通过个人对事物（件）的评价、阐释与推论建构知识，从而在与读者的互动中协商自己的观点。例如：

（13）The first is the empirical **problem** that without formally testing memory in such circumstances it is impossible to know whether there really is a memory impairment in such circumstances and if so how complete it actually is.

（社会科学）

（14）At the same time, however, we have to consider the **possibility** that members of society (consumers?) may not care about "representativeness" setting up, say, a left-of-centre paper is a considerably simpler task than making people read it.

（法学）

可见，作者借助外壳名词表达其对谈论信息内容的立场，在接近话语共同体的允许和规约中与读者劝谏互动。下表4反映每个学科在每个类别中使用频率最高的外壳名词。

表4. 每个学科在每个类别中使用频率最高的外壳名词

		人类学	社会科学	法学	医学	机械工程学	自然科学
本质	事物	report	report	report	report	report	report
	事件	fact	fact	fact	evidence	fact	fact
	话语	claim	claim	conclusion	suggestion	guarantee	conclusion
	认知	view	view	view	hypothesis	assumption	assumption
特征	品质	danger	danger	risk	advantage	advantage	advantage
	状态	possibility	possibility	possibility	possibility	probability	possibility
	方式	extent	way	way	way	way	way
关系		ground	ground	ground	ground	result	result

软学科作者使用最频繁的认知外壳名词是view，表达其对论述题元的个人观点和认识。硬学科作者最常用hypothesis和assumption等一类包含猜测和不肯定性

的外壳名词，构成他们对这种猜测和不确定性进一步实验观察和科学验证的基础。这反映出硬学科归纳式的研究方法和知识认知范式：科学学者往往通过测定实验变量检验现象假定和理论模型（Becher & Trowler 2001；Gilbert & Mulkay 1984）。如下2例：

（15）Part of northern Lewis was ice-free during the last glaciation and the last Scottish ice-sheet did not extend beyond the Outer Hebrides, contrary to the widespread **assumption** that this ice-sheet extended to the edge of the continental shelf.

（自然科学）

（16）Our data therefore do not support the **hypothesis** that transferring to human insulin by itself alters the frequency or experience of hypoglycaemia.

（医学）

在评价事物（件）品质时，硬学科和软学科作者分别倾向选用“正负”不同态度的外壳名词。软学科作者常常使用danger和risk此类否定态度的外壳名词，借以指出前人文献和现实事态的不足，如下2例。这种搭建研究定位的方法十分重要，因为软学科的知识问题与读者群体是不具体、模糊的，甚至是发散性的（Becher & Trowler 2001；Hyland 2004）。

（17）There is a **danger** that a concentration on spatial manifestation masks the realities of social processes, that space itself is fetishised.

（人类学）

（18）Moreover, the **risk** that conventional conflicts may get out of hand and degenerate into nuclear disaster is one with which mankind will have to live for ever, and it strengthens the argument for attempting to keep conflicts within some kinds of bounds.

（法学）

相比之下，advantage在硬学科作者作出态度评价时使用最为频繁。如下2例所示，他（她）们总是对自己的研究和模型给予正面积极的肯定。他（她）们的这种立场选择遵循硬学科线性的知识发展属性，新的科学发现需要通过促进既有的知识技术获得业界的认可（Becher & Trowler 2001；Gilbert & Mulkay 1984）。

（19）SYMAP has the **advantage** that no specialized hardware is required and is thus useful for introductory teaching at degree level.

（机械工程学）

（20）The design has the **advantage** that both solvent and solution compartments are easily rinsed out and the cell does not have to be dismantled if contamination

by permeation of low molar mass solute occurs.

（自然科学）

软硬学科不同的知识属性和认知范式在不同学科高频率的关系类外壳名词中亦得到体现。软学科作者在构建事物（件）之间的关系时多用ground以强调原因，而硬学科作者在论述关系时多用result以突出结果，如下4例。软学科的知识问题和读者群体发散，常涉及广泛的交叉学科，因此学术观点及其理据很大程度上建立在作者思辨阐释的合理性之上，需要作者在主观解说和劝谏措辞时提供有说服力的原因和理据基础。硬学科知识发展呈线性规律，新的论点和发现建立在现有的研究事实和实验证据基础上，因此例（23）和（24）的作者试图把自己新的研究融入于广泛认可的科学事实，将that从句的信息视为result，从而为自己的研究建立科学理据基础。

（21）They present short narratives of women's lives, and claim them for feminist therapy on the **grounds** that they emerge from women's personal experience.

（社会科学）

（22）The United Kingdom and certain other member states contest that view on the **ground** that the E.E.C. Treaty and, in particular, articles 7, 52 and 221 thereof cannot be interpreted as depriving the member states of their competence under public international law with regard to the registration of ships.

（法学）

（23）Numbers are also better suited to computer operations with the **result** that the process of searching the tree is made computationally simple.

（机械工程学）

（24）The above transformation is also related to the well known **result** that, if Z is a solution of Ernst's equation (11.8), then another solution is given by (12.16) where b is a real constant.

（自然科学）

6. 结语

通过本文研究，我们看到外壳名词给予作者丰富立场建构、表达人际意义的机会，在语境和话语共同体的合理规约内与读者劝谏互动，从而发展作者的观点和立论。因此，仅关注外壳名词发挥的语篇衔接作用只是其在功能语篇分析中的一个方面，它所蕴含的人际功能也是非常重要的组成部分。本文提出的对外壳名词基于功能的语义分类，能够帮助我们观察此类名词体现的评价意向，为今后外

壳名词的功能语篇分析提供借鉴。此外，本研究基于BNC的学术体裁子语料库，体现了BNC在学术语篇学科间性研究方面的可操作性及其语料资源价值。

注释

1. 为还原文献观点的完整性，第2部分的例句均摘自原文文献，而非取自本研究所用的学术论文语料库。第3部分及以后的例句均来自本研究的语料库。

2. http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php

3. 本文对例句进行标注：加粗为外壳名词，下划线为外壳名词表达的内容，斜体为前修饰语，并以括号加注语料学科来源。

参考文献

Becher, T. & P. Trowler. 2001. *Academic Tribes and Territories: Intellectual Enquiry and the Culture of Disciplines* [M]. Buckingham: Open University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Written and Spoken English* [M]. Harlow: Longman.

Chafe, W. & J. Nichols. 1986. *Evidentiality: The Linguistic Coding of Epistemology* [M]. Norwood: Ablex Publishing Corporation.

Francis, G. 1986. *Anaphoric Nouns* [M]. Birmingham: University of Birmingham.

Francis, G. 1994. Labelling discourse: An aspect of nominal-group lexical cohesion [A]. In M. Coulthard (ed.). *Advances in Written Text Analysis* [C]. London: Routledge. 83-101.

Flowerdew, J. 2003. Signalling nouns in discourse [J]. *English for Specific Purposes* 22(4): 329-346.

Flowerdew, J. & R. Forest. 2015. *Signalling Nouns in English: A Corpus-based Discourse Approach* [M]. Cambridge: CUP.

Gilbert, G. & M. Mulkay. 1984. *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse* [M]. Cambridge: CUP.

Halliday, M. & R. Hasan. 1976. *Cohesion in English* [M]. London: Longman.

Hyland, K. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing* [M]. Harlow: Longman.

Hyland, K. 2005. *Metadiscourse: Exploring Interaction in Writing* [M]. London: Continuum.

Ivanič, R. 1991. Nouns in search of a context: A study of nouns with both open- and closed-system characteristics [J]. *IRAL* 29(2): 93-114.

Labov, W. 1972. *Language in the Inner City: Studies in the Black English Vernacular* [M]. Pennsylvania: University of Pennsylvania Press.

Schmid, H. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition* [M]. Berlin: Walter de Gruyter.

Thompson, G. 2001. Interaction in academic writing: Learning to argue with the reader [J].

Applied Linguistics 22(1): 58-78.

Thompson, G. & S. Hunston. 2000. Evaluation: An introduction [A]. In S. Hunston & G. Thompson. (eds.). *Evaluation in Text: Authorial Stance and the Construction of Discourse* [C]. Oxford: OUP. 1-27.

Vendler, Z. 1967. *Linguistics in Philosophy* [M]. London: Cornell University Press.

Winter, E. 1982. *Towards a Contextual Grammar of English: The Clause and Its Place in the Definition of Sentence* [M]. London: George Allen & Unwin.

程晓堂，2003，名词化与语用预设 [J],《外语研究》(3)：19-23。

姜　峰，2015，中美学生论说文的立场名词表达——基于语料库的对比研究 [J],《外语与外语教学》(5)：8-14。

李战子，2005，从语气、情态到评价 [J],《外语研究》(6)：14-19。

娄宝翠，2013，基于语料库的研究生学术英语语篇中外壳名词使用分析 [J],《外语教学》(3)：46-49。

朱永生，2006，名词化、动词化与语法隐喻 [J],《外语教学与研究》(2)：83-90。

通讯地址：香港薄扶林道香港大学逸夫教学楼650室

汉语译文的成语特征研究：翻译共性假设再探

北京外国语大学 张汝莹

提要：本文利用汉语译文语料库（ZCTC）与汉语母语语料库（LCMC）对英汉翻译中汉语译文的成语特征进行研究，旨在检验Baker翻译共性假设在汉语译文中的适用性，并探讨该理论的可改进之处。基于肖忠华、戴光荣（2010）对于汉语译文中习语及词簇的研究，本文将两语料库中已标注的习语与词簇，对照《新华成语词典》进一步人工筛选出成语，从宏观和微观两个角度入手，对汉语译入语中的成语在形符与类符总量、词性分布、高频词、词义种类、词形结构等方面的特征进行研究。数据表明，与母语文本相比，汉语译文中的成语形符总量较少、但类符更为丰富、高频词与低频词使用呈现两极分化的趋势；成语意义多显化，以字面义为主；词形结构固定，缺少变换。由此可见，翻译共性理论中的显化、集中化及整齐化假设适用于汉语译入语，而简化假设似由“两极分化”解释更为妥帖。

关键词：成语、汉语译入语、翻译共性、简化原则、语料库

1. 引言

Fernando（1996：32）指出，成语是“由多个词素组成的习惯表达法，且词素不能或只能在有限范围内变化”。成语的意义不是单个词素词义的机械堆砌，而是由母语人士约定俗成的，是每种语言独有的“专利”。正因成语是一种文化的独特产物，成语翻译不仅是译者的“一块心病”，更对时下热议的翻译共性假设提出了挑战。该理论认为，翻译语言表现出独立于源语语言与目标语语言的一些独特规律性特征，又同时具备源语与目标语的一些特征，Frawley（1984）称之为“第三语码”。针对翻译文本表现出的一些共性特征，Baker（1996）将其归纳为四种假设，即明晰化、简单化、规范化与标准化。但是，这四种假设的提出主要是基于词源相近的欧洲语言。源语与目标语的不同可能导致翻译语言呈现不同的特点，有些无法用翻译共性假设解释，有些甚至与之相背。譬如，现有的汉语译入语研究表明，汉语译文虽然的确呈现显化特征，但相比源语却并不一定简化（黄立波2007；柯飞 2005；Xiao & Dai 2014）。

除了不同语种的个性特征外，一些语言特有表达法在译文中的特点也与共性

假设相左。Baker（2007）以英语译文中的习语为例，指出了该语言项在翻译中的复杂性，并总结出译文中习语的五项特征，是现有共性假设难以解释的。如她所说，如果规范化假设是正确的，译者在翻译中会偏好使用那些目标语中典型、安全的表达法，那么习语作为语言中的惯用表达法，可以使译文更加地道、行文更加流畅，因而应是译者的“最爱”，被大量使用。但事实上，译者在翻译中对习语的使用受到了诸多限制，不能像母语人士一样随心所欲地灵活使用。（1）母语人士经常故意变换习语中的一些成分，以此达到特定的效果，比如在广告中更改某些词素来实现独特的修辞效果，而译者则很少这样刻意改动习语成分，在译文语料库中也很难找到这样的例子加以佐证；（2）有些习语深深根植于语言背后的文化传统中，属文化专有项，因而译者在翻译中很难有机会用上这些具有强烈文化特色的习语；（3）Baker发现，英文中的习语具有不透明性，越是地道的习语规范性越差，而译者更倾向于使用规范性的表达法，因此在翻译中较少使用这类习语；（4）有些英文习语在语法上是不规则的，如规范化假设成立，则语法不规范的习语在译文中将很少出现，但相关语料库研究的结果却自相矛盾，有些语法不规则的习语的确较少使用，而有些则不然；（5）习语虽然在母语中使用广泛，但在译文中的使用频率却很低，且多以字面义为主。需要注意的是，Baker的这5点结论是以例证为主，并没有加以量化证实，另外，该研究是基于英语习语的特点，从下文研究可知，有些特征在汉语中并不适用，这也是本文将主要探讨的问题。

2. 成语的定义及意义

Baker研究中所说的idiom究竟是翻译为成语、习语还是惯用语，学界仍是众说纷纭。从词典释义来看，idiom“最为重要的一个义项是其整体意义是不能从单个的词演绎出来的”，包括方言及个人用语等，形式多样、没有字数限制（于薇薇、徐钟 2005：57）。而汉语中成语的定义本身就存在争议。从狭义的角度来说，成语属于熟语的一种，多由四字组成，可从字面义（如万紫千红、大呼小叫）和引申义（如卧薪尝胆）两个层面理解（辞海编辑委员会 1989），而从广义的角度看，成语是一个上义概念，囊括所有习语、俗语、熟语等习惯表达法（史式 1979：12-13）。虽然界定范围不同，但这两类定义都指出了成语有别于短语的独有特征：更加正式、更加抽象、句法结构固定、语法功能等同于句子（马国凡 1978：55-84）。由此可见，Baker研究中的idiom与广义上的汉语成语定义更为相似。然而，其中一个不同之处在于，与英语成语恰恰相反，汉语成语更为正式，更易出现在书面语体中，因而Baker关于习语不透明性的论断在汉语中并不适用。

由于成语的定义本就难以统一，有关成语意义的讨论更是仁者见仁，智者见智。刘洁修（1985：78）将成语意义分为字面义、引申义及比喻义3类。其中，字面义又分直接间接两类，是一切成语意义的基础；引申义是字面义的延伸；比

喻义“已然脱离了字面上的具体内容而发展衍变为更具抽象性和概括性的意义”。史式（1979：326）则完全反对上述分类，认为成语必须具有引申义，否则即为固定短语，成语的形成都有其源头，有些成语同时具有古义和今义。温端政（2006：134-141）则从共时与历时两个维度将成语意义划分为单义与多义、古义与今义、表层义与深层义、基本义与色彩义。综上所述，本文采用与《新华成语词典》（2002）一致的成语狭义定义，将其与习惯表达法区分开来。由于该定义中指出成语意义分为字面义与引申义两种，本文采用刘洁修（1985）的分类，将成语意义分为字面义、比喻义、引申义3类进行探讨。本文将两语料库中已标注的习语与《新华成语词典》一一对照，人工筛选出其中的成语，如这些成语的词典释义即字面意义，则该类成语为字面义成语；如词典释义包含“比喻为、喻为”等字眼，则该类成语为比喻义成语；如释义包含“后用……形容、后指、指”，则该类成语为引申义成语。

3. 肖忠华和戴光荣（2010）对汉语译文中习语与词簇特征的研究

肖忠华、戴光荣（2010：81-82）利用汉语译文语料库（ZCTC）与汉语母语语料库（LCMC）对汉语译文中的习语及词簇特征作了详尽的分析研究。两语料库中习语根据词性不同进行标注，分为nl名词性习语、vl动词性习语、al形容词性习语、dl副词性习语及bl名词修饰性习语5类。该研究依据这些已有标注对每类习语按语体类型进行了统计：

表1. LCMC和ZCTC语料库取样方案

代码	取样类型	代码	取样类型
A	新闻报道	J	学术、科技
B	社论	K	一般小说
C	新闻评论	L	侦探小说
D	宗教	M	科幻小说
E	技术、商贸	N	武侠小说
F	通俗社会生活	P	爱情小说
G	传记和杂文	R	幽默
H	其他：报告和公文等		

肖忠华、戴光荣（2010）研究发现，除侦探小说外，习语在汉语母语中的出现频率均高于汉语译文，因而Baker的规范化假设在汉语译文中显然难以成立，

如前文所述，笔者认为，这主要是汉语与英语中的习语在语域方面的差异所致，前者更为正式，后者则是非正式文体特征。

肖忠华、戴光荣进一步对两语料库中的词簇作了详细研究，但由于词簇相比习语而言涉及范围更广，包括了所有“预制的、反复重现的语言片段”，属于“广义上的习语”（Baker 2007），因而与本文关注的狭义概念上的成语相去甚远，在此便不再赘述。

Xiao（2011）在其研究中也明确指出了其研究受到了语料库标注的限制，“除非语料库有特别标注，否则Baker的（例证法）研究是成语研究的唯一可行途径”（Xiao 2010：6）。如前文所述，由于成语定义本身就存在争议，因此确实难以找到一个统一的标准对狭义概念上的成语进行标注。

鉴于这种局限，本文以肖忠华和戴光荣（2010）的习语研究为基础，利用汉语译文语料库（ZCTC）与汉语母语语料库（LCMC）（第二版），将两语料库中已标注的习语以《新华成语词典》为依据，人工进一步筛选出狭义概念上的成语，从宏观和微观两个角度探究这些成语在汉语母语与译文中表现出的不同特征。另外，由于有些汉语成语的词素位置及个别词素的变化并不影响其意义（如天翻地覆/地覆天翻、独具匠心/匠心独具），但在类符统计时却被记为不同的类符，从而影响对于成语意义的量化研究，因此，为了更好地对比成语在汉语母语与译文中的结构差异，笔者根据苹果iSO平台自行设计了一个统计软件。该软件的原理是先以词表中的一个词为基准，将后续词项与之一一进行比对，如果成语中四个字有三个字都相同，则将这两个成语视为结构相似，一轮比对完成后，再以下一个词为基准，进行下一轮比对……之后，笔者再将软件自动筛选结果进行人工核对，除去其中结构相似、但意义不同的成语项，如有始有终/有始无终、自下而上/自上而下，最后将两语料库中结构相似的成语对进行对比分析。

4. 汉语译文中的成语特征

4.1 汉语译文中成语的宏观特征

表2. LCMC与ZCTC成语归一化形符总量比较（每十万词中的使用频率）

语料库 / 词性	LCMC		ZCTC	
	形符	百分比	形符	百分比
nl	138	24.67%	26	7.58%
al	61	10.86%	54	15.63%

（待续）

（续表）

语料库 / 词性	LCMC		ZCTC	
	形符	百分比	形符	百分比
vl	313	56.13%	245	70.45%
dl	46	8.33%	22	6.34%
成语总量	557	100%	347	100%
习语总量	781		613	
百分比	71.22%		56.66%	

从表2中可以看出，汉语译文中的成语特征与肖忠华、戴光荣（2010）研究中的习语特征趋势一致，即成语在汉语母语中的使用频率高于汉语译文。就词性分布来看，不论是母语还是译入语，动词性成语比例最高，母语中名词性成语紧随其后，而译文中则是形容词性成语位居第二。另外，母语语料库中名词、动词、形容词及副词性成语形符总量比例为2.97 : 6.74 : 1.30 : 1，而译文语料库中的比例则为1.20 : 11.12 : 2.47 : 1，可见，成语在译文中词性分布更加集中，而在母语中则较为分散。这一特征也证实了共性理论中的集中化假设，即“译文文本具有相对较高的同质特征”（Laviosa 2002）。

从成语在习语中所占比例来看，母语中的比例要明显高于译文。由于汉语成语相比一般习语而言更加正式，也更具有文采，因而成语在习语中的比例可以很好地体现出文本中的语言是否地道，但同时，由于其不透明性，读者的理解负荷也随之增加。表2中的数据说明，汉语译文在成语使用方面呈现显化趋势，更方便读者理解，由此可见，共性理论中的显化假设在汉语译文中也是成立的。

表3. LCMC与ZCTC成语类符总量比较

语料库 / 词性	LCMC	ZCTC
nl	214	141
al	195	153
vl	1,569	1,149
dl	74	33
成语类符总量	2,052	1,476
习语类符总量	2,875	1,959
百分比	71.37%	75.34%

表3显示，成语类符特征与表2中的形符特征大体一致——母语都高于译文。最大的不同是两语料库中成语/习语的类符比例相当，即汉语译文中的成语使用频率较低，但种类较多。虽然看似矛盾，但从译者的角度出发，这种特征其实并不难理解：由于成语是文化的特有产物，因而很难在目标语中找到完全对等的翻译，所以译者在翻译过程中使用成语会非常谨慎，除非源语中的成语在意义上与目标语中的翻译高度一致，否则译者一般不会用目标语中一个意义有所偏差的成语来牵强附会地翻译源语。这样，译文中的成语重复率较低，也便不难理解译文成语使用频率明显低于母语这一现象了。

表4. LCMC与ZCTC成语类符频率分布

	LCMC		ZCTC	
	频率	百分比	频率	百分比
[10, + ∞)	29	1.41%	33	2.24%
[5,10)	130	6.34%	101	6.84%
(1,5)	838	40.84%	591	40.04%
=1	1,055	51.41%	792	53.67%

表5. LCMC与ZCTC中的最高频成语

	LCMC	ZCTC
nl	14	11
al	13	16
vl	19	22
dl	25	27

表5中的数据进一步佐证了表4的结论：汉语译文中的低频成语（频率=1）明显高于母语，因而汉语译文中的成语种类更加多样，但使用频率较低。但与此同时，译文中的高频成语（频率≥10）数量与比例也高于母语，且最高频成语使用次数也多于母语。这说明，译者在翻译过程中会经常使用特定成语，这也再次验证了共性理论中集中化假设在汉语译文中的适用性。在笔者看来，汉语译文成语这种两极分化的趋势看似矛盾，但实则统一：译者在翻译过程中为了“保险”会重复使用一些高频成语，而在用成语翻译源语中一些具有文化特色的表达法时则

会非常谨慎，从而导致大量低频成语的出现。

综上所述，从宏观的角度来看，汉语译文中的成语具有以下特征：使用频率较低、种类更丰富、以动词性成语为主、高频词与低频词呈现两极分化的趋势。这主要是源于译者对于目标语的规范化使用以及审慎的遣词造句。由此亦证明，共性理论中的显化、集中化假设在汉语译文中是成立的。

4.2 汉语译文中成语的微观特征

基于上述宏观统计数据，本文从高频及中频成语（频率≥5）入手，依据《新华成语词典》给出的释义将两语料库中的高中频成语分为字面义成语、引申义成语及比喻义成语三类，统计结果如表6。

表6. LCMC与ZCTC中成语意义比较

成语意义类型	LCMC				ZCTC			
	类符		形符		类符		形符	
	频率	百分比	频率	百分比	频率	百分比	频率	百分比
字面义	95	59.38%	679	57.98%	96	72.18%	1,230	80.29%
引申义	43	26.88%	362	30.91%	25	18.80%	205	13.38%
比喻义	22	13.75%	130	11.10%	12	9.02%	97	6.33%
总量	160	100%	1,171	100%	133	100%	1,532	100%

表6表明，不论是母语还是译入语，字面义成语均为主流，其次为引申义成语，比喻义成语使用频率最低。值得一提的是，汉语译文中的字面义成语无论是从种类还是使用频率都远超汉语母语。另外，比照前文中汉语译文成语的高频词表便可发现，高频词多以字面义为主，而低频词则多为比喻义及引申义成语。由此可见，译者在翻译中更倾向于使用成语的字面义，方便读者理解，这与显化假设相符，具体而言，属于Xiao & Dai（2014）所说的语义显化（另外两类为语法显化及逻辑显化）。

为了进一步说明汉语译文中成语多字面义这一特点，笔者以成语“讨价还价”为例，从LCMC与ZCTC中提取了所有包含“讨价还价”一词的语料进行分析，如表7、8所示。

表7. LCMC中“讨价还价”索引行

序号	文件	索引行		
1	A06	命令如山倒，没有任何	讨价还价	的余地。
2	A29	操作员说，这是交易双方在	讨价还价	呢！
3	B05	经过激烈的交锋和	讨价还价	，海部总算相继打掉安倍派和渡边派的要求。
4	E15	交易时，	讨价还价	的声音一浪高过一浪。
5	F36	他们的要价与物品的市场价格明显不符，并在	讨价还价	中以各种花言巧语一再自动降价。
6	G03	外来的投资者开始和傲慢的上海	讨价还价	了，一些土地应有的价格权威开始动摇了。
7	G04	姚红林认为太贵，便与包工头	讨价还价	，想定一个比较合理的价格。
8	G09	出版社反过来保护个人，代表个人利益向国家	讨价还价	？因此，外国专家不得不从版权的ABC讲起。
9	G09	你对国家奉献自己的作品，是不该	讨价还价	的。
10	K01	奖金另发，生活费包干。别再	讨价还价	了，这是那一带保镖的最高待遇。
11	K14	黄瓜茄子熟肉朝鲜泡菜鸡蛋花生仁儿嫩豆腐，叫卖声	讨价还价	声男女老少嚷成一片。

表8. ZCTC中“讨价还价”索引行

序号	文件	索引行		
1	E16	与汽车经纪人	讨价还价	往往是不愉快的经历。
2	F02	买家不得	讨价还价	，除非钻石的重量超过10克拉。
3	F02	从业员多为犹太人因此，善于砍价至关重要。乐纳	讨价还价	时的规律是：……
4	F02	经过一通	讨价还价	，她将出价提高到每克拉5,900美元。
5	F43	卡麦洛已经巴不得立刻加入。卡麦洛小眼睛一眨	讨价还价	：“我会成为一个很出色的小丑，我应该1天3块钱。”
6	J13	这样，小国获得了更多	讨价还价	的权利。

《新华成语词典》释义中，“讨价还价”具有字面义及比喻义两个义项，字面义指“买卖双方一方要价，一方还价”，比喻义则指“谈判或接受任务时提出条件，斤斤计较”。从表7、8中可以看出，汉语母语语料库“讨价还价”的11条索引行中有4条是“讨价还价”的比喻义，比例近三分之一。相比之下，汉语译文语料库中该成语的6条索引中只有1条与比喻义有关，比例只有母语中的一半。这再次证明，译者在翻译中较少使用成语的比喻义或引申义，主要还是以传达成语的字面义为主。

表9. LCMC中结构相似的成语

翻天覆地	天翻地覆
得意洋洋	洋洋得意
闻名遐迩	遐迩闻名
如痴如醉	如醉如痴
阴错阳差	阴差阳错
抖擞精神	精神抖擞
此起彼伏	此伏彼起
人人皆知	尽人皆知
人杰地灵	地灵人杰
迎来送往	送往迎来
离乡背井	背井离乡
独具匠心	匠心独具
无声无息	悄无声息
尽心尽力	尽心竭力
闻所未闻	前所未闻

经笔者自行编写的软件统计，汉语母语语料库中共有15对结构相似、意义相同的成语，而汉语译文语料库中则没有这样的结构相似成语。这种显著的差异再次证明，母语人士在使用成语时更加灵活，使成语在结构上更加富于变换，而译者在翻译时则更为保守，严格遵循成语结构规范，一般不对成语词素顺序及内容进行改动。这也从词形角度验证了规范化假设在汉语译文中是成立的。

综上所述，汉语译文中的成语在宏观及微观上表现出以下特点：与母语相比，使用频率较低、种类更为多样、以字面义为主、缺少结构变换、高频成语与低频成语呈现两极分化趋势。这主要是由译者在翻译中对目标语的规范使用以及遣词造句时的谨慎所致。

5. 讨论

上述汉语母语与译文语料库中有关成语特征的数据看似自相矛盾，实则对立统一，表明译者在遣词造句的过程中其实是受到多重因素的共同影响，这不仅是由于源语与目标语的语言差异，更有译者从读者理解角度出发作出的考量。本文对汉语译文中成语宏观特征的描述，与肖忠华、戴光荣（2010）研究中习语的特征一致，再次验证了共性理论中显化假设在汉语译文中的适用性。即译者在翻译中较少使用具有强烈文化特色的成语，以此减轻读者的理解负担，本文微观数据分析中字面义成语居多这一特征也印证了这一点。

Baker（2007）指出，如果规范化假设成立，那么译者为了使译文更加地道流畅，本应大量使用习语，但在实际操作中，译者使用习语却是慎之又慎，为该假设提供了反证，但笔者以为不然。首先，如Baker所说，英语中的习语主要出现在非正式语体中，用法越是地道，口语化程度越高，而译者在翻译中为了遵循书面语规范，一般较少使用口语化的表达，因此译文中的习语使用频率要比母语低。但汉语则恰恰相反，汉语中狭义概念上的成语较为正式，很多都是出自古代典籍或民间传说，本就多见于书面语中，所以成语的使用频率越高，文本就越正式，而这正是译者追求的目标。笔者认为，汉语译文中成语使用频率低于母语，不是因为成语的非正式性，而是由于英汉语言的不对等，译者很难找到合适的机会来使用具有引申义与比喻义的成语，但为了使译文更正式，只能大量使用一些字面义成语。不过，如果源语中的表达的确与汉语成语的意义对等，那么译者还是会毫不犹豫地使用成语的，这也就造成了大量低频成语的出现，从而很好地解释了前文中令人匪夷所思的高低频成语两极分化现象。

由此可见，简单化假设在汉语译入语中并不完全成立，译者并不是一味地将源语中的文化专有项简化处理，只要能在目标语中找到基本对应的表达法，译者还是会最大程度上地保留源语特征，只有当源语与目标语差异较大，难以找到对等项时，译者才会选用更加抽象概括的表达法进行适当的简化处理，这也致使高频成语使用频率增加，因而翻译共性理论中的简单化假设似用两极分化来解释更为合适。

6. 结论

本文基于肖忠华、戴光荣（2010）有关汉语译文中习语与词簇的研究，利用汉语母语语料库与汉语译文语料库，从宏观和微观角度进一步探讨了狭义概念上的成语在译文中的使用特征。研究表明，与母语相比汉语译文中的成语，使用频率较低，但种类更多样、以字面义成语为主，结构缺少变换、高频成语与低频成语呈现两极分化趋势。基于以上特征，笔者认为Baker（2007）研究中指出的习语使用

与规范化假设之间的矛盾其实是对立统一的，而简单化假设如果改为两极分化似乎更为妥帖。

参考文献

Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead [A]. In H. Somers (ed.). *Terminology, LSP and Translation* [C]. Amsterdam: John Benjamins. 175-186.

Baker, M. 2007. Patterns of idiomaticity in translated vs. non-translated text [J]. *Belgian Journal of Linguistics* 21: 11-21.

Fernando, C. 1996. *Idioms and Idiomaticity* [M]. Oxford: OUP.

Frawley, W. 1984. Prolegomenon to a theory of translation [A]. In W. Frawley (ed.). *Translation: Literary, Linguistic and Philosophical Perspectives* [C]. London: Associated University Press.

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

Xiao R. 2010. Idioms, word clusters, and reformulation markers in translational Chinese: Can "translation universals" survive in Mandarin? [R]. *Paper presented at the 2010 conference of Using Corpora in Contrastive and Translation Studies*. Edge Hill University. 27-29.

Xiao, R. 2011. Word clusters and reformulation markers in Chinese and English: Implications for translation universal hypotheses [J]. *Languages in Contrast* 11(2): 145-171.

Xiao, R. & G. Dai. 2014. Lexical and grammatical properties of translational Chinese translation universal hypotheses reevaluated from the Chinese perspective [J]. *Corpus Linguistics and Linguistic Theory* 10(1): 11-55.

辞海编辑委员会，1989，《辞海》[Z]。上海：上海辞书出版社。

黄立波，2007，《基于汉英/英汉平行语料库的翻译共性》[M]。上海：复旦大学出版社。

柯　飞，2005，翻译中的隐和显 [J]，《外语教学与研究》(4)：303-307。

刘洁修，1985，《成语》[M]。北京：商务印书馆。

马国凡，1978，《成语》[M]。呼和浩特：内蒙古人民出版社。

商务印书馆辞书研究中心，2002，《新华成语词典》[Z]。北京：商务印书馆。

史　式，1979，《汉语成语研究》[M]。四川：四川人民出版社。

温端政，2006，《汉语语汇学教程》[M]。北京：商务印书馆。

肖忠华、戴光荣，2010，汉语译文中习语与词簇的使用特征：基于语料库的研究 [J]，《外语研究》(3)：79-86。

于薇薇、徐　钟，2005，IDIOM是译成"成语"还是"习语"？[J]，《上海翻译》(3)：56-58。

通讯地址：100089　北京市北京外国语大学英语学院

中国英文科技文献中的词束特征调查*

中国科学院大学 钱玉彬

提要：本文参照Biber等人的研究框架，从结构和功能角度系统考察中国英文科技文献中高频4词词束的特征。文章依次讨论词束结构分类、动词词束的类联接、介词词束和名词词束的共现结构、副词词束、词束语篇功能的分类和解释。文章还进一步从结构和功能关联性的角度探讨词束的分布模式。这些发现为深入探讨我国科技工作者学术英语能力的构成和科技论文的写作教学提供启发。

关键词：科技文献、词束、结构、功能、语料库

1. 研究背景

词束（lexical bundle）属于短语学范畴，是语料库驱动研究范式下的一种短语单位。相关研究往往将相对连续的词语序列视为词束，并主要沿着以下3条路线展开分析：（1）词束形式；（2）词束功能；（3）特定语域的短语学特征。

传统研究关注的是词束形式，围绕词束的频次分布和语法结构等话题展开（Altenberg 1998；Biber *et al*. 1999；濮建忠 2003）。这一研究路线仍然是近年二语习得领域的热点，不少课题采用对比分析的视角，揭示学习者对本族语词束的掌握情况（Chen & Baker 2010；马广惠 2009；许先文 2010），或者不同水平学习者所用词束在数量和结构上的差异（Hyland 2008a；徐昉 2012；张霞 2010）。

作为对传统研究的补充，一些学者侧重描述词束功能，调查词束构建文本的作用及其本身的语用含义。例如，有的学者将词束的语篇功能进行归类（Biber & Barbieri 2007；Hyland & Tse 2009），或就词束在组织语篇信息和体现语篇行为的表现进行细致的观察（Poos & Simpson 2002；李晶洁、卫乃兴 2013；娄宝翠 2010）；有的围绕立场（何安平 2011）、身份（徐昉 2011）、话语交互（许家金、许宗瑞 2007）等话题讨论词束的语用功能，或依据词束的语用功能探析词束和语言能力的关系（Liu 2003；Wood 2009；刁琳琳 2004）。

此外，还有不少学者将上述两条研究路线予以结合，考察特定语域中的词束

* 本文受到中国科学院大学青年教师科研启动基金资助（Y551034Y00）。

表现。例如，有学者深入讨论学习者口语中词束的形式分类与功能差异（卫乃兴 2007），有的根据词束分布模式探索学术论文内部的词汇——语法特征（梁茂成、刘霞 2014），有的则从对比视角，分析口笔语（Biber 2006；Biber & Barbieri 2007）、课堂教学与大学教材（Biber *et al.* 2004）、学科之间（Hyland 2008b）等不同语域中词束特征的差异。

这3条研究路线往往相互交叉，揭示词束特征的全貌。不过，结合国内外研究实际，我们不难发现仍有一些问题值得关注：（1）尽管多数研究基于学习者语料库，但是，将中国英文科技文献作为观察语料库的研究并不多见；（2）中国科技工作者既有别于英语学习者，又不同于英语母语者，观察其词束使用特征以揭示这一群体语言特点的研究有待深入。探讨这些问题有助于从宏观上了解我国英文科技文献的语域特征和语言特点，也有助于推动词束研究的横向发展。

本文首先明确词束的定义和提取标准，继而根据Biber *et al.*（1999）对词束语法结构的分类和Biber & Barbieri（2007）对词束语篇功能的分类，系统考察中国英文科技文献中高频词束的特征，进而探讨结构和功能的关联性。

2. 研究方法

2.1 研究问题

本文的研究问题是：（1）中国英文科技文献中高频4词词束在结构和功能上存在哪些特点？（2）中国英文科技文献中高频4词词束的结构和功能存在何种关系？

2.2 语料的采集与整理

本研究采用的语料由1996-2012期间刊发的221篇国内英文学术论文构成。为了提高研究信度，我们对语料来源进行了如下筛选：（1）所选论文涵盖Elsevier ScienceDirect目录索引的20个科技领域；（2）所选论文取自我国自然科学引文索引（SCI）刊物；（3）所选论文由中国学者创作。所有语料均使用头部标签将论文标题、作者姓名和联系方式予以保留，剔除影响索引结果的数学公式和图表，整理后的语料库库容达到811,419形符（词次），词汇总量为36,556类符（词）。

2.3 词束的定义和抽取方案

由于研究角度和对象的不同，词束的界定至今仍存有争议（马广惠 2011）。本文将沿承Biber等人的定义，将词束理解为“高频率重复出现的词语序列”（Biber 2006：133；Biber & Barbieri 2007：263）。结合定义，我们依据如下标准提取词束：（1）词束的提取频点为每百万词（mw）重复出现20次；（2）词束长度为4词；（3）每个词束至少分布在5个不同文本中。这些标准是基于经验而设定的，

例如Biber *et al.*（1999）按照频点为10次/mw，且在不少于5个不同文本中出现，提取3词词束进行研究；Biber等人对40次/mw的4词词束展开分析（Biber 2006；Biber & Barbieri 2007；Biber *et al.* 2004）；Hyland（2008a）则将目标词束设定为20次/mw的4词词束，并且分布于语料库十分之一以上的不同文本中。基于经验而设定的标准并无特定的统计学依据，只是为了得到合理数量的词块，即数量不宜过多，同时得到语料中的典型词块（许家金、许宗瑞 2007：438）。

结合上述提取方案，我们采用以下步骤展开分析：（1）使用AntConc的N-gram功能穷尽语料库中所有4词词束；（2）依据既定标准提取目标词束；（3）对目标词束进行整理，剔除包含在头部标签中的词束，内部以标点（如连接符、逗号）间隔的词束也排除在外；（4）将整理后的目标词束按照语法结构和语篇功能进行分类描述；（5）分析词束结构和功能的关联性。

3. 数据分析与讨论

3.1 中国英文科技文献中高频词束的语法结构

统计发现，国内英文科技文献语料库中在至少5个文本中重复出现20次/mw以上的4词词束共计168类符。我们根据Biber *et al.*（1999）对词束结构的分类，将168类符词束分为5大类，11子类。如表1，最常见的词束是动词词束（43.11%），其次是介词词束（24.55%）和名词词束（23.95%），最后是副词词束（1.8%）。

表1. 语料库中词束结构的分布情况

类别	子类	类符数	比例	范例
动词词束	（动词短语+）that-从句片段	13	7.78%	is well known that
	（动词/形容词+）to-从句片段	3	1.80%	can be used to
	第三人称代词+动词/形容词短语	10	5.99%	it is necessary to
	系动词be+名词/形容词短语	12	7.19%	is one of the
	被动态动词+介词短语	34	20.36%	can be divided into
介词词束	带有of-结构的介词短语片段	26	15.57%	in the case of
	其他介词短语（片段）	15	8.98%	at the same time
名词词束	代词/名词短语+be（+...）	1	0.60%	That is to say
	带有其他后置成分的名词短语片段	7	4.19%	the fact that the
	带有of-结构的名词短语片段	32	19.16%	one of the most
副词词束	状语从句片段	3	1.80%	as shown in Fig
其他词束		12	7.19%	as well as the
总计		168	100.00%	

动词词束分布最广，子类也最为丰富，其中“被动态动词 + 介词短语”结构占比最高。尽管如此，子类存在共性特征，即大多数含有“be + 短语片段”的类联接形式。表2列出了频数排在前50位的类联接形式，从中可以推断动词词束的语用功能，既可以是用于客观描述，例如is one of the、is shown in Fig、is based on the，也可以是表达主观态度，例如can be used to、it should be noted。

表2. “be + 短语片段”类联接

	be	短语片段
	be（is/are/was/were/be/been）	one of the；shown in Fig；listed in Table；seen that the；based on the；well known that；located in the；due to the；proportional to the；consistent with the；related to the；the number of；determined by the；explained by the；used as the；found to be；shown in Figure；assumed to be；widely used in；
can/ should	be	used to；seen that；divided into；expressed as；found in；concluded that；obtained by；written as；described as；noted that
It/it	is/was	necessary to；found that；well known；clear that；easy to；
It/it can/ should	be	seen；concluded；noted
results	are	shown in
that there	is	a
has	been	shown to

介词词束和名词词束具有近似的类符数，并且两者都以“带有*of*-结构的短语片段”为主要结构，尤其值得关注。结合索引行分析，我们发现两类词束具有较强的粘合度，即高频率介词词束与频数相当的名词词束共现，一同组成5词词束。表3列出了这些共现词束，相关的索引行例如：

例1：the potential energy can be expressed *as a function of the* loop inductance and the phase difference cross the JJs in the cubit.

例2：*On the basis of the* CCSM3 climate model, two additional idealized experiments under assumption that the nuclear radiation was leaked…

例1中as a function of系高频介词词束，a function of the为高频名词词束，两者共现组成5词词束as a function of the。同理，例2中的5词词束on the basis of the

则由介词词束on the basis of和名词词束the basis of the构成。

表3. 介词词束和名词词束中的共现结构

介词词束			名词词束		
as	a function	of	a function	of	the
on	the basis	of	the basis	of	the
at	the beginning	of	the beginning	of	the
in	the center	of	the center	of	the
for/in/with	the development	of	the development	of	the
in	the formation	of	the formation	of	the
with	the increase	of	the increase	of	the
for/in	the study	of	the study	of	the
on	the surface	of	the surface	of	the

副词词束的类符数最少，子类别也最为单一，全部由as引导的状语从句片段构成。它要么用于指示图表数据，例如as shown in Fig/Figure/Table，要么作为描述结果的标记语，例如as a function of、as a result of。尽管如此，如表4，从单个词束的频数来看，副词词束as shown in Fig的使用频数却最高。这一现象能够反映科技文献的语域特点，即包含大量图表数据，因而往往需要借助形式固定的短语用于描述数据。例如：

例3：In the traditional clustered sensor network, the nodes are divided into clusters, *as shown in Fig*. 2.

表4. 词束频数表

频数	词束	结构类型
137	as shown in Fig	副词词束
93	on the other hand	介词词束
90	at the same time	介词词束
85	is one of the	动词词束
80	It can be seen	动词词束

需要指出的是，尽管词束结构的分布规律必然受词束功能影响，但是，若要论证两者之间是否存在映射关系，就必须对词束功能进行详细描述，并在此基础上进一步考量结构和功能的关联性。

3.2 中国英文科技文献中高频词束的语篇功能

我们利用索引行观察词束使用的语境，进而依据Biber & Barbieri（2007）的研究成果对词束功能进行归类。目标词束分为3类，包含7个子类别。如表5所示，具有指示功能的词束最为常见，占所有词束类符的64.29%；其次是语篇组织词束，比重为22.03%；立场词束最少，占13.68%。

表5. 语料库中词束功能的分布情况

类别	子类	类符数	比例	范例
指示词束	详述属性词束	76	45.24%	an important role in
	模糊词束	6	3.57%	similar to that of
	时间/地点/文本指示词束	26	15.48%	as shown in Fig
语篇组织词束	识别/聚焦词束	32	19.05%	is one of the
	主题阐述/解释词束	5	2.98%	on the other hand
立场词束	态度/模态词束	8	4.76%	it is easy to
	认知立场词束	15	8.92%	one of the most
总计		168	100%	

指示词束在语料库中占主导，其功能在于指示客观或抽象的对象或者篇章本身，既可以指称对象本身，也可以指称对象的特定属性，包含3个子类：

（1）详述属性词束，用于说明对象数量、体积、形式、抽象特点等属性。这类词束既能够从形式上客观描述研究的内在特征，例如，an important role in、the center of the、of the most important；又能够于逻辑上解释研究的路径，例如，on the basis of、with the increase of、in the presence of。

（2）模糊词束要么用于表示参照物的模糊性，要么用于指示类似的参照物。本语料库中的模糊词束只具有后者功能，例如，is similar to that、the same as the、the same as that。它们利用概念之间的认知框架，采用转喻的修辞手法，用概念A指代概念B，从而使概念B更加容易识别、理解和记忆。

（3）时间/地点/文本指示词束，用于标记时间或空间属性，或者被当作文本

锚点。常见的时间/地点词束包括at the same time、is located in the、in the treatment group等，能够描述对象的时间/空间设置；文本锚点词束包括is shown in Fig、in the present study、at the beginning of，能够前指或回指文献内容。

语篇组织词束用于组织文本，反映文本的前后关系，包含两个子类：

（1）识别/聚焦词束，用于标识重要内容或突出观点。在本语料库中，其功能体现在描述客观事实，例如，The results showed that、be seen that the、It was found that用于描述研究结果；can be obtained by、was used as the、method is used to用于描述研究方法；in the study of、with respect to the、to that of the则用于描述研究内容。

（2）主题阐述/解释词束对命题进行详细的解释和说明。例如，on the one hand、on the other hand用于引导不同话题，从而实现对命题全面完整的论述；as well as the、That is to say、is due to the则通过对命题的相关概念进行补充说明，增强文章的可读性。

立场词束用于表达对语篇内容的态度和评价，包含2个子类：

（1）态度/模态词束，用于表现作者态度。这种态度既可以传达作者对学术研究的责任感，例如，it is necessary to、It should be noted；又可以显示他们对研究客体抽象能力的认识，例如，play an important role、it is difficult to、in good agreement with。

（2）认知立场词束，用于表现作者对话题确定性、可能性的认知评价，并且绝大多数采用“情态动词+被动态动词”结构，例如，It can be seen、can be used to、can be regarded as。利用认知立场词束的模糊性，研究主体可以弱化论述的主观性，而使其观点更加易于接受。

综上，我们推断不同词束功能之所以存在类符数和频数的差异，可能和科技文献的语域特点相关。科技文献侧重报告研究数据和步骤，因此更多地依赖指称词束，用来进行客观指代；语篇组织词束更多地用于识别或聚焦科技文献的研究方法、内容和结果，同样具有较强的客观性。立场词束使用得最少，意在隐蔽研究主体的态度，减少论述的主观性，将科技文献的重心转移至对研究客体的客观描述。正是科技文献的语域特点，潜移默化地影响科技工作者的认知选择，使用那些表达恰当、同时语义客观的词束。

3.3　中国英文科技文献中高频词束结构和功能的关联性

图1展示的是词束功能在不同词束结构中的分布情况。可以看出，指示词束不仅分布最广，而且在各类词束结构中比值最高。尽管如此，词束功能在各类词束结构中的分布仍存在组内和组间的差异。

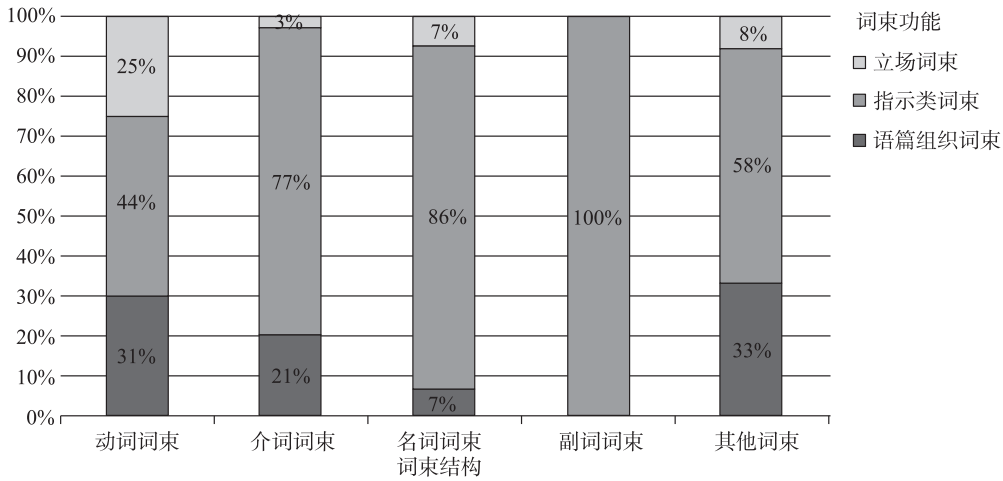


图1. 词束功能在词束结构中的分布情况

从组内对比来看，词束功能在动词词束和副词词束中的分布差异明显：

（1）唯有在动词词束中，立场词束和语篇组织词束的比重之和（56%）高于指示词束（44%）。结合动词词束的分类，可以看出，动词词束较其他词束结构不仅子类别更为丰富，并且功能也更加多变，是使用最为灵活的结构。这也恰好解释了动词词束在科技文献中类符数最多的心理因素，即使用动词词束既能够客观指示研究对象的属性，又能够在指示的同时，有条理地引入和突出话题内容，并灵活融合主观立场态度，因此最能够满足作者体现研究客观性、逻辑性和表现学术观点等多种目的的需求。

（2）唯有副词词束的结构与功能存在单一的对应关系，即副词词束完全由as引导的状语从句片段构成，并且仅仅具有指示功能。这些副词词束是as shown in Fig/Figure/Table、as a function of、as a result of。尽管它们的结构和功能单一，但是单个词束的频数分布却排在前3%，这一矛盾能够解释科技文献的语域特点，也能够为解释副词词束的分布规律提供心理依据，即科技文献包含大量图表数据，以及由数据衍生出的公式和结论，需要寻求相对固定的表达形式进行指代和描述，于是，文献作者便将具有指示功能的副词词束固化为对应的表达，储存于心理词典，一旦需要便能够快速提取。

如表6所示，词束功能在各类词束结构中的分布存在显著的组间差异（χ^2=29.858, p<0.01），具体表现为：（1）立场词束在动词词束中使用过多（实际计数高于期望计数），而在介词词束、名词词束和副词词束中使用过少（实际计数低于期望计数）；（2）指示词束在动词词束中使用过少，而在介词词束、名词词束和副词词束中使用过多；（3）语篇组织词束在动词词束中使用过多，而在介词词束、名词词束和副词词束中使用过少。

这些差异揭示出词束功能在词束结构中的分布模式，同时也反映出中国科技工作者词束使用的心理机制，即更愿意使用动词词束来表达立场和组织语篇，使用介词词束、名词词束和副词词束进行指示。

表6. 词束结构*词束功能交叉制表

χ^2=29.858, df=8, p<0.01		立场词束	指示词束	语篇组织词束	合计
动词词束	实际计数	18	32	22	72
	期望计数	9.9	46.3	15.9	72.0
介词词束	实际计数	1	28	8	37
	期望计数	5.3	23.8	8.1	37.0
名词词束	实际计数	3	36	3	42
	期望计数	5.8	27.0	9.3	42.0
副词词束	实际计数	0	5	0	5
	期望计数	0.7	3.2	1.1	5.0
其他词束	实际计数	1	7	4	12
	期望计数	1.6	7.7	2.6	12.0
合计	实际计数	23	108	37	168
	期望计数	23.0	108.0	37.0	168.0

4．结论和启示

参照Biber等人对词束结构和功能的研究，本文得出以下结论：从结构分类的角度上看，中国英文科技文献的高频词束可分为5个类别，11个子类；动词词束分布最广，且多数采用“be + 短语片段”的类联接结构，用于客观描述或表达主观态度；介词词束和名词词束具有较强的粘合度，通常共现组成5词词束；副词词束的子类别最为单一，且均由as引导的状语从句片段构成。从功能分类的角度上看，中国英文科技文献的高频词束分为3个类别，7个子类；基于词束的语篇功能，我们推测词束在语篇中的分布特征或与科技文献的语域相关。

基于上述观察，本文进一步探讨词束结构和功能的关联性，指出词束功能在各类词束结构中的分布存在明显的组内和组间差异，具体表现为：组内对比揭示，唯有构成动词词束的立场词束与语篇组织词束的比重之和高于其指示词束；唯有副词词束的结构与功能存在单一的对应关系。基于组内差异，我们讨论了动词词束与副词词束频数分布各异的心理依据。组间对比揭示，动词词束更多用于表达立场和组织语篇，而介词词束、名词词束和副词词束更多用于指示。基于组间差异，我们推断词束的分布特征或与科技工作者的心理机制密切相关。

这些发现为深入探讨我国科技工作者英语学术语言能力的构成和科技论文的写作教学提供一些启示。首先，学术语言能力强调的是用语言来展现学术能力，因此其语言特点必须是在学术相关的语言环境中不断积累和培养的，并具有学术语域的共性特征。科技文献中的高频词束更像是一种“程式语”，是科技工作者约定俗成的、长期使用的、固有的语言表达方式，因此无疑是学术语域共性特征的具体表现。词束的类符数和频数、对词束结构类和功能类的选择便能够在一定程度上反映科技工作者的学术语言能力。其次，在科技论文的写作教学当中，可以从词束角度提升学生的文字表达水平。我们不仅应当从产出量上加强学生使用词束的意识，还应当从语言品质上提高学生使用词束的质量，即引导他们使用结构正确、功能恰当，且适用于学术语域的词束。建议将国外英文科技文献制成语料库，与本文的中国英文科技文献语料库形成对比，让学生在比较中外科技工作者词束使用特征的同时，更好地掌握通用学术词束，甄别词束质量。

参考文献

Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations [A]. In A. Cowie (ed.). *Phraseology: Theory, Analysis and Applications* [C]. Oxford: OUP. 101-122.

Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers* [M]. Amsterdam: John Benjamins.

Biber, D. & F. Barbieri. 2007. Lexical bundles in university spoken and written registers [J]. *English for Specific Purposes* 26(3): 263-286.

Biber, D., S. Conrad & V. Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks [J]. *Applied Linguistics* 25(3): 371-405.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English* [M]. Essex: Pearson Education.

Chen, Y. & P. Baker. 2010. Lexical bundles in L1 and L2 academic writing [J]. *Language Learning and Technology* 14(2): 30-49.

Hyland, K. 2008a. Academic clusters: Text patterning in published and postgraduate writing [J]. *International Journal of Applied Linguistics* 18(1): 41-62.

Hyland, K. 2008b. As can be seen: Lexical bundles and disciplinary variation [J]. *English for Specific Purposes* 27(1): 4-21.

Hyland, K., & Tse, P. 2009. Academic lexis and disciplinary practice: Corpus evidence for specificity [J]. *International Journal of English Studies* 9(2): 111-129.

Liu, D. 2003. The most frequently used spoken American English idioms: A corpus analysis and its implications [J]. *TESOL Quarterly* 37(4): 671-700.

Poos, D. & R. Simpson. 2002. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English [A]. In R. Reppen, S. Fitzmaurice & D. Biber (eds.). *Using Corpora to Explore Linguistic Variation* [C]. Amsterdam: John Benjamins. 3-23.

Wood, D. 2009. Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study [J]. *Canadian Journal of Applied Linguistics* 12(1): 39-57.

刁琳琳，2004，英语本科生词块能力调查 [J],《解放军外国语学院学报》(4)：35-38。

何安平，2011，语料库视角的英语口语“立标语块”探究 [J],《外语教学理论与实践》(1)：25-31。

李晶洁、卫乃兴，2013，学术文本中短语序列的语篇行为 [J],《外语教学与研究》(2)：200-213。

梁茂成、刘　霞，2014，语篇内部的短语学特征分布模式探索——以学术论文为例 [J],《解放军外国语学院学报》(4)：1-11。

娄宝翠，2010，学习者硕士学位论文中的词串研究 [J],《当代外语研究》(9)：27-34。

马广惠，2009，英语专业学生二语限时写作中的词块研究 [J],《外语教学与研究》(1)：54-60。

马广惠，2011，词块的界定、分类与识别 [J],《解放军外国语学院学报》(1)：1-4。

濮建忠，2003，英语词汇教学中的类联接、搭配及词块 [J],《外语教学与研究》(6)：438-445。

卫乃兴，2007，中国学生英语口语的短语学特征研究——COLSEC语料库的词块证据分析 [J],《现代外语》(3)：280-291。

徐　昉，2011，中国学生英语学术写作中身份语块的语料库研究 [J],《外语研究》(3)：57-63。

徐　昉，2012，中国学习者英语学术词块的使用及发展特征研究 [J],《中国外语》(4)：51-56。

许家金、许宗瑞，2007，中国大学生英语口语中的互动话语词块研究 [J],《外语教学与研究》(6)：437-443。

许先文，2010，非英语专业研究生二语写作中的词块结构类型研究 [J],《外语界》(5)：42-47。

张　霞，2010，基于语料库的中国高级英语学习者过程词汇使用研究 [J],《当代外语研究》(5)：41-44。

通讯地址：100049　北京市中国科学院大学外语系

农科学术英语论文语料库的创建*

华中农业大学 刘 萍 黄小倩 刘 珊

提要：本文介绍华中农业大学“农科学术英语论文语料库”的创建情况，包括语料收集、文本的转换与清洁、标记、赋码等。借助CQPweb网络语料库系统，将该语料库部署在校园网供博士生和本科生的学术写作教学。语料库应用于教学的效果调查表明：调查对象认为语料库的应用有利于提高学术英语写作水平，有意愿在写作实践中继续运用语料库这一工具与资源；同时调查对象也指出：现有的语料库资源有待充实，语料库的检索操作仍显复杂。本研究旨在为专门用途语料库的建设提供一些参考。

关键词：农科学术英语论文、语料库、学术英语、CQPweb

1. 国内外专业用途英语语料库的建设

20世纪60年代，世界上第一个电子化英语语料库布朗语料库问世后，各种类型、用途、规模的语料库相继建成。按研究目的，可将语料库分为通用英语（EGP）和专门用途英语（ESP）语料库。布朗语料库、英国国家语料库等均属于通用英语语料库。而专门用途英语语料库是特定领域语言的反映（黄大网等2010），包括商务、法律、医学等专业方向的语料库，广泛应用于词典编纂、机辅翻译、语言教学等。专门用途英语中的很多用法在通用英语语料库中未有收集，因而无法检索到例子，那么就需要建立专门用途英语语料库。Sinclair（2003）曾指出大型语料库的建设已趋缓，建设规模较小、专业针对性更强的ESP语料库将是语料库建设的发展趋势。国际上有代表性的ESP语料库，包括Hyland建设的多学科学术期刊论文语料库（含8个学科，240篇论文，130万词）、Swales（2003）开发的学术口语语料库（录音转写170万词）。另外，英国考文垂、雷丁等大学（2004-2007年）联合建设了英国学术英语写作语料BAWE（British Academic Writing of English）库，该库子集life sciences（140万词）的收录涉及农业科学（134篇）、生物科学（169篇）、食品科学（124篇）3个农业学科400多篇，代表

*本研究得到2014国家社科基金项目“农科英语语料库的建设与其在ESP写作教学中的应用研究”（14BYY162）、中央高校基本科研业务费专项资金资助项目（2662015PY193）华中农业大学2014年度校级重点建设课程（科技英语写作）项目的资助。感谢北京外国语大学许家金教授、博士生吴良平对农科学术英语论文语料库的建设与本文的撰写所提供的支持与帮助。

着由高层次英语母语学习者撰写的学术论文。当然，life sciences子语料库并非农科英语专属语料库，它同时也收录了医学、健康、心理学等领域的学术论文。除此之外，未见国外其他农科英语语料库建设的相关文献记载。

在国内，1983年由杨惠中和黄人杰主持建成的上海交大科技英语语料库JDEST是国内建设的第一个学术英语语料库。自20世纪90年代以来，很多学科领域都相继建设了专门用途英语语料库，如军事、海事、法律、商务、医学、计算机等学科的ESP语料库（赵晴 2010；董爱华 2013）。迄今为止，国内有关农科英语语料库建设的文献只有3个检索结果，包括西北农林大学（王敏、李丽霞 2014：6855）建设的动物科学国际期刊论文语料库（100万词）、王景怿（2015：51）主持建设的英汉/汉英双语畜牧业小型语料库，但这两个语料库不仅库容量小，而且只涉及农科领域某一个专业方向。有学者（范晶晶、李丽霞 2014；栗娜 2015）呼吁创建农业学术英语语料库，并提出了建设构想，这表明国内部分学者已经意识到农科英语语料库建设的必要性和重要性。

2. 建设农科英语语料库的必要性

语料库被广泛用于语言教学与研究，正如Leech（1993）所言："从科学方法的角度，语料库研究方法是一种更为强有力的方法，因为其结果是可以验证的。"Johns（1991）提出了"数据驱动学习"（Data-driven Learning，简称DDL）。国内的语料库专家论证了语料库的频率统计、概率分析等功能对于写作中词块、类联接、语义韵律等语言使用方面的研究价值（李文中 2001；王克非、黄立波 2008；王克非、秦洪武 2012），桂诗春等（2010）专家论证了语料库与ESP发展的互动关系，呼吁利用语料库促进ESP教学发展。

农业是涉及国计民生的支柱产业，众多从事农业科技研究的科研人员和高层次的学习者均有发表SCI论文、在国际上推广农科研究成果的需求。因此，农科英语语料库的建设及其在写作教学中的应用具有紧迫的现实意义。依托国家社科基金项目，项目组创建了农科学术英语论文语料库，旨在为农科专业高层次学习者的ESP写作教学及SCI论文写作过程提供资源、工具与方法，提升农科英语论文的写作质量与刊发率，最终促进农科成果在国际上的推广。

农科学术英语论文语料库是根据农学专业分类，收集农科文献中完整的学术英语论文文本而建成的电子文库。该库收集的文本包括已发表的权威期刊论文和农科专业硕士生、博士生撰写的学术论文。该库的创建意义有：1）多学科、跨学科、交叉学科的农科学术英语论文语料库的创建可基本满足ESP写作教学多方面的需求，例如为教材建设、大纲与词表的制定、农科词典编纂、农科专业翻译和语言培训提供资源与工具；2）语料库的应用将促进写作教学的改革。长期以来，写作教学被认为费时低效，枯燥的讲授与单调的操练导致产出格式化、形式化（蔡少莲 2008），语料库数据驱动的写作教学方式可提高写作教学成效，促进ESP写作教学改革。

3. 农科学术英语论文语料库的建设流程

3.1 农科学术英语论文语料库及其网络检索平台

农科学术英语论文语料库包括农科方向的SCI期刊论文语料库（336个完整论文文本，220万词）与学习者语料库，后者收录了硕士生、博士生出于真实SCI发表目的撰写的学术论文（306个完整论文文本，140万词）。这两个平行的语料库有着相同的结构框架，其下是按照学科分类的专业论文子集，包括九个学科：植物科学（ZWKX）、动物科学（DWKX）、生命科学（SMKE）、园艺林学（YYLX）、水产科学（SCKX）、食品科学（SPKX）、农科机械（NKJX）、农业经济（NYJI）、资源环境（ZYHJ）。每篇论文按照“学科名称汉语拼音的首字母组合 + 数字”命名，例如植物科学专业的第34篇论文，命名为ZWKX34。为了凸显传统农科专业的地位，336篇期刊论文中4个传统的农科专业（植科、动科、生科、园林）的文本数量（在50篇以上）比其他学科（在25篇以上）要多。期刊论文语料库代表英语母语者专家语料库，学习者语料库代表汉语母语者语料库。这两个语料库除了按照学科划分的9个学科子语料库之外，还按照论文的部分分类，建立了6个论文部分子语料库，包括摘要ABS（abstract）、引言INT（introduction）、方法材料MET（methodology）、结果RES（results）、讨论DIS（discussion）、结论CON（conclusion）。语料库构架如下图所示：

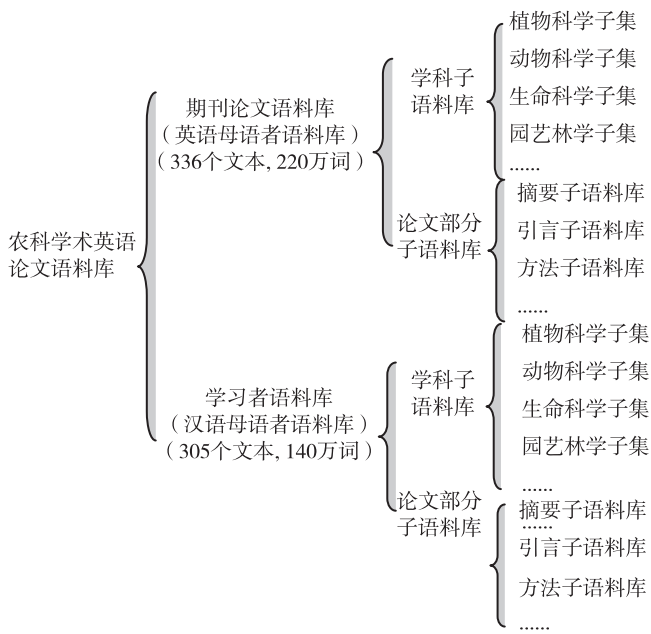


图1. 农科学术英语论文语料库构架

为了实现语料库资源共享与服务写作教学及其研究的目的，项目组在北京外国语大学语料库团队（特别是吴良平老师）的帮助下，架设了基于校园网的语料库检索平台（corpus query processor，简称CQPweb）。该平台主体为农科学术英语论文语料库。同时，为了语言的对比研究，得到授权后，项目组又上传了BAWE语料库。该库收录了高层次英语母语学习者撰写的学术论文，可进行英、汉语母语语言使用的对比比较，亦可进行专家和学习者的语言对比研究。

3.2 语料选取的标准与文本元信息标记

9个农科专业方向相关的期刊论文来自9个农科院系的专家、教授们亲自下载推荐的权威期刊论文。他们对刊源的权威性和论文本身的质量进行了严格把关。期刊论文语料的选择标准包括：1）论文的第一作者必须是英语母语者（依据姓名、国籍、工作单位、作者介绍判断）；2）源PDF论文的分栏排版最好是一栏或两栏；3）文章具有典型的SCI论文结构，即包括摘要、引言、方法材料、结果、讨论、结论几个结构板块，剔除了板块结构不够明显的论文。例如，刊源为*Science*、*Nature*等自然科学顶级期刊上介绍性、报道性或综述性的文章被剔除；4）文章的长度以10页左右为佳，不宜过长，也不宜过短。最终我们收集到9个农科专业满足条件的论文共336篇（220万词）。其中大部分期刊论文的作者来自美国和英国，少部分来自于加拿大、澳大利亚以及新西兰等英语国家，刊源为包括*Nature*和*Cell*等在内的国际权威期刊。

学习者语料库的语料来自华中农业大学9个农科专业方向的硕士生、博士生出于真实SCI论文发表目的而撰写的研究论文。论文由专门提供学术英语论文语言修改服务的华农学术写作工作室提供。项目组从该工作室采集到从2010年1月至2015年3月修改前的完整学术论文共300多篇，经语料加工，最终得到有效文本305篇（140万词）。总的来说，由于来源局限，学习者论文语料比期刊论文语料少，论文的学科分布均衡性也不及期刊论文。由于论文修改遵循自愿原则，并非所有写作者都选择向华农学术写作工作室提交论文修改请求，有的院系的学生直接向国外公司提交论文修改。今后，我们将加大华农学术写作工作室的宣传力度，同时，我们将建立与其他农科院校的合作，争取在更广范围内获取学习者撰写的研究论文，以便拓展语料来源，以求学习者语料库语料的代表性及学科均衡性。

语料采集之后的工作便是文本元信息的统计。论文的编号与命名、文本字数、作者国籍、期刊来源、作者姓名、论文标题等信息被填写在Excel表格中，以便对某些信息数据进行计算分析。

3.3 文本转换与清理、赋码、标记与子语料库的建设

3.3.1 文本转换与清理的两种方法

文本元信息统计之后，便是对语料的加工处理。期刊论文的语料加工往往从PDF格式转换开始，学习者语料库则从DOC格式文本转换开始。在此我们介绍两种从PDF到TXT的文本格式转换方法。

方法一：此法采用北京外国语大学中国外语教育研究中心和外语教学与研究出版社联合举办的暑期语料库培训课程中所介绍的方法，该方法对期刊论文语料进行了两次转换，即从PDF到DOC，再到TXT。文本格式转换需借助的工具软件和操作步骤展示如下：

（1）PDF文件转换成DOC文件。这一步骤所用的工具软件为Adobe Acrobat。首先对页眉、页脚裁剪删除，之后另存为DOC格式文件。然后删除DOC文件中的噪音信息，如刊源信息、作者信息(包括作者姓名、通讯地址等)、图表、注释、参考文献等等，有些文本信息的取舍取决于研究目的，例如是否保留致谢部分取决于研究需要。

（2）将所有由PDF文本转换得到的DOC文件整理到一个文件夹中，利用“DOC to TXT”软件一次性转换成TXT格式文件。但是转换后的文本中乱码现象时有发生。针对这一问题，课题组成员尝试着利用Nitro Reader软件将PDF格式直接转换成TXT格式文本，但是该软件不具备裁剪页眉、页脚的功能，所以增加了手动删除的工作量，但基本上不会出现乱码和正文内容板块顺序错乱的情况，所以两种软件各有利弊。

（3）核对检查TXT文件信息。对照PDF源文件，检查TXT文件。检查对象包括在删除、复制、文本格式转化过程造成的文本内容的遗失、重复、板块结构顺序错乱以及拼写错误。例如，我们发现经过两次格式转换后，有些单词中“fl”和“fi”的字母组合被显示为“?”。如果某类错误有规律可循，便可以使用EditPad Pro软件进行查找和替换，批量处理，或者在PowerGREP软件中逐个修改。

（4）TXT文本的清洁。经过上述检查步骤得到一个初步的TXT基础文本，但是这绝非是最终可以使用的清洁文本，因此需要对TXT基础文本进行清洁。TXT文本清洁工具软件可采用PowerGREP软件。

方法二：

我们不妨把上述文本转换与清理的方法称之为方法一。采用方法一，项目组完成了200多篇期刊论文语料的加工与处理。实践表明，通过方法一加工1篇期刊论文文本的工作，包括从PDF到TXT文本的转换与清洁大致需要花费40-60分钟。语料库建设后期，华农的博士生参与了语料资源的共建共享，有博士生推荐了一种快速有效的文本转换方法，我们将其称之为方法二。方法二加工处理1篇期刊

论文平均所需时间在10分钟以内，是方法一所需时间的1/6或1/5，依靠此种方法得到的TXT文本，基本不需要文本清洁，即文本转化与清洁两项工作一并完成。使用方法二实现从PDF到TXT直接转化的3个简单步骤如下：

步骤1：用以下网址搜索所需要的英文文献：http://www.gfsoso.netscholar；https://scholar.ghbcx.com；https://scholar.wddmz.com。以第一个网址为例，在谷粉搜搜中找到提供全文资源的期刊论文。

步骤2：在网络页面找到相关全文资料后，不需要下载全文，可直接在网页上点击Full Text (HTML)浏览全文。

步骤3：直接选中目标，复制内容，新建TXT文件，把复制的内容直接粘贴到TXT文件中，便得到TXT目标文件。

方法二的优点在于：1）基本上不会出现断行和乱码现象；2）可以避免删除图表及其注释的大量繁琐工作，省时高效；3）操作简便易行。此法得到的文本可以放在PowerGREP软件中进行删除空行的简单处理就能得到我们需要的清洁文本。同时利用谷粉搜搜检索期刊论文也是对期刊论文质量的检验。但是，此法的局限性在于过分依赖网络，仅能加工处理网络上能够检索到的文献，不能处理非网络版的文献。

3.3.2 赋码与标记

文本赋码将有利于文本的检索。利用正则表达式进行的复杂检索对文本赋码提出了要求。不同工具软件甚至要求不同的赋码形式。目前，英语文本的赋码主要有TreeTagger和CLAWS两种赋码，二者皆可借助软件自动完成。总的来说，CLAWS赋码比TreeTagger赋码的精确程度更高。华农语料库对TXT原始文本进行了TreeTagger和CLAWS两种赋码，以便适用于不同的检索工具。

为便于语料的提取，项目组对336篇期刊论文和305篇学习者论文文本（总共641篇）中title、abstract、body 3个部分进行了标记。标记方法是在标注对象的开始位置与结尾位置分别加上一对尖括号。例如，对标题的标记，是在标题前加<title>，在标题尾部加</title>，标记后的标题可提取，而对摘要和正文的标记，同样可以达到提取的目的。

3.3.3 子语料库的建设

为了聚焦论文不同部分的写作教学及其研究，在全文语料库建设的基础上，我们进行了论文各部分（摘要、引言等）子语料库的建设。论文部分子语料库的建设遇到了以下一些问题：

1）不同期刊的论文写作规范要求不一致，导致某些论文6个部分的结构不是很清晰。例如，有的论文将Abstract界定为Summary，其位置可能放在论文的开头，也可能放在论文的结尾；有的论文中Abstract部分甚至缺失；有的论文的Results

部分可能与Discussion部分合并，Discussion也可能与Conclusion部分合二为一。

2）语料的高度专业化给论文部分的切分、提取带来了技术障碍。语料加工者原本是英语专业的学生，其自身的知识完成不了论文章节部分的切分。同时，科技论文并非纯语言文本，里面含有大量的学科专业领域的符号和公式，很多符号是英语语言文学专业学生不曾接触到的，有些符号、公式的删除会影响论文本内容的完整性，那么具体的符号与公式是否能删掉，文本中的某些上下标是否应该恢复，抑或可以删掉等问题的解决需要应用专业学科知识进行识别、判断与处理。

鉴于此，我们把子语料库的建设任务以课后作业的形式分配给参与华农学术英语写作课程学习的60多名博士生，他们来自于植科、动科等不同农科专业，每人分得10篇论文，完成对论文的标记、论文各部分的切分提取以及语料的人工校对。华农60多名博士生经过两个星期的共同努力，在全文语料库建设基础上，我们完成了摘要、引言、方法、结论等6个子语料库的建设。

4. 在线检索平台的建设及在教学中的初步应用

农科学术英语论文语料库建成后，上传到华中农业大学HZAU CQPweb平台（http://211.69.132.28/）。随后，在2个博士班和2个本科生A班（英语成绩优异者组成的班级）的写作教学中开展了语料库应用的教学实验。4个班共124人通过给定的账户与密码登录HZAU CQPweb使用该平台。

语料库检索培训未在写作课程学习中单独增加学时。在QQ学习群上，教师上传了语料库CQPweb检索手册和常见问题及解答，供学生自学，然后布置了语料库检索练习的课后作业。检索练习的设计遵循从易到难、由简入繁的原则，从单个词的检索到短语的搭配、句型的提取，从单库检索到跨库检索，从简单检索到复杂检索。对于复杂检索练习题，我们给予了检索表达式进行提示。在检索作业完成期间，两名教师24小时在QQ群提供检索技术咨询，在线实时解答学生关于语料库检索的各种提问。老师鼓励学生在线提问，并将每周的语料库检索提问与答案收集整理，放到QQ群中与同学们分享。经过4次循序渐进的语料库检索练习，学生基本掌握了语料库检索技术。在此基础上，结合实际写作任务，要求同学们就写作过程中实际遇到的语言困惑，自己提问并通过语料库检索，找到问题的答案。

5. 语料库使用情况的书面访谈反馈信息

语料库应用于写作教学经历了一个学期的教学实验，课程结束时我们对语料库的应用情况进行了书面访谈。访谈围绕“语料库使用的困难与收获”、“对语料库的认识”、“语料库的局限”、“参与语料库创建的感受”4个问题进行了提问。反

馈信息表明：绝大多数学生对语料库在外语教学中的作用持肯定态度。他们认为写作过程中应用语料库有利于英语写作质量的提高，通过语料库检索及其结果分析，他们能够为某些语言困惑自己探求答案。因此语料库的应用有利于提高学生的英语自主学习能力，有利于培养学生发现问题、分析问题、解决问题的能力。跨库检索的对比研究有利于培养学生的批判性思维能力，提高其对语言使用的敏感度。鉴于此，很多学生明确表示在今后实际写作中他们愿意利用语料库这一工具与资源，提高写作质量。

书面反馈也暴露出语料库建设与使用中的一些问题。其中最突出的两个问题分别是：1）现有语料库库容量不够大，农科方向某些专业领域的论文在语料库中未有涉及，影响了语料的代表性，导致某些专业表达在语料库中不能检索到结果；2）语料库检索表达式的编写过于复杂，检索界面不够友好，语料库检索的学习与使用对新手提出了挑战，他们希望语料库的检索能够像Google和百度搜索一样方便。

同时，调查对象对语料库的建设与完善提出了以下建议：1）语料资源须充实。语料库及其子语料库的专业方向须细化，以求语料涵盖面更广、更具代表性。有同学甚至建议教师传授语料库建库流程，以便学生自己下载本专业领域的语料，建设专业领域小型语料库或某个目标期刊论文的语料库，满足个性化语料检索的需求。大部分同学表示愿意参与语料库建设，包括提供专业语料和进行语料加工。2）在语料分类方面，他们建议根据期刊的影响因子的分值范围进行分类，以满足用户对不同档次论文发表的检索之需。3）在检索技术培训方面，调查对象建议：编写更简便易用的CQPweb操作手册；建立网络讨论平台，便于交流互动；开设语料库检索技术培训课程。

6. 结语

本文探讨了农科英语语料库建设的必要性，提出农科学术英语论文语料库的建设及其在学术英语教学中的应用将有利于学术英语写作质量与水平的提高。本文详细介绍了农科英语语料库的建设流程，介绍了两种文本加工的方法。方法一：利用Adobe Acrobat和DOC to TXT两个软件实现从PDF到DOC，再到TXT的两次文本格式转化法，此种方法繁琐耗时，但是具有广普适用性。方法二：利用学术文献的浏览网页，直接复制文本黏贴到TXT文本中，一次性实现从PDF到TXT的格式转换，此法高效省时、出错率低，特别适合已公开发表的学术文本的加工处理。语料库建成后，上传到基于校园网的CQPweb系统，尝试将语料库应用于写作教学。

语料库应用的效果调查表明：经过CQPweb检索手册的自学和多次语料库检索练习，实验对象基本能掌握语料库检索技术，从而解答写作中的部分语言困惑。

调查对象认为语料库有益于写作质量的提高，明确表达了在今后实际写作中将应用语料库的意愿。同时，书面访谈暴露出现有语料库资源仍不够丰富，语料库培训需由专门人员在网络计算机教室进行演示，安排专门课时上机操作。由于语料库研制有一定的技术门槛，可以调动有技术能力的学生参与语料库建设。通过语料库检索技术的学习和应用，学生意识到语料库的价值，他们表示愿意提供专业语料文本并参与语料加工。

参考文献

Hyland, K. 2008. Genre and academic writing in the disciplines [J]. *Language Teaching* 41(4): 543-562.

Johns, T. 1991. Should you be persuaded—Two examples of data-driven learning materials [J]. *English Language Research Journal* 4(1): 1-16.

Leech, G. 1993. Corpus annotation schemes [J]. *Literary and Linguistic Computing* 8(4): 275-281.

Leech, G. 1997. Teaching and language corpora: A convergence [A]. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds.). *Teaching and Language Corpora* [C]. London: Longman. 1-23.

Sinclair, J. 2003. *Reading Concordances* [M]. London: Longman.

Swales J. 1990. *Genre Analysis: English in Academic and Research Settings* [M]. Cambridge：CUP.

蔡少莲，2008，基于语料库的英语写作教学实证研究 [J],《外语教学》(6)：61-68。

曹合建，2008,《基于语料库的商务英语研究》[M]。北京：对外经济贸易大学出版社。

董爱华，2013，专门用途语料库的建设、应用、问题与发展趋势 [J],《北京印刷学院学报》(5)：59-62。

范晶晶、李丽霞，2014，农业学术英语语料库建设构想 [J],《安徽农业科学》(7)：2169-2170。

桂诗春、冯志伟、杨惠中、何安平、卫乃兴、李文中、梁茂成，2010，语料库语言学与中国外语教学 [J],《现代外语》(4)：419-426。

何安平，2010,《语料库辅助英语教学入门》[M]。北京：外语教学与研究出版社。

黄大网、秦 羿、徐赛颖，2010，专门用途英语语料库：挑战、理据与愿景 [J],《宁波大学学报 (人文科学版)》(5)：48-52。

栗 娜，2015，浅析农业学术英语语料库建设思路及设想 [J],《高教学刊》(18)：261-262。

梁茂成、李文中、许家金，2010,《语料库应用教程》[M]。北京：外语教学与研究出版社。

李文中、濮建忠，2001，语料库索引在外语教学中的应用 [J],《解放军外国语学院学报》(2)：20-25。

王景怿，2015，英汉–汉英双语畜牧业小型语料库建设及相关翻译研究初探 [J],《语文学刊 · 外语教育教学》(2)：51-52。

王克非、黄立波，2008，语料库翻译学十五年 [J],《中国外语》(6)：9-14。

王克非、秦洪武，2012，英汉翻译与汉语原创历时语料库的研制 [J],《外语教学与研究》(6)：822-834。

王立非，2008，我国英语写作教学与研究的语料库语言学视角 [A]。载王立非（编），《英语写作教学与研究的中国视角》[C]。北京：外语教学与研究出版社。2-9。

王　敏、李丽霞，2014，动物科学国际期刊论文语料库的创建与应用 [J]，《安徽农业科学》（20）：6854-6856。

卫乃兴、李文中、濮建忠，2005，《语料库应用研究》[C]。上海：上海外语教育出版社。

杨惠中、黄人杰，1982，JDEST科技英语计算机语料库 [J]，《外语教学与研究》（4）：60-62。

杨永林、李　鸣，2004，一种数字化英语学习语料库及其应用 [J]，《外语电化教学》（6）：20-26。

赵　晴，2010，专门用途语料库在ESP教学中的应用 [J]，《重庆科技学院学报（社会科学版）》（19）：182-184。

通讯地址：430070　湖北武汉华中农业大学外国语学院（刘萍、刘珊）

430070　湖北武汉华中农业大学经济管理学院（黄小倩）

《语料库口译研究的垦拓》评介*

北京外国语大学 姚 斌

Francesco Straniero Sergio & Caterina Falbo (eds.). 2012. *Breaking Ground in Corpus-based Interpreting Studies*. Bern: Peter Lang. 254 pp.

1. 引言

语料库口译研究是近些年口译研究方面逐渐兴起的研究视角。Shlesinger（1998）首次提出要将语料库方法应用于口译研究，以解决口译研究的一系列核心问题。相较笔译研究，语料库方法应用于口译研究面临诸多挑战，其关键在于口译材料的获取、转写和数据加工，通常数倍或数十倍于笔译语料库建设的工作量。《语料库口译研究的垦拓》一书，汇集了7位学者在口译语料库建设方面的理论思考与实际运用。7篇论文加上导论较为完整地展现了语料库口译研究的前沿领域与研究现状。在内容上，此书既有对语料库口译研究的详尽回顾，亦有利用语料库针对具体口译问题开展的案例研究，体现了近10年来该领域的最新进展。

此书作者多为既从事学术研究、又进行口译实践的“实践–研究者”。因此无论是对不同场景下口译活动特征的理解，还是利用语料库进行口译研究的选题挖掘都表现出较强的针对性。可以说，此书不仅是对此前欧洲语料库口译研究的总结，也体现出该领域研究的未来动向。该书对国内语料库口译研究一定有其积极的借鉴价值。本文将首先简要介绍该书的内容要点，并在此基础上展望语料库口译研究的未来发展趋势。

2. 主要内容

该书有3大主题：口译语料库的现状与展望、口译语料库的建设方法及口译语料库研究案例，共8个章节（含导论）。以下分述相关章节内容。

2.1 语料库口译研究的现状与展望

该书编者Francesco Sergio和Caterina Falbo撰写了题为“通过语料库研究口译”

* 本文写作得到北京市支持中央高校共建项目青年英才计划（YETP0842）及北京外国语大学2015年基本科研业务费院系自主项目支持。

的导论，实为对语料库口译研究的回顾与展望。导论首先勾勒了从语料库语言学到语料库翻译学的一些核心问题，如语料的代表性及翻译共性等。之后作者重点探讨了口译语料库的建设要点。口译语料库创建之所以滞后于笔译语料库，是因为口译具有不同于笔译的诸多特征。作者指出了口译与笔译在认知、伦理、社会文化及意识形态层面的不同特征。正是由于这些不同特征的存在，语料库口译研究的问题意识也有别于以笔译为对象的语料库翻译学。口译活动中所特有的问题，如面子保全策略、非言语交流及译员的角色变化等，都必须融入到口译语料库的创建中。在展望未来发展时，作者指出，语料库口译研究的最大挑战仍是数据的可比性。目前对特定语料库的研究已经产出了一些成果，但如何实现跨语料库的协同研究和共性研究仍有待于加强研究变量和分析方法的互通性。要实现研究的“生态效力”，不仅要求确保数据的获取和数量，而且要求方法论的统一及研究结果的可比性。

2.2　口译语料库的创建方法

共有5篇论文有关口译语料库的创建，它们分别涉及EPIC、DIRSI-C、CorIT、FOOTIE口译语料库和法庭听证语料库，其中法庭听证语料库主要介绍的是语料的获取方法。

Mariachira Russo, Claudio Bendazzoli, Annalisa Sandrelli和Nicoletta Spinolo介绍了EPIC的研制进展。创建于2004年的EPIC，其主要目标是收集大量真实同传语料，并将实证研究的成果用于改善口译培训。对该语料库的建设方法国内已有学者进行过较详细的介绍（如王克非、黄立波 2012），这里不再赘述。基于该语料库已经开展的研究有口译译文中的词汇密度（lexical density）和词汇变化（lexical variety）、口译中的非流利现象、演讲原文主题、语速及方式对译员表现的影响及文本处理模式等。另外该语料库也给口译博士论文的选题与撰写带来了莫大便利。在EPIC的最新进展方面，除了语料数量的扩增，作者还提及通过SpeechIndexer和Transana 2.41等工具对文本/音频以及文本/视频所作的对齐工作。

Claudio Bendazzoli着重介绍了DIRSI-C口译语料库。该语料库收集了3次国际医学会议的意英同传资料。语料库基于EPIC的理论与方法框架，但由于语料来源与前者完全不同，因而具有自身特征。Claudio重点描述了该语料库所收语料的特征及语料库创建的具体方法。该语料库有四个子库，分别是意大利语和英语的原语库和译语库，因此可以同时被用作可比和平行语料库。目前语料库的规模是136,000词。Claudio借用“交际理论”，从会议结构、演讲事件、会议参与方等角度详尽分析了国际会议的“交际情境”，并在此基础上举例说明DIRSI-C文本头文件的结构与内容。Claudio研究的主要目的在于剖析“译员为媒介的国际会议”的基本构成要素，并将之转化为可以用语料库技术分析的语料。

Annalisa Sandrelli介绍的是2008年欧洲足球锦标赛期间每项赛事前后所有新闻发布会同传语料的FOOTIE口译语料库。作者本人也是参与赛事活动的同传译员之一。建设该语料库的目标是：（1）建设足球赛事中以译者为媒介的新闻发布会语料库；（2）分析足球赛事条件下的对话沟通要素；（3）描述同传译员如何处理口译过程中遇到的挑战。该语料库与EPIC和DIRSI-C都不相同，因为它是对话性质的，而非演讲性质。该语料库包含16场新闻发布会，涉及意、英、法、西等四种语言，同时可用于可比和平行语料库研究。

作者首先介绍了足球翻译的特点，描述了足球赛事译员的工作场景、地点及形式等，随后具体介绍了2008年欧锦赛的翻译情况，之后是对FOOTIE语料库数据获取、转写等具体情况的介绍。她特别提到，虽然译者本人参与翻译有可能导致研究中的偏见，但这仍是目前获得真实会议数据的最佳途径之一。在转写方面，她特别强调，转写的语料特征应根据研究目的而定，避免标注不相关的特征。为确保跨语料库的衔接与协同研究，采用了EPIC的转写方法，但未作词性标注，因为研究目的是沟通以及特定的词汇模式。由于其对话特征，因此采用了特殊的转写方式。转写时生成两个文件，一个是不含任何标注的txt版本，可以用于WordSmith工具。另一个是Excel版本，其中一页是新闻发布会的文本头，另外一页为两种（或三种）语言的平行对照版本。文本头记录了所有与特定新闻发布会相关的信息。该文最后一部分，作者从“机构性沟通”[1]的角度分析了新闻发布会的沟通特征，包括其语言特点、参与者角色、权力关系、视角、议题选择等。作者还指出，未来的研究方向是译员在面对语速挑战时采用的口译策略。

Caterina Falbo介绍的是CorIT（意大利电视口译语料库）的分类标准。CorIT语料库主要收集的是意大利国有电视台RAI自1960年代以来的全部口译语料，同时还加入了来自其他商业电视台的语料，时间跨度大、收集内容多。该库时间跨度达50年之久，包括2700多场口译，为多种外语译入意大利语的交传和同传。该语料库被定义为开放式、多媒体平行语料库，并不断更新。该库建设的重点是：研究CorIT的主要特征及其对“语料库”、“分类标准”等既有概念重新定义的影响，以及如何重新定义研究中的核心概念，如口译模式、互动类型以及电视特征等。

本章主要了介绍该语料库的分类方法。首先介绍了文本头设计。该库文本头设计非常细致，包括：译员姓名、口译模式、互动类型、交际事件参与方姓名、日期、意大利节目名称及播放频道、外国电视台频道、语体风格、电视风格及文本类型（原文/译文）。具体到电视口译的特征，作者指出电视交传更类似于联络口译，译员一般是不带纸笔的，不记笔记，因此发言人的讲话也会比较简明。而在电视同传中，译员往往与讲话人不在同一现场，甚至有时与现场沟通无关。另外作者还提到话轮转换在源语和译语之间的差异，并通过戛纳电影节上多人发言

变为译员两人之间的顺序翻译这一案例说明，电视语体风格的分类结构不再固定，而是灵活和开放的，可以不断更新。在转写方面，作者强调应保持文本文件与声音/视频文件之间的永久性联系，建议使用的工具为winpitch。CorIT采用正字法的转写方式，去除了所有标点，以显示与书面文本的不同。可以用它开展翻译、互动及语言层面的各种研究。

Marta Biagini主要介绍了法庭口译的数据收集。他从对话口译的特点谈起，提到数据的收集并非中立，而是与研究者的研究目标相关。他研究的是在法庭这样一个高度形式化和仪式化的场景中，以译员为媒介的互动行为如何发生并产生什么影响。他发现在对译员角色的规定与译员实际行为之间存在差距。在简要介绍意大利的法庭口译情况后，作者主要介绍了数据收集的情况。一种是法庭提供影像资料，可能不符合研究者要求；一种是研究者自己现场录制，但存在审讯时间很长、高品质录像设备昂贵等问题。作者现已收集6名译员9小时法语–意大利语的对话口译视频。他指出今后的研究目标应为：（1）从语料中观察到的行为能否被普遍化，并成为规范，还应只是被视为个人行为？（2）有没有可能思考对“中立”概念的重构，分析所谓的“中立性的影响”？

2.3 口译语料库研究案例

本书有两个章节是应用语料库进行口译研究的实际案例，分别研究了“主题连贯性”（topical coherence）和“译员风格”问题。

Eugenia Dal Fovo呈现了一个利用电视口译语料库研究主题连贯性的案例。这是她正在进行的博士研究。利用的是1988-2004年美国总统大选辩论的口译版本，共计640分钟。其研究目的是探讨译员如何处理对话型及问答结构的原文，及其对译员的影响。

她首先介绍了硕士论文的研究。她将原文中出现的问题分为4大类：（1）Yes/No问题；（2）特殊问句问题；（3）引导性问题；（4）陈述性问题。她对每一种问题在语料库中出现的频次及其译文实现的连贯性程度都作出量化统计。译文分析主要从问题类型、译文连贯性程度、译员采用策略、原文结构保留程度及译文具体内容这几项入手。

她提出将在博士研究中就文本语言学和互动与对话分析等方法论问题进行更加深入的探讨，希望能够解决对话题连贯性的定义问题。由于电视口译的特殊性，译员不在现场，现场沟通也不依赖于译员，因此对于话题连贯性也需有新的定义。她将聚焦于在听众并不知道原文的情况下译员对译文的话题重建问题。

Francesco Straniero Sergio利用CorIT语料库进行了对译员风格的探索性研究。他指出，与利用笔译语料库研究译者风格相比，利用语料库研究口译员风格的研

究极少，但较之口译研究中常见的定性研究，基于语料库的译员风格研究更符合描写性译学的原则。他利用前述CorIT电视口译语料库从词汇选择、语言使用、话语标志及听说时间差的角度研究了几名意大利译员的传译风格，通过大量实例证明译员在从事电视口译时的确存在独特风格，例如常用语汇、比喻、结构及语域等。

3. 简评

口译语料库的建设与研究一向被学术界视为畏途。很多研究与其说是对语料库的研究，倒不如说是对口译语料库建设困难的研究。在我国，对于口译语料库的语料加工、对齐级别等问题探讨较多，而对口译活动自身的特点研究较少。其原因可能是多数从事相关研究者并非口译的实际从业者。

本文集的特点之一便是多数研究者均为口译从业者，这既使得他们可以较为便利地获取研究数据，又使得他们能够将自身实践中的问题带入研究中，从而在很大程度上跨越了实践与理论的鸿沟。本文集的另一特点是非常强调跨库研究及方法论的一致性。这一点值得推荐。目前不少语料库研究往往各行其道，不能实现跨库共享，缺少像文集中几位意大利学者这样的协同协作研究。例如DIRSI-C就明确采用了EPIC所使用的分类方法，以便实现研究成果的可比性。

正如本文集中不少学者指出的那样，口译语料库研究的难点在于口译语料发生场景的多样性和复杂性。因此研究中首先要对“以译员为媒介”的交际事件特点进行分析，考察特定场景下的交际事件特点。例如，文集中涉及的EPIC（欧洲议会）、DIRSI-C（医学会议）、FOOTIE（欧洲足球赛事）、CorIT（电视口译）等都是不同场景下的口译语料，因此具有不同特征。我们注意到，文集作者均不惜笔墨地分析特定场景下交际事件的特点。而对此，国内相关研究作得还比较少。国内研究更多关注点是在语料收集、语料加工和标注及检索工具开发上。然而，正如文集多位作者所强调的那样，口译语料库的建设及语料的加工标注，均应以语料库的研究目的而定，因此在建设口译语料库及开展基于语料库的口译研究时，首先需要详尽分析的是口译语料的发生场景及其特征。在这方面，需要跨学科理论，如言语行为理论、交际理论等的适时介入。

如前所述，本文集除介绍语料库构建过程，包括数据收集方法等，也展示了基于语料库进行口译译文及译员研究分析的实例。这两项研究都利用了CorIT，即意大利电视媒体口译语料库。该语料库时间跨度大、涉及译员多，提供了不可多得的研究数据。两项研究都体现了量化与质性研究相结合的特征，尤其是译者风格研究。在语料库翻译学中，译者风格研究已经比较普遍，但在语料库口译研究中尚属罕见。可以说，这两项研究为未来基于语料库的口译研究提供了方法和操

作层面的参考方案。

本文集还提出了对一些核心概念的再认识，这是由口译交际的场景特殊性所决定的。例如，电视口译的特点就使得“以译员为媒介的交际事件”有了新的定义。通常情况下，译员是交际双方之间沟通的媒介。然而，在电视口译中，译员往往不在现场，而只是针对收看电视节目的本国听众。文集作者对译员的此种角色定位进行了重新定义。又如，在法庭口译数据的收集过程中，研究者发现“中立”这一口译职业道德中的核心概念在口译规范与口译实践之间存在差距，由此尝试对口译中的这个核心问题进行语料库的探索。

从文集内容来看，对比国外口译语料库发展情况，国内语料库的语料来源比较单一、加工方法与层次比较初级（张威 2012）。仅就语料来源而言，国内目前的口译语料库主要有学习者语料库、记者招待会语料库和英专口译考试语料库，新近还出现了小型的电视口译语料库。以上几种语料库虽有其个性特征，也值得研究，但从使用范围和频率来看，占据目前口译市场90%以上比例的会议口译却几乎没有成规模的语料库。记者招待会语料库的场合比较特殊，译员身份也有特殊性，其语料的代表性不足。而国内电视口译语料库与国外也有差异，国内电视口译员几乎没有专职者，都是临时为之，他们的口译策略和风格，应该是在会议口译实践中养成的，尚未形成独特的电视口译特征。对口译市场的核心群体，即职业会议口译员的语料库建设也才刚刚起步。

来自文集的另一启示是，由于实际会议存在多种类型，如金融、环保、医疗、专业技术等，而各种类型会议的口译又各有特征，因此都有建设语料库进行研究的必要。但在语料库建设过程中，首要问题应该是对这些特定场景下的交际行为进行深入细致的分析，同时还要注意到跨库研究的可行性和协调性。

本文集中的语料库建设与研究也还存在一些值得进一步推进的地方。例如，不少研究还处于未完成状态。其实口译语料库的研究大有用武之地。对于从事译员培养的人来说，最关心的还是语料库研究可以给口译教学带来什么。对于口译学习者，在学习过程中，关注的是有哪些可以从语料库中得到的经验。未来值得探索的问题还有，在译员口译策略的习得和口译风格的形成方面，语料库研究究竟可以带来什么样的启示？对比职业译员语料库和学习者语料库，结合专家–新手研究又能呈现出怎样的差异化特征？语料库研究如何做到描写与解释互补，通过将量化和质性研究相结合的方法，跨越理论与实践的鸿沟，构建一种口译研究的新范式？这些都是我们心中的问题，有待语料库口译研究的进一步垦拓，并给出答案。

注释

1. Galatolo（2002:137）将“机构性沟通”定义为“在沟通中至少有一方通过特定的沟通行为扮演了机构性的角色”。

参考文献

Shlesinger, M. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies [J]. *Meta* 43(4): 486-493.

王克非、黄立波，2012，国外双语库研制与应用评析 [J],《外语电化教学》(6)：3-10。

黄立波，2013,《语料库翻译学：研究与应用》评介 [J],《外语教学与研究》(4)：623-628。

张　威，2012，近十年来口译语料库研究现状及发展趋势 [J],《浙江大学学报（人文社会科学版）》(2)：193-205。

通讯地址：100089　北京市北京外国语大学高级翻译学院

第三届亚太语料库语言学大会征文通知

第三届亚太语料库语言学（The 3rd Asia Pacific Corpus Linguistics Conference）暨第三届中国语料库语言学大会（The 3rd Corpus Linguistics in China Conference）大会将于2016年10月21日至23日在北京航空航天大学举行。

会议将邀请国内外语料库语言学知名学者做主旨演讲。预计将有数十个国家和地区的语料库语言学研究人员出席宣读研究成果。

我们热忱欢迎国内语料库语言学同行参会宣读论文、切磋交流、建立友谊、促成合作，共同推进学科事业发展。

请大家就下列主题或相关话题准备论文：

1. 语料库语言学新技术、新方法
2. 语料库语言学新视野、新方向
3. 语料库与语言对比及翻译研究：肖忠华教授纪念专题
4. 语料库与话语研究
5. 语料库与二语习得及外语教学研究
6. 语料库与词典编纂
7. 其他语料库相关话题

有关大会的最新动态可关注大会网站（http://fld.buaa.edu.cn/Science/View.aspx?id=2194）。

第三届亚太语料库语言学大会组委会

2015年12月

English Abstracts

Some reflections on Corpus Linguistics upon request

Dr. Richard Xiao is Reader at the Department of Linguistics and English Language, Lancaster University, United Kingdom. His major research interests cover corpus-based language studies, contrastive linguistics, translation studies, Chinese linguistics, English language and linguistics, tense and aspect theory, and teaching Chinese as a second language. It is with great pleasure that Dr. Xiao has completed a written interview with our journal around such topics as his views on the recent advances in corpus research home and abroad as well as his personal academic career.

Liang Maocheng's views on corpus research and computer technology

Professor Liang is Professor of corpus linguistics and applied linguistics at the National Research Centre for Foreign Language Education, Beijing Foreign Studies University, China. His research interests include computational linguistics, corpus linguistics, automated essay scoring, corpus-based interlanguage analysis, and second language acquisition. He is vice Chairman and a founding member of Corpus Linguistics Society of China. In this interview, Liang writes about, around eight topics suggested by the Journal, the impact of computer technology on corpus-based language studies, and looks into the prospect of the field in the age of big data.

Xing Fukun's views on corpus research and computer technology

Dr. Xing is affiliated to the PLA University of Foreign Languages, China. His major research interests include computational linguistics and corpus linguistics. In this interview, Xing addresses too the eight topics suggested by the Journal concerning the impact of computer technology on corpus based language studies, and how the technology in the time of big data will shape the future of corpus studies.

Demystifying Zipf's law and Zipfian linguistic economy theory

Zipf argues that, because the speaker's and the auditor's behaviors are both governed by the principle of least effort, the speaker's and the auditor's economies are in conflict, a speaker-based force of unification is in opposition with an auditor-based force of diversification, and the two forces manufactures in speech stream a balanced vocabulary distribution, which is termed by Zipf himself rank-frequency distribution and afterward famously known as Zipf's law. According to relevant mathematical studies which have established consensus, Zipf's law is not a product of the principle of least effort. Therefore Zipfian theory of linguistic economy is starved of its vitally important evidence. Moreover and under linguistically rational scrutiny, Zipf's theory on the two forces and the two-sided economies is far from infallible. This paper intends to develop an intuitive exposition that the Zipf's law is not a product of the principle of least effort and to undertake a demystifying and critical review of Zipfian theory of linguistic economy.

A study of the collocational behaviour of Chinese time words: *Nian* 'year', *yue* 'month' and *tian* 'day'

This paper attempts to investigate the time words *nian* 'year', *yue* 'month', and *tian* 'day' collocated with numbers on the basis of corpus data. The previously syntax-based studies are updated with lexical frequency and collocational behaviour of the words. The word *nian* 'year' has significantly higher frequency than *yue* 'month' and *tian* 'day'; likewise, *nian* 'year' has nine meanings different from *yue* 'month' and *tian* 'day'. The nominal classifier *ge* has to be added to *yue* 'month' when it is used as in cardinal number contexts. The multi-word units *youyitian* 'one day or someday' and *diertian* 'the second day' are interpreted against their pragmatic and textual functions. The paper concludes that among a few high frequency words of a lexical field, sometimes only partial homogeneity in terms of collocational, syntactic and lexico-semantic properties is shared.

Entity, attribute and relation: The trichotomy of shell nouns and their interpersonal functions

Shell nouns metaphorically refer to a type of nouns which carry and represent propositional information, a key feature of lexical cohesion in discourse. However its affordances of stance construction are less noticed than deserved, which is caused at least by unclear semantic

classification and interpersonal function of shell nouns. This paper is based on a corpus of 60 research articles and comes up with a new function-based classification of shell nouns concerning entity, attribute and relation in an attempt to explore how shell nouns allow stance construction and social interaction along discourse community in the disciplines.

Idioms and idiomaticity in translational Chinese: Translation Universals hypotheses revisited

The present study intends to explore the idiomatic properties of idioms in translated Chinese texts (i.e. ZCTC) to verify and revisit Baker's Translation Universals hypotheses from the perspective of Chinese. As a follow-up study of Xiao and Dai (2010) on word clusters in translated Chinese, the present paper manually filtered the annotated word clusters in the two corpora into idioms based on *Xinhua Chengyu Cidian* (2002) dictionary and examined the total counts of token, type, part-of-speech distribution, high-frequency idioms as well as semantic and structural properties of idioms in the two corpora. It is observed that compared with their native counterparts (i.e. as in LCMC), idioms in translated Chinese are less in total amount, more diversified in type and exhibit a polarisation of high-frequency and low-frequency idioms. Their meaning tends to be more literal and explicit and their structure demonstrates more fixedness, revealing translators' prudence in conforming to the standard norms of the target language. These results reconfirm the hypothesis of explicitation, standardisation and leveling out in the Translation Universals theory, yet leaving the simplification hypothesis largely unmatched with Chinese. Thus, the present paper revises this simplification hypothesis in the TUs theory to a polarisation hypothesis, hoping to better account for the perplexing phenomenon in idiomatic usage in translated Chinese.

Lexical bundles in China-based English journal articles of science and engineering

The present paper, in conformity with Biber and his colleagues' framework, sets out to describe and discuss the structures and functions of four-word lexical bundles in China-based English journal articles of science and engineering. The target bundles are investigated in terms of structural categories, colligation of VP-based bundles, co-occurrence of PP-based and NP-based bundles, AP-based bundles, functional categories and their pragmatic connotations. A further analysis on the relations between structures and functions has revealed the distribution pattern of bundles. These findings have implications for studying Chinese scientists' and engineers' language competence in academic research as well as improving academic writing pedagogy.

Constructing an agricultural research article corpus of English

LIU Ping, HUANG Xiaoqian & LIU Shan (97)

This paper discusses the necessity of agricultural academic English corpus construction, and introduces the procedures of construction, which include the collection of data, the conversion and cleaning of texts, annotation, tagging, and the extraction of texts. It compares two text cleaning methods. The agricultural academic English corpus was mounted to Hua Zhong Agricultural University (HZAU) intranet using the CQPweb (Corpus Query Processor) system. HZAU CQPweb was tried out in the Ph.D. students' and undergraduates' academic writing courses. The written feedback of the trial indicates that most students acknowledge that corpus-assisted learning helps improve English academic writing quality and claim that they will use corpora in writing up their academic papers. In the meanwhile, the feedback also shows that the existing corpus resources need to be enriched, and that corpus search is still too complicated and not user-friendly enough.