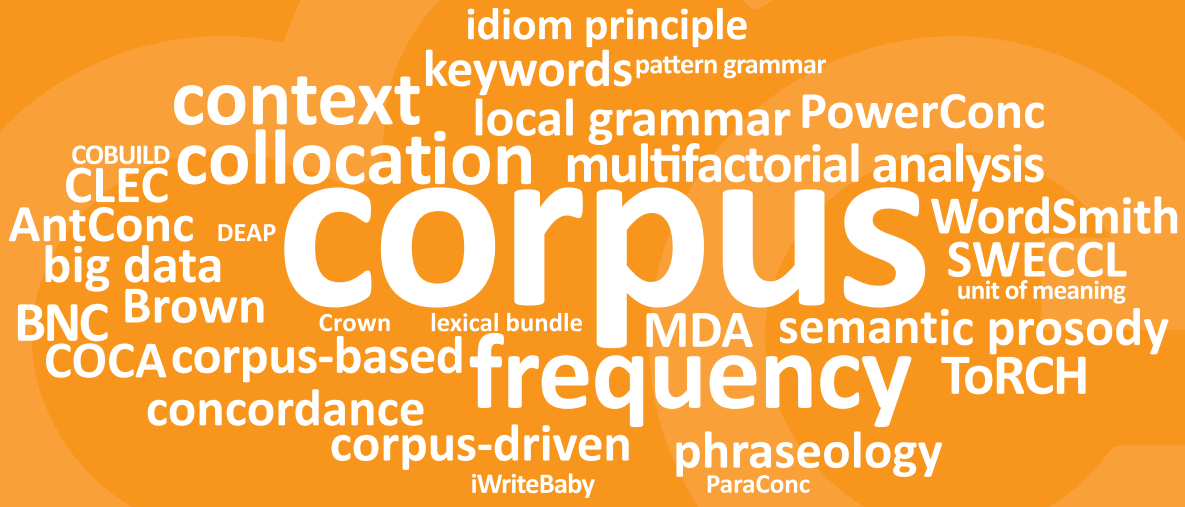
《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

语料库语言学

CORPUS LINGUISTICS

第12辑
2019

北京外国语大学中国外语与教育研究中心
中国英汉语比较研究会语料库语言学专业委员会
许家金　主编



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：中国外语与教育研究中心
　　中国英汉语比较研究会
　　语料库语言学专业委员会
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 and Corpus Linguistics Society of China, China Association for Comparative Studies of English and Chinese
Published by Foreign Language Teaching and Research Press

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

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

《语料库语言学》
2019年　第6卷　第2期

目　录

同题共议

研究论文

研制开发

书刊评介

CORPUS LINGUISTICS

Volume 6, Number 2, 2019

Table of Contents

王立非谈语料库与ESP研究

北京语言大学 王立非

编者按：

本刊自2014年第2期推出“同题共议”栏目。每期由《语料库语言学》提出一定的选题，邀请两位从事语料库语言学相关研究的学者，就共同约定的议题进行笔谈。两人各自阐述，互不交流，以此保证观点的独立性。本栏目鼓励和而不同的学术争鸣。学术思潮诸流并进，方才是兴盛的气象。

本期“同题共议”立意语料库语言学在ESP/EAP中的应用，由王立非、姜峰两位学者就此议题陈述各自观点。本刊不作评点，亦不选择立场。本刊欢迎诸位同仁就同样的议题来稿研讨。

1. 在您印象中，国内外较早开展的ESP语料库研究有哪些？您如何评价这些研究？

（1）ESP研究的四个发展阶段

专门用途英语（English for Specific Purposes，简称ESP）是应用语言学的一个重要分支领域，主要包括学术英语（English for Academic Purposes，简称EAP）、专业英语（English for Professional Purposes，简称EPP）和职业英语（English for Occupational Purposes，简称EOP）三类（Belcher *et al.* 2011）。ESP产生于20世纪60年代，大致经历了体裁分析、语篇修辞分析、目标情景分析及以学习者需求为中心等几个重要发展阶段。第一阶段为初创期，大致为1962—1981年，以John Swales等为代表的学者专注于体裁分析，总结分析了科技文章中的词汇、句法和语篇特征；第二阶段为发展壮大期，大致为1981—1990年，介绍了ESP广受关注的一些核心概念，比如目标情景分析、学习者需求分析、体裁分析、修辞语步分析等。Hutchinson & Waters（1987）等分别提出了目标情景分析和学习者需求等ESP的核心概念。第三阶段为学术高涨期，大致为1990—2011年，创立了ESP学术期刊*English for Specific Purposes*（《专门用途英语》）和*Journal of English for Academic Purposes*（《学术英语学刊》），也出现了一些LSP研究的成果等；第四阶段为相对成熟期，大致为2011年至今，国外的ESP研究逐步成熟和深入，重点从语言学、语用学、心理认知、语料库等不同视角对学术英语、ESP行业英语和职业语言特

征描写和ESP教学与应用。

在国内专门用途外语领域，从2007—2017年的10年间，以第一作者身份发表过论文的作者共有364位。对专门用途外语发文活跃度综合指数统计显示，复旦大学（蔡基刚、陈宁阳 2013；蔡基刚 2015）在学术英语研究方面十分活跃，发表论文27篇，总被引频次高达1,450次。复旦大学专门用途外语研究领域活跃度在全国领先，综合指数、总被引次数、总下载量、篇均下载量均最高。此外，北京外国语大学、上海外国语大学、对外经济贸易大学、上海交通大学等几所高校也在全国位于前列。

根据我对ESP的研究，ESP学科体系尚未完全建立，而是一种教学理念和教学方法。目标是专门针对学科专业和职业需要，开设英语课程，帮助学生熟练掌握专业领域有效沟通的英语技能，ESP特征鲜明，面向学生需求，与学科、职业及专业相关，专业语篇与体裁特点鲜明，教材和学习方法不同，教学理念和方法不同。本文为了和学术英语（EAP）区别开来，专门探讨除EAP以外的ESP，因此，这里的ESP语料库研究专指学科英语、行业英语和职业英语研究，不包含EAP。

（2）国外ESP语料库研究

纵观ESP研究，呈现出的特点是EAP研究要远远多于ESP研究，以J. Swales和K. Hyland等人为代表的英国学派主要关注学术英语体裁和教学。2004年，剑桥大学出版社出版的J. Swales著作*Research Genres: Exploration and Applications*探讨学术体裁与应用，涉及教材、讲座、授课等体裁，引起广泛关注。该书奠定了ESP研究的重要概念、基础知识边界及研究对象，引用率极高，至今体裁理论仍是国际ESP研究最重要的理论基础或研究范式。2000年，K. Hyland 出版著作*Disciplinary Discourses: Social Interactions in Academic Writing*，由J. Swales作序，提出了“学科话语”（Disciplinary Discourse）的概念，研究了学术共同体文化与ESP话语之间的关系，并指出学科话语理论是作者与读者相互理解的基本框架。以上两部著作堪称ESP研究的奠基之作。另一位学科英语的代表性人物是V. Bhatia，他早期关注职场和商务话语，近年来转向法律话语研究。他2006年发表了“Discursive practices in disciplinary and professional context”一文，指出ESP研究只分析文本特点不够，还应该与行业、职业和企业的语境与话语惯例相结合，话语和体裁分析与组织分析和职场沟通相结合。

纵观ESP语料库研究，海外较早的代表人物为Lynne Flowerdew。他1995年在《专门用途英语》期刊发表了题为“Designing CALL courseware for an ESP situation: A report on a case study”的论文，以香港理工大学本科生和研究生的求职技能课程为案例，通过需求调查，提出语言和学习为中心的课程设计思路，探讨如何根据学生需求和求职需要，将ESP方法应用于计算机辅助外语教学课件设计。

（3）国内ESP语料库研究

国内基于语料库的ESP研究起步较晚，开始于2000年。从CNKI能够检索到的最早一篇论文是发表在《外语研究》上，题为《ESP与语料库建设》（陈明瑶2000）。该文论述了ESP语料库的作用（谈不上是研究），呼吁开展ESP语料库研究。作者认为，ESP语料库有助于描写专门用途英语的特点，有助于改进ESP的研究方法，有助于ESP教学，有助于语料库向专用语料库方向发展。从2000—2018年，国内一共发表ESP语料库论文33篇，其中，发表在CSSCI期刊的论文5篇，都是围绕ESP语料库在英语教学中的应用。

总体而言，我国的ESP语料库研究还处在初级阶段，主要表现在五个方面：第一，选题比较窄，只关注大学英语教学，对ESP语言特点和体裁特点的描写远远不够。第二，ESP语料库的建设还相对落后，我因主持项目需要，建设过2亿词的“中国商务英语语料库”（Business English Corpus of China）和3,000万词的“国家标准英汉文本双语语料库”。原解放军外国语学院建设过军事英语语料库。此外，全国部分院校还零星建设过法律英语或新闻英语语料库，但规模和研究成果不详，至今还没有出现包含多学科领域的大型ESP综合性语料库。第三，语料库方法在ESP研究中的应用还有待深入，除了研究词汇以外，许多研究句法、语篇结构、语用、语义的语料库软件和工具都没有得到应用。第四，多语种ESP语料库建设和研究几乎还是空白。第五，国内的ESP语料库研究主要集中在学术英语、商务英语等领域；艺术、历史、法律、新闻、理工、农林、医药等ESP领域的语料库建设和学术发表相对薄弱，主要原因是许多外语老师的专业背景是语言文学，对理、工、农、医、经、管、法等学科领域了解少，对ESP语料库建设和研究感到力不从心。

2. 请问您是如何开始对基于语料库的ESP研究产生兴趣的？怎样的时机（或者某些学者或作品）促使您走上ESP语料库研究之路？

我从上大学开始就接受ESP教育，对ESP产生了比较浓厚的兴趣。早期我学习军事外交英语，参加过联合国维和行动，实际接触到“维和英语”这种特殊的ESP英语。过去10多年来，我开始从事商务英语教学和研究，接触商务英语较多，努力开始了我个人的学术转型，从军事英语转向商务英语。商务英语是ESP最大的分支，主要研究英语作为国际商务通用语的特点、功能和规律。我通过这些年的努力，取得了一批学术成果，发表了多篇论文，撰写了专著，申报了国家社科基金课题，指导了一批博士生和硕士生。我对语料库语言学理论和方法比较熟悉，以前一直用它来研究中介语，发表过许多论文。

我尝试用语料库方法研究商务话语是因为2010年申报了国家社科基金项

目“商务话语名物化的语料库考察与研究”，获得立项，便开始了ESP语料库研究。国内对商务话语的名物化现象研究较少，名物化是语言语法化的一种现象，是反映人类认知世界的一种方式，对商务话语中的名物化现象进行考察，有助于我们进一步认识语言所反映出的经济全球化趋势，而采用商务语料库，涵盖经贸、金融、法律、新闻、演讲等各种类型，提取名物化特征进行系统的定量和定性分析，也有助于发现和总结出商务话语的名物化特点，促进商务英语教学。

我基于自建的商务英语语料库，开展了语料库商务词汇名物化和句法名物化研究，分析了商务法律话语名物化，对比了商务口头与书面话语名物化和英汉双语名物化现象，考察了我国英语学习者名物化使用的特点，先后发表了15篇论文，出版了专著《商务话语名物化研究》。通过研究发现，ESP语料库建设在商务英语研究中大有可为，许多领域的语料库还有待开发和建设，如医学、法律、金融、贸易、旅游、新闻、外交、体育、计算机、工程、石油、电力、航空、军事等多个领域。语料库对比分析方法对考察ESP体裁的特点和用法十分有效，通过ESP和BNC等通用语料库对比，我们发现了许多商务话语特性。

3. 您认为基于语料库的ESP研究的理论和实践意义体现在哪些方面?

语料库ESP研究的理论意义在于以功能语言学的理论视角描写语言（Coffin & Donohue 2012），基于语言使用和语言事实，应用定量和定性的研究范式，为我们从大规模语言数据中归纳出语言规律提供了认识论和方法论的基础。语言在各行各业的应用十分广泛，衍生出海量数据，语言特点各不相同，用乔姆斯基语言学的语言直觉，只能判断语言深层的共性特征和规则，却很难穷尽语言表层的个性特征和差异，而ESP语言个性特征千差万别，通过基于语料库的ESP研究，可以自下而上地归纳出专门用途语言与通用语言之间的互文性（intertextuality）和互语性（interdiscursivity）特征，从而更有效地运用某种专门用途语言。

ESP语料库的实践意义在于研究实践和教学实践两个方面。在研究实践方面，通过基于ESP语料库（corpus-based）和语料库驱动（corpus-driven）两种不同的方法，开展多维度的对比分析。基于语料库的方法采用一种自上而下的演绎性范式（deductive），我们对某种语言规则的直觉和假设是否正确，可以通过海量的ESP语料库得到验证，去寻找语料库中是否存在语言例证或相关用法，从而验证语言规则和用法的合法性和正确性。因此，演绎性研究范式就需要我们提出理论观点或研究问题在先，然后采用语料库依托的方法去查找范例，自上而下开展定量语言实证研究或定性案例分析。语料库驱动方法则不同，它是一种自下而上的归纳性（inductive）范式，从真实语言使用的数据出发，通过对语言事实的详细观察和描写，从中归纳和提炼出某种语言规则或特点，在观察语料数据之前，研

究者没有任何预设的问题或观点，这个过程可以概括为提取、观察、概括、解释。归纳性研究范式需要我们善于从语料库中发现语言事实和用法，总结出规则或定律。

在教学实践方面，ESP语料库为开展商务英语、法律英语、医学英语等ESP教学提供了丰富的资源和教学手段。基于语料库的ESP教学与传统的外语教学的最大区别是实现了以学生为中心的教学理念和方法，教师可以让学生围绕一个重点词汇、重点句型、语法现象等去探索性学习，学生在教师的指导下，主动和自主地去随时查阅和调用丰富的ESP语料资源，通过观察和对比无数不同例句以及中心词或词组出现的前后语境，了解搭配关系和强弱度，以及哪些词汇或用法更常用、更地道和如何用，彻底改变过去教师讲解靠单个举例的局限和不足。

我认为，ESP教学要加强词汇和体裁对比，ESP话语具有独特的学科话语、行业话语、专业话语特征。在词汇层面，ESP核心词表编制十分重要，ESP大量使用专业词汇，核心专业词汇代表核心知识本体，也是ESP英语的一大特征。以专门用途英语词汇表的编写为例，我们可以在专门用途英语语料库的基础上，借助语料库工具生成该语料库的英语词汇总表，然后计算出各词出现频率的同时，与普通用途英语语料库作对比，将出现频率高但却与普通用途英语语料库高频词汇意义差别很大的那些词汇，按照从高到低的方式挑选，最终形成专门用途英语核心词汇表，方便学生掌握和教学。在体裁风格上，ESP体裁十分丰富多样，而且差异很大，各类体裁都有自身的特点和格式，体现专业和行业话语的特征。因此，建立ESP体裁教学案例库十分必要，是大学英语教学改革的一个方向。

4．能否请您简单介绍一下当前国内外ESP语料库研究的主要热点？

基于Web of Science文献数据库，采用CiteSpace软件对国外的七本ESP和语料库相关的学术期刊进行分析，结果显示，过去10年，国外共发表ESP语料库研究的论文270篇。其中出现了10个热点领域，分别是讲座风格、体裁分析、词束、科技词汇、语步构念（step construct）、短语动词、语言描写、句子连接词、语法隐喻、立场词束，其中，最受关注的领域是讲座风格，立场词束受关注最少。可见国外ESP语料库研究热点仍以学术英语为主，其次涉及词束（lexical bundle）、程式序列语（formulaic sequence）、短语学（phraseology）、ESP学习者、二语ESP学习、ESP学习者语料库、二语ESP写作、ESP元话语、ESP立场词与态度短语、ESP模糊限制语（hedging）、ESP语篇衔接、法律英语、名物化、互文性等，这些都是目前国际ESP语料库的研究热点和趋势。根据Elsevier论文数据库检索发现，近30年来，国外发表ESP语料库实证研究论文1,417篇，主要发表在*English for Specific Purposes*、*Journal of Pragmatics*和*Procedia-Social and Behavioral Sciences*三

本期刊上，其中*English for Specific Purposes*发文最多，达303篇，其次为*Journal of Pragmatics*，发表79篇，*Procedia - Social and Behavioral Sciences*发表77篇。

我以语料库+ESP主要领域为关键词对知网文献库检索发现[1]，国内过去40年来ESP语料库研究共发表论文549篇，所有论文都是近20年发表的，可以看出，国内从2000年开始关注ESP语料库建设和研究。相比国外而言，国内ESP语料库论文总量不少，但研究水平总体不高，CSSCI期刊论文为74篇，占论文总发表量的13.4%。研究热点主要涉及以下几大类：1）各类ESP语料库建设；2）ESP语料库在教学中的应用；3）基于语料库的ESP翻译研究，如译文的显化特征等；4）基于语料库的ESP语言特征描写，包括词汇、句法结构、语篇功能、文体特征、名物化、情态动词、话语标记、语用特征等；5）基于语料库的批评话语分析，如对中西方新闻报道的批评话语分析。可以看出，国内的ESP语料库研究还有待进一步从学术英语向更广阔的专业领域拓展，聚焦法律、经济、外交、传媒、教育、医疗、旅游、交通、电信、能源、机械、电子、航空、农业等领域的行业英语研究，围绕我国高等外语教育新一轮改革提出的“一精多会，一专多能”的方向，建设ESP教学语料库，产出更多更好的研究成果，推动教育教学改革。

5. 基于语料库的ESP研究前景如何？哪些方面值得我们进一步挖掘？

我认为，ESP语料库研究具有广阔的前景，是外语教学改革的一个重要方向，具体可以关注以下几个领域：

1）跨语种和跨语类“一带一路”ESP语料库建设与应用

语料库研究前景广阔，近年来，国家社科基金和教育部社科基金的语料库立项逐年增长。下一步，我们要从服务国家战略和行业急需出发，面向“一带一路”国别语言和文化，建设跨语种和跨语类的语料库，采集“一带一路”政治、经济、法律等各领域的语料和不同类别样本，开展基于语料库的“一带一路”多语言特征研究，助力“一带一路”建设。

2）基于语料库的ESP技能教学研究

探讨学术或职场背景下ESP与听说教学的关系。总体而言，ESP书面语研究较多，建设综合性、大规模、多学科的ESP语料库成为可能。如何结合专业语境，找到ESP语篇和体裁的核心特征，将ESP读写教学融合到特定学术和职业文化中去，如在法律和商务行业背景下，提高学生的跨文化读写能力，是ESP阅读和写作教学今后努力的方向。随着互联网技术和语音识别技术的发展，ESP听说语料

库建设难度降低，今后语料库建设的重点是ESP听说教学语料库，采集更多职场和专业领域的口语语料库和ESP学习者有声语料，研究我国学生学习ESP的重点和难点，探讨听力技能发展过程中ESP学科知识与元认知开发的关系。

3）ESP语料库驱动的“新工科”英语教学与研究

教育部根据新一轮科技革命与产业变革，支撑服务创新驱动发展、“中国制造2025”等一系列国家战略，提出了“新工科”“新医科”“新农科”“新文科”建设的总体思路，对ESP教学和研究提出了新的方向和挑战。新工科指针对新兴产业的专业，以互联网和工业智能为核心，包括大数据、云计算、人工智能、区块链、虚拟现实、智能科学与技术等相关工科专业。新工科专业涵盖新型工科专业、新生工科专业、新兴工科专业[2]。随着大数据和人工智能的快速发展，新工科将成为学科建设的重点，新工科ESP教学也成为新一轮英语教学改革的重点和方向。基于新工科ESP语料库研究将会成为热点，全面加强ESP语料库的词汇短语、句法结构、语篇结构、语用特点、体裁特征、多模态特征研究是一个新的增长点。

就“新文科”语料库研究而言，可以关注数字人文、语言大数据、机器翻译、技术哲学、语言智能、区块链金融、机器人伦理等领域，建设“新文科”英语教学语料库，为课程开发、教材编写、教学资源配套、教师培训等提供专业化支持。

4）基于ESP语料库的专业核心词表研发

ESP专业词汇是掌握ESP的一个突破口，分为通用ESP词汇和专用ESP词汇。研究首先要回答的就是ESP学习者究竟需要什么词的问题。这个看似简单的问题却引发了更多的深入问题，比如学习者目标、语言水平、学习者情景以及学习者时间等。一些日常词汇在特定语境下的专业意义，比如，日常词汇monitor在计算机科学中的意义就明显不同。尽管每个领域分类或层级多样，但是ESP词汇在本质上主要是指某个特定领域的词汇，专业化或特定领域是其主要特征。确定ESP核心词表，可在时间宝贵的课堂ESP教学中，有针对性地开展教学，更好地满足大学生的英语学习需求，特别是非外语专业学生的学习需求。国外Nation（2001）等学者曾经针对学术英语开发出核心词表，我曾按照商务英语专业教学要求，组织团队研发过商务英语核心词表，并以《高等学校商务英语词汇学习手册》（王立非 2013）出版。该核心词表分为普通商务英语词汇和专业商务英语词汇两部分，普通商务英语词汇12,114个，专业商务英语词汇520个，总计12,634词，基本涵盖了商务英语学习应该掌握的常用词汇。普通词汇是学生必须掌握的常用核心词汇，专业商务英语词汇涉及经济、贸易、管理、金融、营销、法律等领域的核心专业词汇。我认为，下一步应该按照《大学英语教学指南》的总体思路，组织力

量建设ESP语料库，提取ESP英语核心词表，供ESP课程教学和学生学习参考，要完成这个艰巨的任务，还有许多工作要做。

6. 能否请您谈谈我国学者如何能作出具有本土特色的ESP语料库研究？

中国是个外语教育大国，外语教育对中国对外开放至关重要，新一轮外语教育改革的重点是服务国家战略，面向行业需求，培养国际化人才，帮助学生用外语从事专业工作的能力。我认为，ESP语料库研究要体现中国特色，面向中国实际，解决中国问题，培养中国人才，需要重点加强以下四个方面：

1）结合《大学英语教学指南》对ESP课程的要求开展语料库研究

《大学英语教学指南》明确指出，高校开设大学英语课程，一方面是满足国家战略需求，为国家改革开放和经济社会发展服务；另一方面，是满足学生专业学习、国际交流、继续深造、工作就业等方面的需要。对三类英语课程提出明确要求，专门用途英语课程以英语使用领域为指向，以增强学生运用英语进行专业和学术交流、从事工作的能力，提升学生学术和职业素养为目的，具体包括学术英语、职业英语两大课程群。专门用途英语课程将特定的学科内容与语言教学目标相结合，教学活动着重解决学生学科知识学习过程中所遇到的语言问题，以培养与专业相关的英语能力为教学重点。ESP语料库研究要根据《指南》提出的这个目标和要求，开展学术英语和职业英语语料库资源的建设。职业英语语料库建设和研究是体现本土化和学校特色的一个重点，各学校教师要根据本校的学科定位、专业特色和优势，确定重点建设的“金课”，组建团队，持续开展职业英语语料库建设和学科英语特点与体裁研究，特别是具有中国本土特色的英汉双语ESP可比语料库和平行语料库建设，为打造基于语料库的大学职业英语“金课”提供资源保障。如，法律院校要大力开发“法律英语”“法律文献翻译”“法律写作”“法庭口译”“知识产权英语”等法律英语综合语料库和教学案例库，将语料库资源和信息技术融合，建设一批高质量、大容量、有深度的“金课群”。

2）依托ESP语料库开展“一带一路”语言服务研究

语言服务是近年来兴起的一个新兴学科领域，根据中国翻译研究院和中国翻译协会（2017）的定义，语言服务指以语言能力为核心，促进跨语言、跨文化交流为目标，提供语际信息转化服务和产品，以及相关研究咨询、技术研发、工具应用、资产管理、教育培训等专业化服务的现代服务业。而语言服务研究以语言能力为核心，以语言行业研究和应用为特定目标和对象，系统研究如何将语言理论应用于行业实践，解决语言规划、语言大数据、语言智能、语言技术、语言管

理、语言培训、语言标准化、语言产业、语言服务贸易、文化外译等领域实际问题的一门学问。由此可见，“一带一路”语言服务是“一带一路”五通的基础和保障。在当前“一带一路”建设中，如何面向商务语言服务、法律语言服务、企业网站语言服务、行业标准国际化、跨境电子商务贸易服务、文化对外传播服务，建设ESP语言服务大数据库，为“一带一路”提供准确高效的政治、经济、法律、宗教、文化、媒体信息，为走进“一带一路”防范国别潜在风险，调查我国的海外影响力和接受度，维护我国的国际形象，提升对外语言服务能力，在这些方面，ESP语料库都大有可为，希望更多的外语学者和师生关注和思考。

注　释

1. 检索关键词为语料库+商务、法律、医学、工程、计算机、航空、航海、新闻传媒、军事、农林、旅游、艺术、体育、出版14类。
2. 新工科专业具体包括：物联网工程专业、光电信息科学与工程专业、计算机科学与技术专业、数字媒体技术专业、数据科学与大数据技术专业、机器人工程专业、智能电网信息工程专业、智能科学与技术专业、智能建造专业、智能制造工程专业、智能医学工程专业、航空航天工程专业、飞行器设计与工程专业、飞行器制造工程、飞行器动力工程专业、飞行器环境与生命保障工程专业、飞行器质量与可靠性专业、飞行器适航技术专业、船舶与海洋工程专业、新能源科学与工程专业、机械设计制造及其自动化专业、车辆工程专业、生物制药专业。

参考文献

Belcher, D., A. Johns & B. Paltridge (eds.). 2011. *New Directions in English for Specific Purposes Research* [C]. Michigan: The University of Michigan Press.

Bhatia, V. 2006. Discursive practices in disciplinary and professional contexts [J]. *Linguistics and the Human Sciences* 2(1): 5-28.

Coffin, C. & J. Donohue. 2012. Academic literacies and systemic functional linguistics: How do they relate? [J]. *Journal of English for Academic Purposes* 11(1): 64-75.

Flowerdew, L. 1995. Designing CALL courseware for an ESP situation: A report on a case study [J]. *English for Specific Purposes* 14(1): 19-35.

Hutchinson, T. & A. Waters. 1987. *English for Specific Purposes: A Learning-centered Approach* [M]. Cambridge: CUP.

Hyland, K. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing* [M]. London: Longman.

Nation, I. 2001. *Learning Vocabulary in Another Language* [M]. Cambridge: CUP.

Swales, J. 2004. *Research Genres: Explorations and Applications* [M]. Cambridge: CUP.

蔡基刚、陈宁阳，2013，高等教育国际化背景下的专门用途英语需求分析 [J],《外语电化教学》(5)：3-9。

蔡基刚，2015，中国专门用途英语教学发展回顾、问题和任务 [J],《西安外国语大学学报》（1）：68-72。

陈明瑶，2000，ESP与语料库建设 [J],《外语研究》(2)：60-61。

王立非，2013,《高等学校商务英语词汇学习手册》[M]。北京：高等教育出版社。

通信地址： 100083 北京市北京语言大学高级翻译学院

姜峰谈语料库与EAP研究

吉林大学 姜 峰

1. 在您印象中，国内外较早开展的EAP语料库研究有哪些？您如何评价这些研究？

EAP，全称为English for Academic Purposes，即学术英语或学术用途英语，泛指学术语境下开展各类学习和交流所需的英语。EAP是ESP（English for Specific Purposes，特殊用途英语）的重要分支，起源于20世纪60年代。当时，英国文化委员会（British Council）资助大量发展中国家学生赴英国学习，短期培训课程增加。以Michael Halliday、Angus McIntosh和Peter Strevens为代表的学者们开始探索传统英语教学中出现的新需求，提出此时语言学的任务是分析基于特定人群和特定语境下的语言实例，包括学术交流在内的特殊用途语言特征。根 据Swales（1985），Charles Barber（1962）“Some measurable characteristics of modern scientific prose”标志着学术英语研究的开端。Barber基于两万三千余词的三篇科技文本，分析科技语篇的句子结构、动词形式以及词汇特征，发现28%的实义动词用于被动语态，而被动语态的最常见形式是现在时，约占25%。可见，EAP研究自起步阶段便与语料文本和语言观察结合在一起。正如Hyland（2012a：30）的评价：“很难想到比学术英语受语料库影响更大的应用语言学研究”。

EAP研究大致可以归纳为“语域分析”（register analysis）“语用分析”（pragmatic analysis）、“语境分析”（contextual analysis）和“体裁分析”（genre analysis）四个发展阶段。国外早期的EAP语域分析围绕科技文本展开，基于小样本语料报告某类语言形式的出现频率，如Barber（1962）和Trimble *et al.*（1978）。这类语域分析都是频率描述性的报告，没有对交际功能作出解释。到了80年代，语用和语境分析受到关注，前者关注语用功能，探索各类功能（如下定义、模糊表达等）的语言实现形式，而后者强调在语境中深入细致分析语言形式的交际功能。例如Holmes（1988）分析教材中表示质疑和肯定的语言形式，Tarone *et al.*（1981）分析被动语态在科学论文中的交际意图与功能。早期的这些研究一方面所用语料样本的规模较小，另一方面语料文本的体裁与学科分类混杂，但是为后续的EAP研究提出了一系列值得思考的问题，比如语料样本选择与构成、分析方法和数据阐释等。直到90年代初期，随着Swales（1990）和Bhatia（1993）系统论述体裁

概念，体裁分析迅速成为EAP研究的重要方法。与此同时，计算机技术发展促使EAP研究采用大规模语料库，实现对语言形式或语用功能的批量检索。Swales（1990）关于论文引言的语步分析和Hyland（1994）关于学术论文中模糊限制语的论述，无疑是这一时期最具代表性的EAP语料库研究。他们不仅明示运用语料库作为研究工具，借助计算机实现半自动检索，而且在语料库建设时强调语料文本体裁的单一性。我们不难看出，国外早期EAP研究几乎都是基于小型自建语料库，语料文本由“杂”逐渐演变到“专”，对语料库的依赖和定位也从隐含描述过渡到明示表达。

根据徐秀玲、许家金（2017），我国语料库研究始于20世纪70年代末80年代初，首先重点发展的即是科技英语语料库的建设及应用。当时，清华大学、大连工学院等院校相继建立了百万词级的科技外语文本语料库，考察科技英语的词频与分布，编写了《科技英语常用词汇3000》等书。1983年，上海交通大学创建了世界上第一个EAP语料库——上海交大科技英语语料库（Jiao Da English for Science and Technology，简称JDEST），语料涵盖计算机、化工、电工、机械等10个理工科专业英语文本，建成时容量为100万词，其后规模不断扩充。该语料库用于编制科技英语常用词表（内含通用词汇、技术词汇、半技术词汇等词项）（杨惠中、黄人杰 1982），为大学英语教学改革提供了可靠的量化依据。直到90年代末至2000年初，国内开始增加基于语料库的EAP研究，主要围绕语法特征（如雷秀云 2000）和语用功能（如余千华、秦傲松 2001）。可见，国内早期EAP语料库研究的状况是语料库建设工作多于应用实践，且在相当长的一段时间里以统计词频、制定词表为主。

2. 请问您是如何开始对基于语料库的EAP研究产生兴趣的？怎样的时机（或者某些学者或作品）促使您走上EAP语料库研究之路？

我对EAP语料库研究产生兴趣得益于有幸师从Ken Hyland教授攻读应用语言学博士学位。其实，我的本硕教育背景是国际经济与贸易专业，在2013年攻读博士学位之前一直从事商务英语教学，对商务语篇研究感兴趣。在2012年申请香港大学博士研究生学习时，我提交的研究计划是分析反倾销起诉书的语篇特征。在准备过程中，我广泛阅读了关于ESP语篇研究的文献，丰富了对体裁分析理念与方法的了解。此外，我也细读了Ken Hyland教授一系列的文章与著作，对*Disciplinary Discourses: Social Interactions in Academic Writing*一书特别感兴趣，惊讶于学术语篇的学科差异如此之大，更是被Ken Hyland教授对于学科差异的社会学阐释深深吸引。2013年暑期，我参加了外语教学与研究出版社举办的“语料库在外语教学与研究中的应用”研修班。在研修班上，许家金老师介绍了“外壳名词”（shell nouns）这一概念以及Schmid（2000）提出的语义分类。顿时，我便思考一

个问题：这些不同语义的外壳名词应该在所谓的“软学科”和“硬学科”语篇中有不一样的分布。

在第一次与导师见面的时候，我表达对考察外壳名词学科差异的兴趣。导师很认同，但是也提出我需要重新考虑外壳名词的分类，因为Schmid的分类不清，子类之间互有重叠。例如，fact一词既出现在“事实类”（factual）又在“情态类”（modal）划分中。之后，我便阅读大量关于此类名词的文献，寻找解决方案，最终提出了自己的“本质、特征和关系”分类框架（姜峰 2015）。此外，我发现这类名词在语篇中实际上起到的是类似元话语的功能，于是我称之为“元话语名词”（metadiscursive nouns），并基于Hyland（2005）的元话语框架，界定该类名词的“引导”（interactive）和“互动”（interactional）两方面功能。我通过正则表达式在自建多学科语料库中检索此类名词，并统计其在句式结构、文理学科、论文组成部分等方面的分布差异，然后结合学术英语、体裁分析、学科知识论等不同领域文献，解读差异产生的可能原因。以上是我接触并走上EAP语料库研究之路的大概经过。

反思过往，Ken Hyland教授的三本专著*Disciplinary Discourses: Social Interactions in Academic Writing*、*Metadiscourse: Exploring Interaction in Writing*、*Disciplinary Identities: Individuality and Community in Academic Discourse*对我影响很大，丰富了我对EAP语料库研究的认识。我认为这三本专著也是从事基于语料库的EAP研究必读的经典文献。它们都采用社会建构主义视角，基于自建多学科EAP语料库，调查不同的修辞资源和语言标记，系统论述学术语篇中蕴含的人际互动。在上述研究中，作者并没有采用复杂的语料库技术，仅通过量化统计语言特征的发生频率和分布规律，呈现不同话语共同体享有的学术话语实践和语篇范式。第一本讲述了不同学科话语共同体之间创造知识、建构话语的差异；第二本聚焦元话语，讨论其丰富的人际意义与劝谏功能；第三本论述了个体与话语共同体之间的社会互动关系。很重要的是，作者向我们展示了如何汲取应用语言学、社会学，聚焦小组讨论和个体访谈等不同视角与方法，对语料库统计结果进行充分而深刻的阐释。

3. 您认为基于语料库的EAP研究的理论和实践意义体现在哪些方面？

一般说来，EAP研究具有明显的外语教学取向，它突出目标话语共同体，常以文本为切入点，通过实证描述目标话语共同体共享的特定语言特征、体裁资源和话语实践，并以此组织教学，提高学生体裁与修辞意识，使其了解交际目的和学科文化，从而更好地加入目标共同体，构建学术身份。EAP研究通常建立在分析学生对学术语言需求的基础上，探究特定体裁的文本特征和话语实践，以此指导课堂教学与测评。例如，学生用英语撰写论文摘要时存在诸多困难，需要增强把握该体裁的话语能力，因此学术英语教师基于该需求，了解论文摘要的交际

目的和目标话语共同体，分析该体裁的语言及修辞特征，进而安排教学内容，提高学生驾驭该体裁的体裁意识和话语能力，撰写有说服力且符合语篇规约的论文摘要。可见，EAP研究的理念是，话语共同体约定俗成的话语实践起到把关（gatekeeping）作用，新手或者二语学习者需要增强发挥该话语实践的学术话语能力，使自己被共同体接纳，建立共同体成员身份（Swales 1990；Hyland 2012b）。因此，基于语料库的EAP研究帮助我们揭示学术语言的本体特征，为教学提供实证参考。

语料库为EAP研究提供大量真实的语言使用实例，呈现学术英语在词汇、短语以及句式层面的分布。Biber *et al.*（1999）基于大规模朗文英语口笔语语料库，报告学术英语与其他语域（如小说、报纸、会话等）在各类词汇和句式方面的差别。例如，学术语篇中名词使用比例高，动词比例低，短语结构比例高，从句结构比例低。Gardner & Davis（2013）和Dang *et al.*（2017）报告学术英语口笔语语境中常用词汇列表；Ackermann & Chen（2013）编纂学术英语常用搭配列表；Biber *et al.*（2004）和Hyland（2008）考察语块在学术语篇中的结构特点和交际功能；Hewings & Hewings（2002）和Hyland & Tse（2005）分别讨论“引入型it从句”和“评价型that从句”的句式结构在学术语篇中的分布和使用特点。

另一方面，基于语料库的EAP研究在真实学术文本中挖掘语言形式与功能之间的关系。受功能语言学影响，EAP研究强调学术语言的形式和功能之间的关联，特别是人际和篇章功能与其语言实现形式之间的关系。传统观点认为，学术语篇是完全客观、脱离主观因素的数据报告，但是大量的EAP语料库研究表明，学术语篇不仅传递科学信息，而且表达丰富的人际意义（Hyland 2000，2005a；Hewings & Hewings 2002）。立场评价（stance）和读者带入（engagement）是学术语篇人际意义最为突出的两个方面（Hyland 2005b）。语料库通过语言实例印证了Latour & Woolgar（1986）等科学知识社会学研究的结论，让我们认识到，学术语篇的说服力不仅在于词句正确的语言形式，而且也建立在人际互动与劝谏修辞的语言功能上。

此外，学术英语语言的学科间性以及体裁导向性无疑是EAP语料库研究另一个重要的理论意义。Jordan（1997）将EAP进一步划分为EGAP（English for General Academic Purposes，通用学术英语）和ESAP（English for Specific Academic Purposes，特殊学术英语）。前者指不同学科专业通用的语言和话语实践共核，后者指针对特定的学科与专业，学生完成知识学习与学术交流时所使用的英语语言。但是EAP语料库研究的数据表明，EGAP文本几乎是不存在的，不论语言形式还是修辞功能都表现出学科特殊性。Biber & Gray（2016）运用大型语料库示例理工学科文本的语言形式偏向短语结构，而人文学科文本偏向小句结构。Hyland（2000，2005a）展示人文学科文本的人际互动与修辞策略明显多于理工学

科。因此，基于语料库的EAP研究使我们不断认识到学术语篇的社会建构属性，学术语言与特定学科共同体的知识建构和论辩方式紧密相关。

语言本体研究对教学实践的应用启示一直是EAP研究的立足点。正如我们上面谈到的，EAP研究的指导思想是通过挖掘和描述学术话语的语言特征，指导课堂教学，增强其针对性与有效性。通过与EAP密切相关的两家国际期刊*Journal of English for Academic Purposes*与*English for Specific Purposes*的宗旨和兴趣，我们也可以认识到EAP研究的这一立足点，即“对学业和学术交流语境下的英语使用给予语言学描述”，用于“基于特定共同体话语实践的教学与学习”。接下来，我们围绕上述三方面的理论意义，分别讨论各自的实践意义。

首先，在教学中我们要提高学生的语域和体裁意识，认识到学术语言与日常用语的差别。例如，借助Mark Davis创建的网站https://www.wordandphrase.info，学生可以对比并观察不同文本所使用的学术词汇（Gardner & Davis 2013），也能够反思自己的写作文本是否符合学术语域的语言形式。第二，我们还要增强学生的读者意识以及学术话语人际互动能力。根据对读者群的理解和预期，学生应该不仅可以表达恰当的观点评价，也能够运用适当的语言资源将读者带入语篇，与读者对话。第三，教学内容和课程设置要充分体现EAP的学科特殊性。在EAP语料库研究的影响下，香港大学的学术英语课程以学科为单位进行安排和实施，课程总名称为“English in the Disciplines”（Hyland 2017），这样能够使学生充分接触和学习特定学科的语言形式与话语表达。此外，在教学中，我们也要培养学生使用语料库进行自主学习的习惯，使其从EAP学习者转换到EAP观察者和分析者。Johns（1991）称之为数据驱动学习（data-driven learning，DDL），让学生在教室或课下自学时使用语料索引，探索学术英语的规律和模式，在活动和练习中，体现出以学习者为中心的发现式学习。

4. 能否请您简单介绍一下当前国内外EAP语料库研究的主要热点？

我认为，当前EAP语料库研究主要围绕语言资源、体裁实践和学科共同体展开。通过语料库的丰富语言实例，我们观察到学术话语参与者有规律性且反复使用的语言选择和体裁结构。因此，语料库展现的不是一个人的语言习惯，而是背后话语共同体的语言规约和认知方式。

首先，在语言资源方面，EAP语料库研究不断挖掘学术文本中各种类型的语言资源，探讨语言形式与功能之间的关系，了解学术语境下知识建构与信息呈现的语言依托。以引用（citation）为例，它是学术文本特有的语篇特征。如何使学生的引用实践具有合理性和说服力是EAP研究的任务。语料库提供大量证据，呈现给我们不同引用形式的频率和功能。比如，我们对比不同学科学术论文语料库，

发现嵌入式（integral）引用在人文学科的比例高于理工学科，理工学科更倾向使用剥离式（non-integral）引用。试比较如下两例。（1）Kaplan（1972）claims that problems of organization in academic writing by ESL students are due to cross-cultural differences in rhetoric（嵌入式）。（2）all assembling components serve as potential individual engineering components（Elowitz and Leibler，2000，Gardner *et al*. 2000 and Wang *et al*. 2011）（剥离式）。此外，就嵌入式而言，除claim一词转述被引者观点外，还有哪些转述动词，分别表达何种功能与意义？类似这样带有教学用意的语言资源问题，都是EAP语料库研究的热点。目前随着MAT tagger工具的开发，且能够有效复制Biber（1995）的多维分析（Multidimensional Analysis），越来越多的研究运用该方法考察不同学术文本的语言特征。

体裁实践是EAP乃至整个ESP研究的持久话题。一方面，研究者基于语料库分析某类体裁的语步（move）及语阶（step），或者聚焦某个关键语步，考察其语言实现形式。例如，Hu & Liu（2018）关注当前热议的三分钟演讲（3 Minutes Thesis，3MT）的体裁实践，归纳参与者普遍采用的体裁序列，并分析每一语步的修辞特征。体裁分析对语言教学有重要的指导作用，给予学生提高话语能力所需的支架式（scaffolding）指导，因此在相当一段时间内它将仍然是EAP语料库研究的热点。另一方面，研究者基于语料库对比不同的体裁实践，了解不同交际目的与读者对象的语篇结构和修辞策略。根据Swales（1990），体裁的核心是交际目的和语篇结构，影响作者的话语风格和语言选择。通过对比不同的体裁实践，我们能够培养学生的读者意识或者体裁意识，即如何针对不同的读者，采用不同的修辞策略，作出不同的语言选择。譬如，目前对比科普文本（popular science）与专业科学文本（professional research）的体裁实践是一个热点，Zhang（2015）考察两种体裁在被动语态方面的使用差异，Liu & Deng（2017）分析抽象名词结构的使用差异。

此外，我们前面谈到，EAP研究突出目标话语共同体，因此话语共同体是EAP语料库研究的另外一个热点，至少体现在两个方面。首先，不同话语共同体之间的对比研究，最显著的当属学科话语对比研究。研究者从2000年初期的比较“软学科”（soft disciplines）和“硬学科”（hard disciplines）（如Hyland 2000），逐渐发展到对比某一学科领域的不同范式，例如McGrath & Kuteeva（2012）聚焦理论数学这一细致的理科方向；Hu & Cao（2015）则比较应用语言学定量与定性研究范式的学术论文在元话语使用的差异。此外，另一个热点是研究作者个体与话语共同体之间的关系，最显著的当属作者身份研究。身份是一个多面体概念，表现主义者认为身份是个体的自我表现，建构主义者则表示身份是社会协商过程中的话语展现。Hyland（2012）无疑是基于语料库研究学术身份的代表著作，通过量化分析作者立场、读者带入、语篇衔接等语言标记，探讨作者展现自己对概念

内容和话语共同体的学术姿态（positioning），但是同时，学科身份还体现在作者靠近话语共同体的亲缘过程（proximity）。

5. 基于语料库的EAP研究前景如何？哪些方面值得我们进一步挖掘？

EAP领域的旗舰刊物*Journal of English for Academic Purposes*于2002年创刊，迅速进入SSCI索引数据库，并已经成为应用语言学领域重要的学术期刊。可见，EAP语料库研究的前景十分广阔。但是当前一个普遍的问题是，现有研究同质化现象严重，常常机械地套用国外理论或分析框架，缺少对学术英语研究的本质、语料库方法的适用性与解释力以及教学指导意义等问题的深入思考。比如，有人认为Hyland（2005）的立场框架是一个热点，便随意选取某个学术体裁文本建立语料库，分析其立场标记的频率和分布。但是可能无法回答审稿人或者读者可能产生的质疑：为什么要在该体裁中考察立场而非其他修辞策略？既然研究作者的态度，为什么基于Hyland（2005）的立场概念和框架，而非Hunston（2011）的“评价”（evaluation）或Martin & White（2005）的“评价”（appraisal）？我认为，对EAP研究理念和文献脉络的理解与认识是思考这些问题的前提，也关乎EAP语料库研究的前景。

说到学术英语语料库研究的前景，我们逐一从“学术”“英语”和“语料库”进行讨论，分别指涉学术语境、英语本体和语料库方法。首先，就学术语境而言，以往的EAP语料库研究通常以共时研究为主，或者对比学术语篇的学科差异，或者比较一语和二语作者的语篇差异，或者比较新手和资深学者的话语实践，或者对比不同体裁的学术文本，缺少对动态语境的考量（朱永生 2005）。因此，基于语料库的学术英语历时研究应该是重要的前景之一。一方面，通过语料库的文本实据，我们观察到学术语言选择的历时变化，进而推知学术语境以及社会文化宏观环境的重要演变。另一方面，我们亦可纵向跟踪并收集作者产出文本，建立历时语料库，考察作者在不同时期的语言选择，不仅可以分析作者学术身份的动态变化，而且了解显性教学指导的影响。此外，科学传播大众化是当今学术语境的一个重要特色，因此考察新兴学术体裁（例如学术博客、多媒体论文、TED演讲等）也不失为一个有价值的方向。Maria Kuteeva和Anna Mauranen最近在*Discourse, Context & Media*期刊组织的Digital Academic Discourse专刊中有六篇论文论述这类体裁，给我们提供了不错的借鉴。

第二，英语作为通用语无疑是英语语体在当今语境下的一个重要特征。学术英语亦是如此。英语在学术交流中作为一种有效接触性语言，能够使拥有不同母语、不同文化的人之间互相沟通。也如同其他领域一样，英语非本族语使用者远远超过本族语者。因此，英语不是学术交流群体的第一语言，而是第二或者附加语言。随着Anna Mauranen发起建立英语作为学术语境的通用语语料库（English

as Lingua Franca in Academic settings，ELFA语料库），EAP研究者开始关注学术语境下英语通用语研究。尽管ELFA语料库与后期建立的学术书面通用英语语料库（Written English as a Lingua Franca in Academic settings，WrELFA语料库）对外免费开放，但是相关研究目前仅限于欧洲。因此，可以预见基于语料库的英语作为学术通用语研究会具有广阔的前景。

第三，就语料库而言，其前景应该体现在具有新意的EAP研究方法和语料。目前EAP语料库研究多以书面语料为主，例如期刊论文、课程作业、学位论文和教材等，应该加强学术语境下的口语交际语料的建设与研究。在2017年于香港大学举办的“专业及学术英语国际研讨会”的学术英语论坛上，Ken Hyland、John Swales、Anna Johns和John Flowerdew一致表示，学术英语口语语料库的建设和开发是学术英语研究的一个瓶颈，致使目前涉及口语和听力的学术英语研究缺乏，同时也是EAP研究未来的一个重要方向。在研究方法方面，学术英语研究目前大体上是基于语料库的研究，语料库驱动的研究相对较少。词语共选、语义趋向和语义韵、短语序列等语料库驱动方法有待在学术英语研究中进行系统考察。此外，目前基于语料库的学术英语研究主要通过研究者的客位视角（etic）解读语言本体特征，即依靠研究者的主观认知与分析经验阐释特定语言特征的形式与功能。因此，基于语料库的EAP研究应该加强引入语篇参与者的主位（emic）认知和见解，或者通过话语共同体的内部视角阐释语言使用的特点和目标，弥补研究者的客位观察。与语料库分析进行三角互证的常见方法包括个体访谈、聚焦小组讨论、有声思维、反思日志等。

6. 能否请您谈谈我国学者如何能作出具有本土特色的EAP语料库研究？

向世界讲好中国学术故事无疑是学术英语研究的时代命题。因此，具有本土特色的EAP语料库研究就要紧紧围绕中国学者与学习者的本土学术语境，关切国情，聚焦他们的学术语言能力。譬如，国内EAP语料库研究在考察学科差异与特殊性时，常常跟随国外文献的学科选择，粗略地划分以人文和社会科学为代表的软学科和以自然科学和生命科学为代表的硬学科，缺少关切我国本土学科属性和专业设置。这导致所建EAP语料库及其研究发现对我国本硕博各阶段学生以及年轻学者的指导意义受限。2016年，中国外语与教育研究中心牵头设立了“中国外语教育基金专用英语语料库建设项目”，设计库容不少于一亿词次，开创性地按照我国学科专业设置的实情，涵盖人文社会科学、自然和生命科学各主要领域的一级和二级学科，建设学术用途英语语料库（Database of English for Academic Purposes，DEAP）。DEAP无疑是关切我国多学科学术英语教学实情、服务本土教学与研究的学术英语语料库，产生了许家金（2017）、Jiang & Wang（2018）等

一系列特色学术成果，期待更多的适合我国学科专业国情的EAP语料库建设与研究。另一方面，我们也缺少通过语料库考察我国学者学术话语能力的动态发展研究。例如，未来研究可以收集我国学者在不同时间段产出的学术文本，以此建立我国学者EAP语料库，并用语料库分析结果设计深入访谈、民族志等质性研究方法。

随着我国高等教育国际化的持续深入，特别是在“双一流”高校建设背景下，EAP研究的规模也将不断扩大。因此，我国新时期外语教育需要更多高水平、接地气的EAP语料库研究，为国际化人才学术话语能力的培养提供必要的实证参考和教学启示。

注 释

1. 网址为：https://sites.google.com/site/multidimensionaltagger/home

参考文献

Ackermann, K. & Y. Chen. 2013. Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach [J]. *Journal of English for Academic Purposes* 12(4): 235-247.

Barber, C. 1962. Some measurable characteristics of modern scientific prose [A]. In F. Behre (ed.). *Contributions to English Syntax and Philology* [C]. Gothenburg: Almqvist and Wiksell. 21-43.

Bhatia, V. 1993. *Analysing Genre: Language Use in Professional Settings* [M]. London: Longman.

Biber, D. 1995. *Dimensions of Register Variation: A Cross Linguistic Comparison* [M]. Cambridge: CUP.

Biber, D. & B. Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing* [M]. Cambridge: CUP.

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 Written and Spoken English* [M]. Harlow: Longman.

Dang, T. N. Y., A. Coxhead & S. Webb. 2017. The academic spoken word list [J]. *Language Learning* 67(4): 959-997.

Gardner, D. & M. Davies. 2013. A new academic vocabulary list [J]. *Applied Linguistics* 35(3): 305-327.

Hewings, M. & A. Hewings. 2002. “*It is interesting to note that…*”: A comparative study of anticipatory “it” in student and published writing [J]. *English for Specific Purposes* 21(4): 367-383.

Holmes, J. 1988. Doubt and certainty in ESL textbooks [J]. *Applied Linguistics* 9(1): 21-44.

Hu, G. & F. Cao. 2015. Disciplinary and paradigmatic influences on interactional metadiscourse in research articles [J]. *English for Specific Purposes* 39: 12-25.

Hu, G. & Y. Liu. 2018. Three Minute Thesis presentations as an academic genre: A cross-disciplinary study of genre moves [J]. *Journal of English for Academic Purposes* 35: 16-30.

Hunston, S. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language* [M]. New York: Routledge.

Hyland, K. 1994. Hedging in academic writing and EAP textbooks [J]. *English for Specific Purposes* 13(3): 239-256.

Hyland, K. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing* [M]. Harlow: Longman.

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

Hyland, K. 2005b. Stance and engagement: A model of interaction in academic discourse [J]. *Discourse Studies* 7(2): 173-192.

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

Hyland, K. 2012a. Academic discourse [A]. In K. Hyland, C. M. Huat & M. Handford (eds.). *Corpus Applications in Applied Linguistics* [C]. London: Continuum. 30-46.

Hyland, K. 2012b. *Disciplinary Identities: Individuality and Community in Academic Discourse* [M]. Cambridge: CUP.

Hyland, K. 2017. English in the disciplines: Language provision in Hong Kong's new university curriculum [A]. In E. Park & B. Spolsky (eds.). *English Education at the Tertiary Level in Asia* [C]. New York: Routledge. 27-45.

Hyland, K. & P. Tse. 2005. Evaluative *that* constructions: Signalling stance in research abstracts [J]. *Functions of Language* 12(1): 39-63.

Jiang, F. & F. Wang. 2018. *"This is because..."*: Authorial practice of (un)attending this in academic prose across disciplines [J]. *Australian Journal of Linguistics* 38(2): 162-182.

Johns, T. 1991. *Should You Be Persuaded: Two Samples of Data-driven Learning Materials* [M]. Birmingham: University of Birmingham English Language Research.

Jordan, R. R. 1997. *English for Academic Purposes: A Guide and Resource Book for Teachers* [M]. Cambridge: CUP.

Latour, B. & S. Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts* [M]. Princeton, N.J.: Princeton University Press.

Liu, Q. & L. Deng. 2017. A genre-based study of shell-noun use in the N-be-that construction in popular and professional science articles [J]. *English for Specific Purposes* 48: 32-43.

Martin, J. & P. White. 2005. *The Language of Evaluation: Appraisal in English* [M]. New York: Palgrave Macmillan.

McGrath, L. & M. Kuteeva. 2012. Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices [J]. *English for Specific Purposes* 31(3): 161-173.

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

Swales, J. 1985. *Episodes in ESP: A Source and Reference Book on the Development of English for Science and Technology* [M]. Oxford: Pergamon Institute of English.

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

Tarone, E., S. Dwyer, S. Gillette & V. Icke. 1981. On the use of the passive in two astrophysics journal papers [J]. *The ESP Journal* 1(2): 123-140.

Trimble, M., L. Trimble & K. Drobnic. 1978. *English for Specific Purposes: Science and Technology* [M]. Corvallis: English Language Institute, Oregon State University.

Zhang, G. 2015. *It is suggested that… or it is better to…?* Forms and meanings of subject it-extraposition in academic and popular writing [J]. *Journal of English for Academic Purposes* 20: 1-13.

姜 峰，2015，本质，特征，关系：外壳名词三分法及人际功能研究 [J]，《语料库语言学》（2）：62-74。

雷秀云，2000，基于语料库的学术英语语法的频率特征 [J]，《上海交通大学学报（哲学社会科学版）》（1）：117-122。

许家金，2017，体裁短语学视角下的医学学术英语词典研编 [J]，《外语与外语教学》（6）：52-60。

徐秀玲、许家金，2017，我国外语教学中的语料库应用40年 [J]，《中国外语教育》（4）：62-68。

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。

余千华、秦傲松，2001，英语科技论文中的模糊限制语 [J]，《华中科技大学学报（社会科学版）》（4）：121-123。

朱永生，2005，《语境动态研究》[M]。北京：北京大学出版社。

通信地址： 130012 吉林省长春市吉林大学公共外语教育学院

中国英语学习者口语句法复杂性多维分析*

安徽工程大学 徐 鹏

提要：本研究在梳理已有句法复杂性测量指标的基础上，通过因子分析的方法尝试建立一个多维测量体系，并据此对中国大学生在记叙与议论性英语口语产出上进行了跨年级比较，以揭示在从属结构比率、单位结构长度、并列短语和名词性结构四个维度上，中国英语学习者口语产出句法复杂性上的发展规律。方差分析显示：除了代表从属结构比率维度的C/T指标在年级水平上具有稳定性以外，其他三个维度上的指标均受到年级和文体共同影响。三个维度上句式复杂性与年级水平正相关，且具有较强的文体区分性。口语议论性文体复杂性普遍高于记叙文。

关键词：口语产出、句法复杂性、测量、文体、年级水平

1. 引言

句法能力是学习者语言发展的重要表现，是二语习得领域的研究重点之一。句法复杂性作为句法能力的一个重要组成部分，涉及语言产出单位长度、从句嵌套的比率、结构类型的多样化以及特定结构的复杂化程度等。它们是二语发展的重要指标（Ortega 2003）。长久以来，句法复杂性研究集中在二语写作层面。部分研究从理论构建和研究方法角度，探寻二语写作句法复杂性的组成维度及测量指标；另一部分探寻句法复杂性与语言水平和写作质量的关系及句法复杂性的发展过程；还有探讨写作任务、学习者背景、教学写作环境等外部变量对句法复杂性的影响（陆小飞、许琪 2016）。而英语学习者口语句法能力的研究屈指可数。国外的学习者口语研究多集中于横向外部变量影响研究，即揭示准备时、任务类型、是否限时等条件对流利度、准确度、句法复杂性的影响（Skehan 1996；Yuan & Ellis 2003；Wigglesworth 1997；Ortega 1999；Skehan & Foster 1997；Ellis 2009），展示了口语句法表现对外界变量的敏感性。而纵向口语句法发展方面，Loban（1976）曾调查了幼儿园到12年级儿童语言口语、书面语流利度和句法结构的发

* 本文为教育部人文社科青年基金项目“中国大学生英语口语语体特征多维度研究（15YJC740107）”及“2016安徽省高校优秀中青年骨干人才国内外访学研修重点项目（gxfxZD2016106）”阶段性成果。

展。除了动词密度以外，其他的指标没有明显的发展阶段。Nippold *et al.*（2005）考察了不同年龄者英语母语的发展程度，揭示出说明性口语复杂度高于日常会话。且口语句法复杂性从儿童发展到青壮年时期，在中年时期保持稳定。国内也有部分学者研究了口语句法的二语发展（李茶、隋明才2007；文秋芳2010；彭慧2012），以及不同水平者英语口语句法复杂性之间的横向比较（鲍贵2010）。

以上口语研究肯定了语言水平、写作任务，以及是否限时等变量对口语复杂度的影响。然而在测量指标和测量方法上，沿用的依然是书面语句法复杂性的测量范式。然而，口笔语料分属不同的语体，对于口语句法复杂的测量直接套用笔语的测量范式是否妥当值得商榷。鲍贵（2010）仅仅用了T单位长度和子句密度两个指标初步揭示了不同水平英语学习者在书面语和口语体句法复杂性上体现出了不同的特征。Nippole *et al.*（2005）用T单位长度和关系从句两个指标证实了英语母语者在谈话和说明类口语上呈现不同的句法复杂性。而Biber（2011）也通过大型语料库证明，考察T单位为主的附属小句成分并非对所有语域都具有区分度。大部分小句附属结构的使用其实是口语化的特征，并非高级书面语的典型特征。高水平写作者倾向于使用更简化的句子结构（Cooper 1976）。语言使用到最高阶段是名词化所体现的语法隐喻（Halliday & Matthiessen 1999）。可见，句法复杂性对语域具有较强的敏感性。因此在口语句法复杂性的研究上，我们应当考虑语体的差异性、测量指标的全面性，以及口语语料的代表性。即目前现行的广泛应用于笔语测量指标对口语测量是否具用同样的实用性？二语口语句法测量应当选用何种指标群？

前人的实证研究习惯性地套用T单位作为句法复杂性的指标，最早由Hunt（1965）在检验英语本族语儿童写作发展而提出，其定义为包含一个主句，以及附加和嵌入的所有从句和非从属结构“不可分割的最小单位”。Hunt（1970）认为测量学生句法发展的最好指标是T单位长度。而T单位的广泛使用，均是基于一个前提：子句从属性代表着句法的复杂程度。然而在高级学习者学术写作中，常倾向于使用名词化短语结构代替从句，我们并不能因此认为学术英语的复杂性不如口语对话；从另一面来看，中国英语学习者在口语产出中存在着例如“I lived in a house which is very big.”之类无意义低功能的从属结构，由此可以发现单纯的T单位为主体的附属小句考察有其局限。

2. 研究设计

2.1　研究问题

针对中国英语学习者这一群体，笔者认为，句法复杂性的研究应：1）建立更加全面的句法复杂性测量体系，确立适合二语口语句法复杂性的考察指标群；2）句法复杂性的考察要考虑不同文体的差异性。在上述框架下，本研究试图回答：

（1）各测量指标及其组合效应对二语句法复杂性影响程度如何？

（2）年级和文体如何影响中国学习者二语口语句法复杂性的变化？

2.2 语料收集

本研究随机选取国内某一本理工科高校英语专业大二、大四学生各50名，且母语均为汉语，无出国经历。分为低年级组和高年级组。在口语任务上使用组图分别诱导采集记叙和议论两种口语语料。第一组图为四幅图片构成的故事，使用组图诱发产出不仅统一了作业内容，而且可以有效规避其他文本带来的文字干扰。为方便与国内其他研究对比，本研究统一要求学生准备三分钟，并独白三分钟，讲述故事的经过。第二组图为一个话题“Should the Old People Enjoy Free Bus Ride”，让学生对其进行评论。同样给予三分钟时间准备，独白三分钟。在得到两个年级组不同文体的口语语料之后，我们进行了口语转写以及文本净化，剔除口语中重复、自我修正等影响句法复杂性分析的内容。

表1 语料构成

	记叙性口语	议论性口语
低年级组	40篇	40篇
高年级组	40篇	40篇

3. 研究结果

3.1 二语句法复杂性测量指标的优化和筛选

在测量指标方面，目前不论口语、笔语，几乎所有的研究在都围绕T单位展开，其他指标有从属小句和句子等。但由于自动标注和分析工具的缺乏，2005—2015十年的前人研究中，大部分的测量指标均在1-4个之间，使得该领域的研究缺乏全面性和系统性（陆小飞、许琪2016）。为了弥补前人测量指标上的不足，本研究采用陆小飞（2010）开发的L2SCA二语句法复杂性分析器进行标注分析。汇集了Wolfe-Quintero *et al.*（1998）和Ortega（2003）文献综述中提到的14个测量指标，以及其他9个表层指标在内的23个频数和比例变量，进行因子分析，对涵盖的23个测量指标进行相关性的考察，拟试图揭示在二语口语复杂度测量中，哪几种变量组合的因子解释的方差最大。之所以使用这23个指标，是为了和张艳敏等（2015）使用写作文本测量指标所作的因子分析进行对比。通过二语句法复杂性分析软件L2SCA分析出句法复杂性各类指标数据。然后运用SPSS进行因子分析。

提取出主要的解释因子，根据因子将所有的23个测量指标分为几大主要维度。从几大维度层面择选少量代表性的指标进行跨年级和跨语域的考察和比较。但笔者认为，中国大学生口语可能会产出不同于笔语的语法特点，在指标相关性上也会有不同的呈现。

通过KMO测度和Bartlett球性检验得出，表2中KMO的值为0.69，显著水平p<0.01，表明变量间有共同因子，各变量有一定相关性，变量间存在显著关系，适合开展因子分析。

表2　KMO和Barlett检验

KMO检验		.698
Bartlett球性检验	卡方值	8358.318
	自由度	253
	显著性	.000

表3　因子特征值

成分	总载荷	解释方差%	累积解释方差%
1	7.683	33.405	33.405
2	6.136	26.679	60.084
3	3.166	13.767	73.851
4	2.432	10.573	84.423
5	1.261	5.482	89.905

表4　因子构成及载荷

	成分				
	1	2	3	4	5
DC/T	**.960**				.104
C/T	**.955**				
DC/C	**.884**	.163			.129

（待续）

（续表）

	成分				
	1	2	3	4	5
CT/T	**.865**	.112	.114		
VP/T	**.864**			.112	.224
C	.107	**.926**	-.228	-.124	-.185
VP	.148	**.901**	-.223		
W		**.765**	.402		.368
T	-.508	**.755**	-.259	-.106	-.184
CT	.391	**.755**	-.138	-.125	-.120
S	-.440	**.684**	-.452		
DC	.663	**.680**	-.119		
MLS	.115	-.176	**.919**		.228
T/S	-.112		**.875**		-.238
MLC		-.285	**.811**		.414
MLT	.384	-.243	**.716**		.423
C/S	.658		**.675**		-.151
CP		.177		**.953**	
CP/C		-.260		**.937**	.155
CP/T	.320	-.228		**.878**	.152
CN/C	.117	-.297	.246	.235	**.839**
CN	.254	.499		.201	**.699**
CN/T	.602	-.203	.196	.229	**.678**

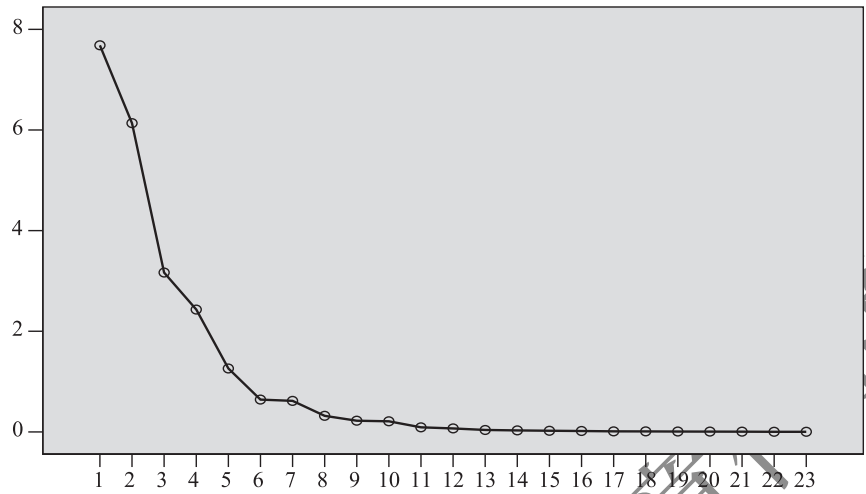


图1　碎石图

表5　测量指标五个维度

类别	缩写	指标名称
1. 从属结构比率维度	DC/T	T单位平均从属子句数
	C/T	**T单位中子句数量**
	DC/C	各子句中从属子句数量
	CT/T	复杂T单位比率
	VP/T	T单位中动词短语数量
2. 单纯结构频数维度	C	子句数
	VP	动词短语数
	W	文本长度
	T	T单位数
	CT	复杂T单位数
	S	句子数
	DC	从属子句数
3. 单位结构密度维度	MLS	平均句子长度
	TS	各句中T单位数量
	MLC	平均子句长度

（待续）

（续表）

类别	缩写	指标名称
	MLT	**平均T单位长度**
	C/S	各句子中子句数量
4. 并列短语维度	CP	并列短语数
	CP/C	各子句中并列短语数量
	CP/T	**各T单位中并列短语的数量**
5. 名词性短语维度	CN/C	各子句中复杂名词性短语数量
	CN	名词性短语数量
	CN/T	**各T单位中复杂名词性短语数量**

（其中子句指的是任何带有一个主语和一个限定性谓语动词的结构，包括独立句、形容词性、副词性或名词性子句，不包括非限定性动词短语。复杂T单位指的是有一个或多个从句的T单位。）

根据因子分析的结果，尤其是图1的碎石图可以清晰地看出，从第五个因子之后的曲线变得平缓，再根据表3中有五个初始特征值超过1.0的可接受值，其累计解释方差达到89.905，遂决定抽取5个因子作为23个变量背后共同的作用因子。图6为Varimax旋转后成分矩阵结果，根据各变量的载荷大小，我们将所统计的23个变量根据五个因子分为五个维度：

因子一：主要由DC/T（T单位平均从属子句数）、C/T（T单位子句数量）、DC/C（各子句中的从属子句数量）、CT/T（复杂T单位比率）、VP/T（T单位中动词短语数量）五个变量组成，且这五个变量都有大于0.86的超高载荷，解释方差达到33.405，可命名为从属结构比率因子。

因子二：主要由C（子句数）、VP（动词短语数）、W（文本长度）、T（T单位数）、CT（复杂T单位数）、S（句子数）、DC（从属子句数）七个变量组成，解释方差为26.679，命名为单纯结构频数因子。

因子三：主要由MLS（平均句长）、TS（各句中T单位数量）、MLC（平均子句长度）、MLT（平均T单位长度）、C/S（各句子中的子句数量）五个变量组成，解释方差达到13.767，可命名为单位结构密度因子。

因子四：主要由CP（并列短语数）、CP/C（各子句中并列短语数量）、CP/T（各T单位中的并列短语数量）组成，解释方差为10.573，虽然只支撑三个变量，但各变量载荷均高于0.878，可命名为并列短语因子。

因子五：支撑CN（复杂名词性短语数量）、CN/C（各子句中复杂名词性短语数量）、CN/T（各T单位中复杂名词性短语数量），可命名为名词性短语因子。

以上五个因子中第一、二、四、五因子的抽取与张艳敏等（2015）从144篇议论文提取的四个因子比较吻合，而第三单位结构密度因子的出现很可能是因为口笔语语体差异造成的。五个因子较好地揭示了句法复杂性考察的维度，为下一步样本的比较提供了较为科学的选择依据。

我们选取每个维度中的代表性指标进行了考察。依据前人文献，第一个从属结构比率维度中的C/T（T单位中子句数量），被广泛用于句法复杂性测量。秦晓晴（2007）、鲍贵（2010）均用此指标测量过句子复杂性，并认为C/T能很好地测量与年级水平或教学水平相关的写作能力。基于此，我们选取该指标作为第一维度的比较对象。第三个单位结构密度维度中的MLT（平均T单位长度）也是一项被广泛使用，能体现学习者水平的敏感指标（陈慧媛 2010），故在此被选为第三维度的代表指标。由于中国大学生在写作中简单句使用高达52%，句式结构较为单一，并列结构的使用也可以作为复杂度的考察的一个维度。第四并列短语维度中我们选择CP/T（各T单位中并列短语数）；第五名词性短语维度中的CN/T（各T单位中复杂名词性短语数量）也可以有效地测量学习者名词性结构的使用情况。名词性结构是学习者句法发展的一个阶段性的产物，尤其是语言使用的最高级阶段，为名词化体现的语法隐喻。由于第二维度中均为一些表层的频数指标，且未能很好地反映句法的复杂度，此处则不作考察。

3.2　年级水平和文体如何影响学习者口语产出的句法复杂性

为了建立年级水平对于英语学习者口语产出句法复杂性的正确评价，我们平行比较低年级口语记叙、高年级口语记叙、低年级口语评论、以及高年级口语评论四个语料库在上述C/T，MLT，CP/T，CN/T四个指标上的平均值。试图揭示年级段和文体两个因素对口语产出复杂度的影响。由于两年级水平组分别进行两次不同的数据采集，因此分析上采用2×2双因素混合设计，以此揭示年级和文体两个因素在上述四个指标所代表的四个维度上是如何影响口语句法复杂性的。

表6　各组样本指标值

语料	C/T	MLT	CP/T	CN/T
低年级口语记叙	1.61	12.21	0.22	0.89
低年级口语评论	1.78	13.09	0.24	1.03
高年级口语记叙	1.54	13.14	0.37	1.02
高年级口语评论	1.93	22.78	0.56	2.09

1）C/T（T单位中子句数量）

表7　年级水平和文体影响C/T指标方差分析结果

C/T	自由度	均方	F值	显著性
文体	1，78	3.049	17.726	.000
文体 * 年级水平	1，78	.502	2.918	.092
年级水平	1，78	.073	.415	.522

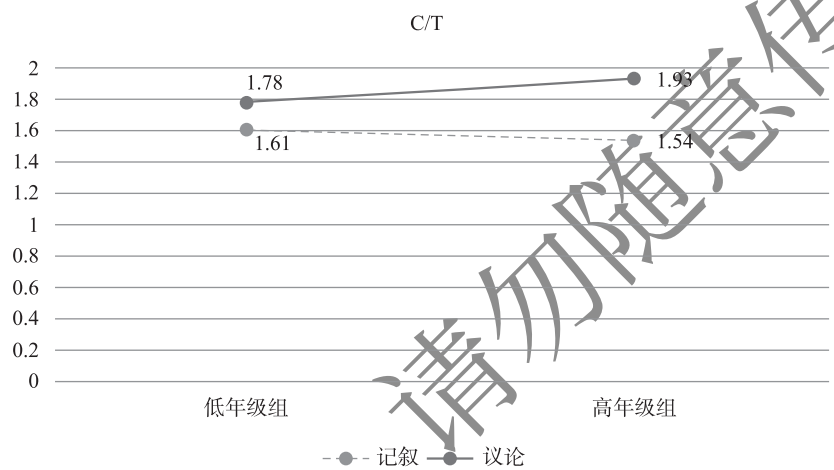


图2　不同年级水平和文体上C/T变化趋势图

由图2可以看出，记叙文体上，高年级学习者口语在各T单位子句数指标上，有着略微下降的水平，而在议论文体上却有着微弱的升高。低年级组在不同文体上C/T指标值区分不大，而到了高年级，文体对C/T值的影响较大，记叙文和议论文明显拉开了差距。方差分析显示：文体因素对于口语产出中C/T值影响明显，主效应显著，F（1，78）=17.726，p<0.01，年级水平没有发现主效应，F（1，78）=0.415<1，p>0.05. 可以认为，在学习者口语产出中，年级水平对各T单位子句数没有影响，而文体影响显著。

前人对写作研究均认定C/T指标会伴随着年级水平呈线性增长。但当涉及中国英语学习者时，秦晓晴（2007）对写作产出的研究发现C/T并非按照线性趋势增长，大一至大四四个阶段经历了升高减少再升高的曲折趋势，且大一水平最低，大二水平达到顶峰。本研究虽然未能跟踪四个年级的指标变化，但从大二、大四的两个水平来看，C/T的变化和秦晓晴对写作的考察结果相似。单看议论性口语的上升趋势和文秋芳、胡健（2010）用LSECCL语料库中224篇议论性口语文本得出的结论相一致。但在记叙文体上，彭慧、卢慧玲（2012）的研究证明二年级到

四年级C/T指标在上升，但却未呈显著性差异，这一发现侧面支撑了本研究中年级水平对C/T指标没有主效应的结果。即便是同年级不同水平口语产出之间，C/T作为子句密度的考量上，也没有显著差异。（鲍贵 2010）

对于记叙性口语产出随着年级增加而T单位中子句比率下降，可能的解释在于低年级水平时T单位的形成主要依赖于从句的使用，而随着学习者语言水平的提高，更倾向于使用短语或紧缩子句而造成从句的使用减少，从而降低了T单位中子句的比率。

2）MLT（平均T单位长度）

表8　年级水平和文体影响MLT指标方差分析结果

MLT	自由度	均方	F值	显著性
文体	1，78	1109.350	25.724	.000
文体 * 年级水平	1，78	768.183	17.813	.000
年级水平	1，78	1127.428	22.255	.000

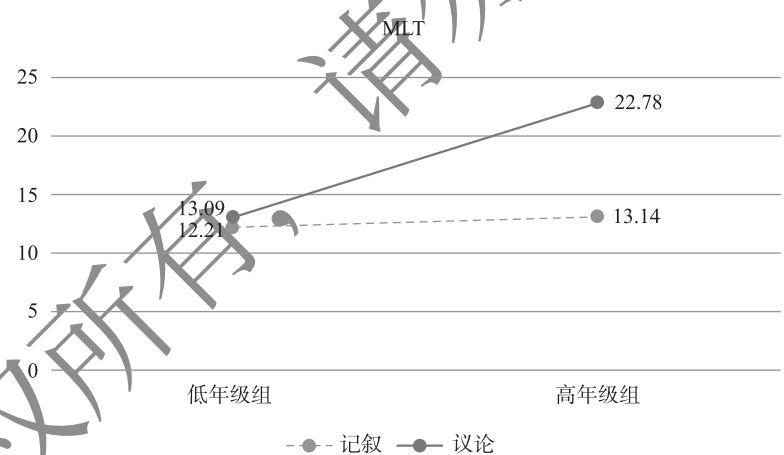


图3　不同年级水平和文体上MLT变化趋势图

和C/T一样，T单位长度也是被广泛用于评价句法复杂性的指标。由图3可以看出，低年级学习者在记叙和议论文体上的口语产出的平均T单位长度区分不大，但随着学习年数的增加，到达高年级时，该指标在两文体上均有升高，且差异急剧增大。尤其是高年级议论性口语的T单位长度已经达到22.78，但在记叙文体上的口语T单位长度仅为13.14，上升轻微。通过双因素混合方差分析，年级水平F（1，78）=22.255，p<0.01，说明年级水平对于平均T单位长度具有主效应；在不同的文体口语产出上，F（1，78）=25.724，p<0.01，主效应显著，且年级水平和

文体之间存在着交互效应，F=17.813，p<0.01。

口语中T单位长度随着年级的升高而增长这一发现与之前文秋芳（2006），鲍贵（2010），彭慧、卢慧玲（2012）的结论相一致，再次佐证了Wolfe-Quintero *et al.*（1998）在对学习者作文的研究中发现：不管水平如何定义，如果不考虑统计结果的显著性以及语言任务和目标语的不同，这两个指标的发展均与水平呈线性关系，即随水平的提高而有不同程度的增加的论断。随着年级的升高，文体差异增大这一现象说明了学习者的语体意识有了进一步的增强，在选择句式方面有了更深刻的考量。学习者随着语言水平的提高以及认知水平的提高，为了表达的更深层次，倾向于使用各种复杂的语法结构。学习者对于T单位结构更深刻地掌握，在议论性口语产出上得到了充分的体现。

而高年级记叙性口语中T单位长度的微弱变化，很大程度上收到了体裁的限制。在记叙文体上，学生可能会更多地倾向于使用简单句和并列句等流水句式。

3）CP/T（T单位中并列短语数量）

表9　年级水平和文体影响CP/T指标方差分析结果

CP/T	自由度	均方	F值	显著性
文体	1，78	.393	5.998	.017
文体 * 年级水平	1，78	.282	4.305	.041
年级水平	1，78	2.192	42.616	.000

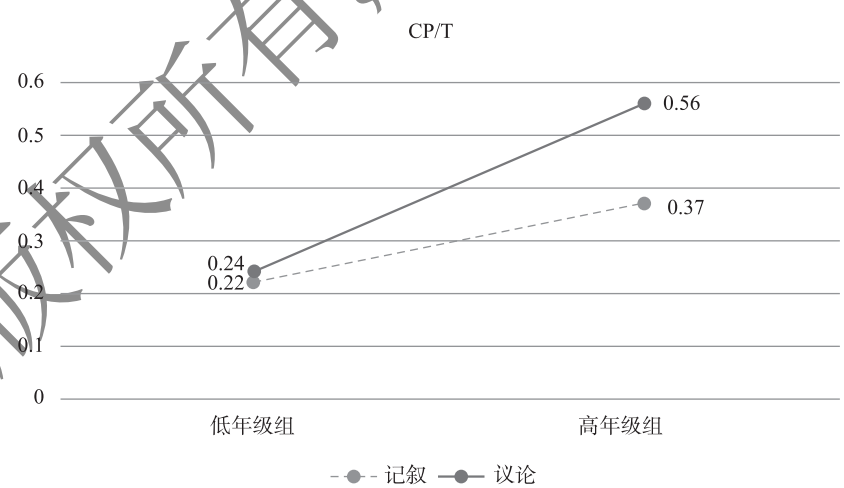


图4　不同年级水平和文体上CP/T变化趋势图

本研究中，并列短语被界定为并列的形容词、副词、名词、以及动词短语

（Lu 2011）。在各类指标中，分别代表着并列短语维度和名词性维度的CP/T和CN/T长期以来未被重视，唯一检验该二指标的学习水平的研究只有Cooper（1976），且认为CP/T和CN/T在临近水平组上区分度较弱，非临近水平组有显著性差异，且一定程度上可以反映发展的阶段性变化（Copper 1976）。而Lu（2011）通过写作文本验证认为CP/T指标不仅在非相邻年级上具有显著性差异，而且在议论文体和记叙文体之间也有显著性差异，属于年级、文体敏感指标。

在口语表现上，由图4可以看出，T单位并列短语数量指标和平均T单位长度一样，在低年级文体区分不明显，随着年级的升高，均表现出增长的趋势，且文体上差异变得显著，议论文体T单位并列短语数量密度更大。说明随着年级的增长，英语学习者对并列短语结构使用更多，掌握得更好，而非Bardovi-Harlig（1992）所说的并列结构指标只适用于测量初级二语学习者句法复杂性。短语结构的使用，尤其是并列短语动词的使用，有效地减少了简单句的使用。图4中的方差分析也显示年级和文体因素对于CP/T的使用均有显著的主效应，文体F（1，78）=5.998，p<0.05；年级F（1，78）=42.616，p<0.01，且两因素具有交互效应，F（1，78）=4.305，p<0.05。

4）CN/T（T单位中复杂名词性短语）

复杂名词性短语在这里界定为1）含有形容词、所有格、介词短语、形容词性子句，分词，以及同位语的名词；2）名词性从句；3）主语中的现在分词和不定式。Lu（2011）指出代表着名词性维度的T单位复杂名词性短语数量在记叙文体和议论文体的口语表达上有着不同的发展模式。由图5可以看到从低年级到高年级，记叙文体上，这一指标只有轻微的增长；而在议论文体上，却有了翻倍的提升。从理论上说，二语学习者语法复杂度的发展顺序可大致描绘为“不完整的句子”⟶“独立的子句”⟶“并列句”⟶“副词性从句”⟶“形容词和名词性从句”⟶“形容词短语、副词短语和名性动词短语”（Wolfe-Quintero *et al.* 1998）。因此本研究中，学习者到了高年级阶段，复杂名词性短语的比率增加恰恰说明了学生句法的成熟性。

再结合方差分析，可以看出文体和年级均影响CN/T的使用，且二因素之间也有交互作用。口语研究的结果和Lu（2011），张丽丽（2016）对写作句法发展的考察相一致，也再次证明CN/T作为句法复杂性指标的敏感性，不仅适用于写作文本，也适用于口语文本，且能有效地预示着学习者句法的发展阶段。

表10　年级水平和文体影响CN/T指标方差分析结果

CN/T	自由度	均方	F值	显著性
文体	1,78	14.710	69.364	.000
文体 * 年级水平	1,78	8.632	40.704	.000
年级水平	1,78	13.840	76.178	.000

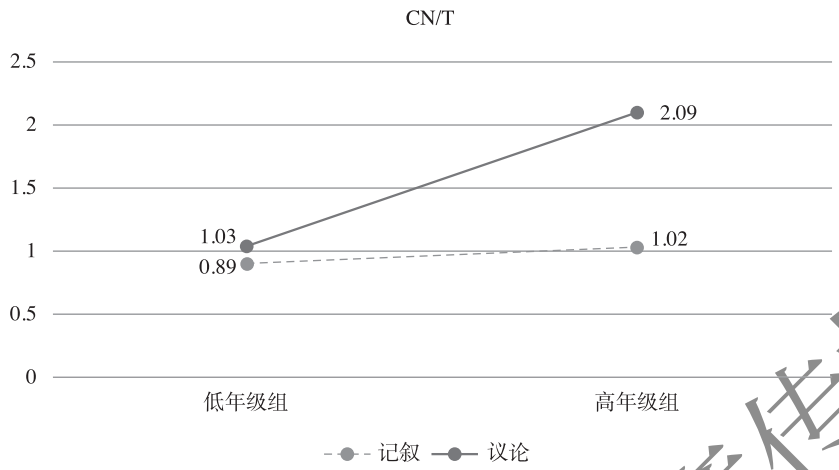


图5　不同年级水平和文体上C/T变化趋势图

4．结论和思考

句法复杂性的测量指标长久以来一直难以达成统一，似乎任何一个指标的使用均有偏颇。本文收集了Wolfe-Quintero *et al.*（1998）推荐最常用的23个测量指标。这些指标几乎都是依然围绕T单位和子句这两个核心结构而衍生出来的。这些指标均或多或少地、直接或间接被前人的研究所认可。在L2SCA的软件帮助下，第一次实现了大规模的文本综合指标分析。本文无意选出最优指标，而是建立指标群的测量体系，建立了五个维度。其中第一维度中的指标紧紧围绕在特定单位中子句等核心结构的比率上。这些比率可以看作是句法复杂性最核心的一个维度，指标间存在高度相关性，且高达33%的解释力对复杂度预测起到了至关重要的作用。另一个核心的维度为第三维度，涵盖了核心结构的单位长度。这两个维度的构建和文秋芳、鲍贵等人常采用的“T单位长度”和“子句密度”不谋而合，从而肯定了该二维度的合理性。第四和第五维度考察了并列结构和名词性结构的使用程度。虽然这两个结构在国外研究中重视不够，但对于中国二语学习者来说却有不同的揭示意义。二语句法发展到高级阶段最终依然要简化句子结构，回归到名词性、紧凑型短语结构的语法隐喻上来。通过因子分析也发现了第二维度中单纯频数的指标，这些指标大多数无法反映句法的嵌套、从属特征，建议在后续的研究中可以规避排除。综上最终建立适用于句法复杂性测量的四大维度分别是“从属结构比率维度”“单位结构长度维度”“并列短语维度”“名词性短语维度”。

从这四个维度框架下考量不同年级大学生句法能力的发展，以及在不同文体下的表现发现：1）除第一个子句密度维度C/T指标以外，后三个维度上，年级水平和文体对于指标MLT、CP/T、CN/T均有显著的主效应，三个指标基本上均呈

现出随着年级提高而逐渐增大的趋势，并且两种文体在高年级时的差异更加显著，议论文体的复杂度明显高于记叙文体；2）代表子句密度维度中的C/T在年级水平上未表现出显著差异，只有文体对其产生主效应；可能的解释是T单位的构成主要依赖于子句，子句的掌握水平在低年级到高年级没有显著的变化，即便T单位长度随着年级水平增长，但作为核心要素的子句也会同比例的增加。随着T单位的增长，其他的成分不可避免地出现了愈来愈多的并列结构和名词结构，而这些结构的使用在一定程度上因替代而抵消掉T单位中子句成分的增加，维持了C/T的稳定性。

中国的高年级大学生由于缺少口语交际环境，因此随着英语水平的增高，口语中具有较强的书面语倾向，不可避免地在论述型口语中显露出议论文书面语的特征。另一方面，记叙文注重故事的描述，而非观点的思辨。在句子成分之间的逻辑关系上并没有很高的要求，学习者在句式选择上，也会采用认知负担较轻的简单句和并列句。此两方面共同导致了高年级议论口语句法较之记叙文体，从属结构使用较多，句式更加复杂的局面。

概言之，中国大学生英语学习者的口语句法复杂性的发展可总结为随年级增长倾向使用更长的复杂句式，擅于采用更密集的并列短语，而名词性结构的使用也加深了句子的紧凑性。大学生的语域意识有所增强。

参考文献

Bardovi-Harlig, K. 1992. A second look at T-unit analysis: Reconsidering the sentence [J]. *TESOL Quarterly* 26(2): 390-395.

Biber, D. & K. Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? [J]. *TESOL Quarterly* 45(1): 5-35.

Cooper, T. 1976. Measuring written syntactic patterns of second language learners of German [J]. *Journal of Education Research* 69(5): 176-183.

Ellis, R. 2009. The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production [J]. *Applied Linguistics* 30(4): 474-509.

Halliday, M. & C. Matthiessen. 1999. *Construing Experience through Meaning: A Language-based Approach to Cognition* [M]. London: Continuum.

Hunt, K. 1970. Do sentences in the second language grow like those in the first? [J]. *TESOL Quarterly* 4(3): 195-202.

Loban, W. 1976. Language Development: Kindergarten through Grade Twelve [R]. Urbana NCTE Committee on Research Report, No. 18.

Lu，X. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development [J]. *TESOL Quarterly* 45(1): 36-62.

Nippold, M., L. Hesketh, J. Duthie & T. Mansfield 2005. Conversational and expository discourse: A study of syntactic development in children, adolescents and adults [J]. *Journal of Speech,*

Language, and Hearing Research 48(5): 1048-1064.

Ortega, L. 1999. Planning and focus on form in L2 oral performance [J]. *Studies in Second Language Acquisition* 21(1): 109-148.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college level of L2 writing [J]. *Applied Linguistics* 24: 429-518.

Skehan, P. 1996. A framework for the implementation of task-based instruction [J]. *Applied Linguistics* 17(1): 38-62.

Skehan, P. & P. Foster. 1997. Task type and task processing conditions as influences on foreign language performance [J]. *Language Teaching Research* 1(3): 185-211.

Wigglesworth, G. 1997. An investigation of planning time and proficiency level on oral test discourse [J]. *Language Testing* 14(1): 85-106.

Wolfe-Quintero, K., S. Inagaki & H. Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity* [M]. Hawai'i: University of Hawai'i Press.

Yuan, F. & R. Ellis. 2003. The effects of pre-task panning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production [J]. *Applied Linguistics* 24(1): 1-27.

鲍　贵，2010，英语学习者语言复杂性变化对比研究 [J],《现代外语》(2)：166-219。

陈慧媛，2010，英语写作表现测量指标的类别及特性研究 [J],《现代外语》(1)：72-80。

李　茶、隋铭才，2017，基于复杂理论的英语学习者口语复杂度、准确度、流利度发展研究 [J].《外语教学与研究》(3)：392-404。

陆小飞、许　琪，2016，二语句法复杂性分析器及其在二语写作研究中的应用 [J],《外语教学与研究》(3)：409-420。

彭　慧、卢慧玲，2012，英语学习者口语句法复杂性发展性研究 [J],《西安外国语大学学报》(1)：72-76.

秦晓晴，2007,《中国大学生英语写作能力发展规律与特点研究》[M]。北京：中国社会科学出版社。

文秋芳、胡　健，2010,《中国大学生口语能力发展的规律与特点》[M]。北京：外语教学与研究出版社。

张丽丽，2016，中国EFL学习者句法复杂性测量研究 [J],《贵州大学学报（社会科学版）》(5)：143-149。

张艳敏、王　涛、侯　旭，2015，基于语料库的二语写作句法复杂性测量因子研究 [J],《天津外国语大学学报》(3)：56-62。

通信地址： 241000 安徽省芜湖市安徽工程大学外国语学院

汉译英新闻语篇时态不一致搭配的介入资源研究

浙江外国语学院 郁伟伟

提要：本研究以汉译英翻译体新闻语篇中间接引语的时态为研究对象，聚焦时态不一致现象的介入资源，应用评价理论中的介入系统对报道者和被报道者的立场进行分析。结果发现，时态不一致搭配的介入资源搭配中，汉译英语主要依靠过去时报道动词与意图绝对时态来表明[对话扩展][对话扩展]的立场，而英语母语主要依靠现在时、现在完成时报道动词与意图相对时态搭配来表明[对话扩展][对话扩展]的立场。汉译英语更频繁地持有中性及消极报道立场，拒绝为表征说话人消极或积极立场承担责任。而英语母语报道者则倾向于采用中性报道立场，承认表征说话人的消极立场是可能的声音之一，并且不反对其他声音介入。汉译英语报道者更善于运用时态不一致搭配这一语言策略采取对话收缩立场，挑战、限制其他声音和立场，进而压缩对话空间，将报道者的态度引入对话空间。汉译英语报道者对于表征说话人明确表明立场、压缩对话空间的策略会利用外部声音采取公开支持的态度和立场，而本族语者则会引用权威声音否定先前引入的肯定命题，同时彰显自身鲜明的消极立场。

关键词：汉译英翻译体新闻语篇、时态不一致现象、介入资源

1. 引言

新闻语篇时态使用的不一致或非连续现象在国内外已有人关注。如Leech & Short（1981：327）、辛斌（2011）称这种现象为间接引语指示中心统一或分离；牛新生（1985）、王伟（1992）描述现在时用于过去报道动词后的现象；Declerck & Tanaka（1996）引入相对时态和绝对时态来探讨这种现象；马景秀（2008）将此现象称为新闻语篇间接引语时态非连续现象；Vandelanotte & Davidse（2009）、Davidse & Vandelanotte（2011）引入以报道说话人的时间零点为参照的第二个指称中心来分析此种现象；赖彦（2014，2015）称其为违背“逆移”规则的时态变异序列情况，而且间接引语时态使用得到越来越广泛关注。

对于时态不一致所起的功能，前人有如下研究。Leech & Short（1981：326-327）认为，这种转述方式既方便了转述者对被转述话语的介入又向读者呈现了被转述话语的一些原有风格，这种做法非常有利于给所转述的话语添上嘲讽的色彩”。最后Leech & Short建议，在考察间接引语指示中心统一或分离时，要综合

考虑“转述者的表义动机和对原话语的理解上加以解释”。

牛新生（1985）、王伟（1992）从增加现场感和活跃氛围来解释现在时用在过去报道动词后的现象。任绍曾（1995）提到了体育报道里现在时态的使用是为了及时向观众传达信息，而不考虑行为的实现，这样减少与观众的距离感。

辛斌（2011）首先从间接引语指示中心的统一和分离入手，讨论了转述者语境和被转述者语境的两个含义，即言语事件和语境定位功能。其次，作者分析了间接引语中的时态组合与指示中心统一，分为与主句转述者时态一致以被转述者的说话时间为参照的相对时态和以实际说话者或转述者为指示中心的绝对时态。第三，作者讨论了间接引语中指示中心的分离及其语义动机，认为有时为“出于某种语义动机经常故意在一定程度上保留被转述者指示中心从而制造一种双声效果”。

赖彦（2015）同样考察了新闻语篇间接转述言语的时体变异，主要考察了转述动词为过去时，从句为一般现在时、现在完成时和现在将来时。从新闻的真实性与一般现在时、新闻的时效性与现在完成体、新闻的客观性与现在将来时、信息突显与时体变异四个方面分析语用动因。通过分析得出，间接转述言语时体变异的动因在于，“句法进入语篇以后，其句法关系的形式规则被打破，而依从语篇层面某些特定语用功能的制约”。

Munday（2012）曾指出，翻译领域对评价性语言并未给予足够重视。有鉴于此，本研究主要运用Martin & White（2005）评价理论介入系统来分析报道动词和从句动词，进而分析时态不一致搭配所起到的功能。

2. 理论基础

Martin & White（2005：94）的评价理论与协商系统（negotiation）和参与系统（involvement）一起，构成了语篇语义学层面表达人际意义的三个系统，主要包括三部分：态度（attitude）、介入（engagement）和级差（graduation）。其中介入主要指：在当前交际语境中作者声音融入其他声音的所有言语手段。归属进一步分为间接引语中的承认（acknowledge）和疏远（distance）。主体间立场（intersubjective stance）由作者/说话人对文本内真值的协商及作者/说话人将虚拟受话人写进文本的手段组成。此过程被称为“固化”（solidarity）。因此对齐（alignment）和偏离（disalignment）主要指价值立场选择上的一致或分歧。介入分为否认（disclaim）、公告（proclaim）、接纳（entertain）和归属（attribute）。根据它们是否指代其他声音将其分为单声（monogloss）和多声（heterogloss），多声分为对话收缩（contraction）或者称之为收缩对话空间以阻挡潜在的导致不足信的其他态度，如报道动词：demonstrate或show。而对话扩展（expansion）指相

反的情形，即开放对待疑问或对话的其他可能性，如：与主体间（intersubjective）功能有关的claim。接纳和归属属于对话扩展，而否认和公告属于对话收缩。接纳指作者的声音表明为其他声音开放对话空间，或者作者的立场只是众多可能性中的一种的情景，通过使用“情态助动词、情态附加语、情态定语及in my view这类情形，及特定的心理动词、附加语投射”（同上：105）。在这种用法中，说话人/作者暗示了价值立场的可能性或者是对真值的不确定性或对真值不负责任。归属下面的承认指没有明显的立场或价值判断的情形，如：report, believe, say, according to。与此相对的是疏远，它使得作者的声音疏远外部的声音，如：claim。承认的次范畴为读者“呈现了一个相对无人称的或中立的场景”（同上：115）。对话收缩的范畴包含否认和公告。否认又进一步分为否定（deny）和反对（counter）。否定预设了一些读者持有相反的观点，为了争取他们，或者把他们与作者一方结盟，作者/说话人首先引入积极的观点然后否定它，有时会与带有虚拟受话人的第三方相对。反对也激发了一个相反的命题，此命题随后如同否认一般不再被坚持。另一方面，否认限制了对话选择，公告次范畴包括同意（concur），宣告（pronounce）和背书（endorse）。同意指一种作者-读者关系，指和说话人有同样的了解，通过of course，naturally，not surprisingly，admittedly and certainly等话语明显地同意对话同伴或虚拟读者。背书讨论了说话人与外部声音的赞同关系，用类似show，prove，demonstrate、find and point out等表明事实性的词汇。宣告指的作者明显的干预，他们挑战、对抗或是抵制这个特定的对话选择。宣告的实例可以进一步划分为主观的、客观的及明显的和暗含的，明显主观化的例子如：it is absolutely clear to me that，明显客观化的例子如：the facts of the matter are that，暗含的主观化的例子如：not，暗含的客观化的例子如：really。

本研究将上述评价理论介入系统用来分析间接引语时态不一致所承担的功能，分别从上述各个范畴和次范畴对部分语料进行文本分析，以从功能角度对间接引语时态不一致现象作一分析，并找出背后所隐含的作者或报道者的意图。

3. 研究语料和方法

本研究基于的汉英平行语料库，取自Yixiao Ma创建并分别于2007年9月17日、2008年7月16日及2008年9月15日发布的“GALE一期汉语广播新闻平行文本-第一部分、第二部分及第三部分”（Ma & Strassel 2007，2008a，2008b），包含中国中央电视台（CCTV）-中国大陆的电视台和凤凰卫视（设在香港的卫星电视台）的23.3、21.9及19.1小时的汉语广播新闻节目转写及相应英文译文，二者播报形式和内容时段基本对应。所有音频文件由LDC标注器及专业转写机构完成。人工句子单位/语段标注也算作转写工作的一部分。最后识别出三类句子单位：（1）陈述句子单位；（2）疑问句子单位；（3）不完整句子单位。转写和句子

单位标注后，文档重新调整为人可读的格式，由专业译者进行仔细翻译。译者遵守LDC的GALE翻译指导原则，旨在描写翻译团队的构成、源数据格式、翻译数据格式、翻译特定语言特征（如名字和言语不流利）的最优译法以及对完成译文的质量控制步骤。语料代表了聚焦当前事件的新闻节目，如2004年和2005年度CCTV的“每日新闻”、凤凰卫视的“全球大视野”和“新闻早班车”。（见表1）

表1　GALE汉英平行语料库CCTV新闻及CNN本族语语料库构成

来源	节目名称	时间跨度
CCTV4	CCTV4 Daily news	2004.06—2006.11
CCTV4	CCTV4 news	2005.04—2006.01
Phoenix	GLOBAL REPORT	2005.04—2005.12
Phoenix	GOOD MORNING CHINA	2005.09—2006.01
CNN	Situation Room	2006.12—2007.12

其中，第一部分包含38个文本，351,594汉字，译文240,481词次。第二部分包含38个文本，328,760 汉字，译文含有227,532词次。第三部分包含34个文本，227,330汉字，相应译文包含160,040词次。平行语料库三部分共计包含907,654汉字和628,053英文单词。

4．结果和分析

汉译英语从句动词的介入资源共出现583次，见表2，其中包括71种对话收缩资源和512种对话扩展资源，按照频次由高到低依次为：承认（464）、背书（61）、接纳（40）、疏远（8）、否定（5）、反对（3）、同意（2）。英语母语从句动词的介入资源共出现125次，其中对话收缩资源有23种，对话扩展资源有102种，按照频次由高到低依次为：承认（88）、背书（13）、接纳（10）、疏远（4）、否定（7）、反对（3）。汉译英语在从句动词介入资源总数（$p = 0.000 < 0.05$）、对话收缩（$p = 0.000 < 0.05$）和对话扩展（$p = 0.000 < 0.05$）资源频数上都显著多于英语母语，英语母语仅在否定次范畴上多于汉译英语，但不存在显著差异，并且在否定、反对、同意、疏远次范畴上数量相当。可见无论是报道动词还是从句动词，介入资源排序都是承认、背书、接纳、疏远、否定、反对及同意。这与新闻语篇追求客观公正性不谋而合。接下来，我们将按照出现频率排序分析介入资源与时态分布的共现关系。

表2　汉译英语和英语母语特有报道动词与从句动词的介入资源频数比较

			汉译英语特有报道动词	英语母语特有报道动词	显著性	汉译英语特有从句动词	英语母语特有从句动词	显著性
介入资源			105	27	0.000*	583	125	0.000*
对话收缩	否认	否定	3	0	-	5	7	0.564
		反对	2	1	0.559	3	3	0.999
	宣称	同意	0	0	-	2	0	-
		宣称	0	0	-	0	0	-
		背书	24	6	0.001*	61	13	0.000*
对话扩展	接纳		21	5	0.001*	40	10	0.000*
	归属	承认	49	15	0.000*	464	88	0.000*
		疏远	6	0	-	8	4	0.243

4.1　报道动词和从句动词介入资源搭配类型

我们将在本节分析17种报道动词和从句动词时态不一致搭配中介入资源的出现规律。本节讨论中，介入资源搭配由两部分构成，前面中括号代表报道动词介入资源，表明的是作者或报道者的立场，后面中括号代表从句动词介入资源，表明的是表征说话人的立场。依据评价理论，接纳和归属中的承认和疏远属于对话扩展，而否认中的否定和反对，公告中的同意、宣告和背书属于对话收缩。根据介入资源搭配类型，可分为[对话扩展][对话扩展]、[对话扩展][对话收缩]、[对话收缩][对话扩展]、[对话收缩][对话收缩]。同时，我们可以把介入资源根据报道动词是否有报道者或表征说话人明显的立场分为积极的、中性的和消极的三类，积极地表明立场的有：背书、同意；中性的并未显露立场，表示可能有其他可能性的有：承认、接纳；明确表达消极立场的有：否定、反对、疏远。接下来将按上面两种分类方法对汉译英语和英语母语共有的和特有的各时态不一致搭配类型进行讨论。

4.1.1　[对话扩展][对话扩展]

对话扩展指作者运用言语策略引发、容纳其他声音和立场，开启对话空间（岳颖 2011：31）。这里，我们考察报道者采取对话扩展立场，表征说话人采取对

话扩展立场的情形，由前文可知，对话扩展具体包括以下搭配类型：接纳、承认和疏远，因此下文将讨论各种不一致时态搭配中出现的[对话扩展][对话扩展]的8种搭配，分别为：[承认][承认]、[承认][接纳]、[承认][疏远]、[接纳][承认]、[接纳][接纳]、[疏远][承认]、[疏远][接纳]和[疏远][疏远]。

[承认][承认]搭配中报道者和表征说话人共同持有中性立场，分布规律如下，汉译英语中出现频数显著多于英语母语的情形出现在以下时态搭配类型：过去时报道动词和意图绝对现在时从句动词搭配、过去时报道动词和意图绝对过去时从句动词搭配、过去时报道动词和意图现在完成时从句动词搭配、过去时报道动词和意图将来时从句动词、过去时报道动词和真正现在时从句动词搭配。英语母语显著多于汉译英语的情形出现在以下时态搭配类型：现在时报道动词与意图过去完成时从句动词、现在时报道动词和意图过去将来时从句动词。仅在汉译英语以下时态搭配类型中出现的搭配包括：过去时报道动词和真正现在完成时从句动词（2例）和过去时报道动词和真正将来时从句动词搭配（3例）。仅在英语母语以下时态搭配类型中出现的搭配包括：现在时报道动词和意图过去将来完成时从句动词搭配（2例）、现在完成时报道动词与意图相对过去时从句动词搭配（3例）和过去将来时报道动词与意图绝对现在时搭配（2例）。

[承认][接纳]这种搭配表达报道者对表征说话人观点的支持，此时表征说话人采取的立场为：承认命题内容仅是一种可能的情况，不排除有其他声音的存在。其中汉译英语中出现频数显著多于英语母语的包括以下时态搭配类型：过去时报道动词和意图绝对现在时从句动词。仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词和意图现在完成时从句动词、过去时报道动词和意图将来时从句动词（8例）、过去时报道动词和真正现在时从句动词（3例）和过去完成时报道动词与意图现在时从句动词（1例）。仅在英语母语以下时态搭配类型中出现，包括：现在完成时报道动词和意图相对过去时从句动词搭配（1例）。

[承认][疏远]搭配表明报道者持支持立场，而表征说话人持疏远态度，避免对命题内容承担责任。此种搭配仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（1例）和过去时报道动词和意图绝对过去时从句动词（6例）。此外，同时出现在过去时报道动词和意图将来时从句动词搭配的[承认][疏远]搭配无显著差异。

[接纳][承认]搭配表达的是报道者持有开放对话空间的态度，承认表征说话人的命题内容仅是一种可能性，还存在其他声音，持中立态度。而从句表征说话人持有的是肯定命题内容的立场。其中汉译英语显著多于英语母语的情况出现在以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词、过去时报道动词与意图绝对过去时从句动词、过去时报道动词与意图现在完成时从句动词、过去时报道动词和意图将来时从句动词。

[接纳][接纳]指报道者和表征说话人都持对话开放的态度，都认为命题内容仅是一种可能的情形，承认其他声音的存在。其中汉译英语显著多于英语母语的情况出现在以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词。仅在汉译英语以下时态搭配类型中出现，包括：现在时报道动词与意图相对过去时从句动词。

[疏远][承认]搭配指在表征说话人支持命题内容正确性的情况下，报道者持疏远立场，避免对命题内容真实性承担责任。其中汉译英语显著多于英语母语的情况出现在以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词、过去时报道动词和意图绝对过去时从句动词、过去时报道动词与意图现在完成时从句动词。[疏远][承认]仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词和意图将来时从句动词。仅在英语母语以下时态搭配类型中出现，包括：现在时报道动词与意图相对过去时从句动词（1例）。

[疏远][接纳]指表征说话人承认命题仅是一种可能性，还存在其他声音，持中立态度，而报道者并不赞同，采取疏远态度。仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（3例）。

[疏远][疏远]指表征说话人对命题内容持疏远态度，而报道者对命题内容持疏远态度，并隐含了不同意表征说话人的态度。仅在汉译英语以下时态搭配类型中出现，为现在时报道动词与意图过去完成时从句动词（1例）。

4.1.2 [对话扩展][对话收缩]

对话收缩指作者通过言语策略挑战、限制其他声音与立场，压缩对话空间（岳颖 2011：31）。由前文可知，否定、反对、同意和背书属于对话收缩，因此下文将讨论各种不一致时态搭配中出现的[对话扩展][对话收缩]的9种搭配：[承认][否定]、[承认][反对]、[承认][同意]、[承认][背书]、[接纳][否定]、[接纳][背书]、[疏远][否定]、[疏远][反对]和[疏远][背书]。

[承认][否定]指表征说话人引入命题内容，并进一步持否定态度，报道者赋予表征说话人态度正确性，提高可信度和可靠性。英语母语显著多于汉译英语的情况出现在以下时态搭配类型：过去时报道动词与意图绝对现在时从句动词。仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词和意图现在完成时从句动词（3例）。仅在英语母语以下时态搭配类型中出现，包括：过去时报道动词和意图绝对过去时从句动词（6例）、现在时报道动词与意图过去完成时从句动词（1例）和现在时报道动词和意图过去将来时从句动词（5例）。

[承认][反对]搭配指表征说话人引入并承认一种肯定的观点，但在语篇发展的过程中提出合理的理据否认命题，以达到争取读者的目的。此搭配中报道者持有肯定并支持表征说话人态度的立场。[承认][反对]搭配仅在本族语以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（1例）。出现

在其他时态类型中的汉译英语和英语母语不存在显著差异。

[承认][同意]搭配指表征说话人对命题内容公开表明同意态度，而报道者也暗中支持这一立场。此搭配仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词和意图现在完成时从句动词（7例）。仅在英语母语以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（1例）。

[承认][背书]搭配表征说话人通过外部声音说明所述命题正确、理据充分或不容置疑，同时也映射了表征说话人自己的声音，此时报道者持中性的肯定态度，如消息来源是权威人士，则报道者认为命题内容是可靠真实的。其中在汉译英语中显著多于英语母语的情况出现在以下时态搭配类型中，包括：过去时报道动词与意图绝对现在时从句动词、过去时报道动词和意图绝对过去时从句动词、过去时报道动词和意图现在完成时从句动词、过去时报道动词和意图将来时从句动词。仅在汉译英语以下时态搭配类型中出现，包括：现在时报道动词与意图过去完成时从句动词。仅在英语母语以下时态搭配类型中出现，包括：现在时报道动词和意图过去将来时从句动词（3例）。

[接纳][否定]搭配指表征说话人把肯定命题引入语篇，然后以权威声音否定其可能性。而报道者对此持保留态度，承认表征说话人对命题内容的态度只是一种可能性，可能存在其他声音。仅在英语母语以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（1例）、过去时报道动词和意图将来时从句动词（2例）。

[接纳][背书]搭配指表征说话人通过外部声音说明所述命题正确、理据充分或不容置疑，同时也映射了表征说话人自己的声音，而报道者对此持保留态度，承认表征说话人对命题内容的态度只是一种可能性，可能存在其他声音。仅在汉译英语中以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（3例）、过去时报道动词与意图现在完成时从句动词（1例）、过去时报道动词与意图将来时从句动词（3例）。仅在英语母语中以下时态搭配类型中出现，包括：过去时报道动词和意图绝对过去时从句动词（1例）。

[疏远][否定]搭配指表征说话人把肯定命题引入语篇，然后以权威声音否定其可能性，而此时报道者对此持疏远态度，避免对表征说话人立场承担责任。仅在汉译英语的过去时报道动词与意图绝对现在时从句动词搭配中出现1例。

[疏远][反对]指表征说话人引入并承认一种肯定的观点，但在语篇发展的过程中提出合理的理据否认命题，以达到争取读者的目的。此搭配中报道者持疏远的立场，尽量避免承担责任。[疏远][反对]仅在过去时报道动词和意图现在完成时从句动词搭配中出现1例。

[疏远][背书]搭配指表征说话人通过外部声音说明所述命题正确、理据充分或不容置疑，同时也映射了表征说话人自己的声音，而报道者对此持疏远态度，

尽量避免承担责任。仅在汉译英语中以下时态搭配类型中出现，包括：过去时报道动词和意图绝对过去时从句动词（3例）、过去时报道动词和意图现在完成时从句动词（1例）、过去时报道动词和意图将来时从句动词（2例）。

4.1.3 [对话收缩][对话扩展]

[对话收缩][对话扩展]主要是报道者运用报道动词压制对话空间，避免其他声音的介入，表征说话人运用从句动词开启对话空间，通过这种方式，报道者从中发出自己的声音，表明自己的态度。接下来分析8种[对话收缩][对话扩展]搭配类型：[背书][承认]、[同意][承认]、[否定][承认]、[反对][承认]、[背书][接纳]、[同意][接纳]、[背书][疏远]和[否定][疏远]。

[背书][承认]搭配表征说话人虽然没有明确表明立场，但会通过间接方式赋予命题可信与否的性质，若源于权威人士，则可靠性提高，相反则降低（岳颖2011：33）。此时报道者通过外部声音说明所述命题正确、理据充分或不容置疑。其中汉译英语频次显著多于英语母语的情况主要出现在以下时态搭配中，包括：过去时报道动词与意图绝对现在时从句动词、过去时报道动词和意图绝对过去时从句动词、过去时报道动词和意图现在完成时从句动词、过去时报道动词和意图将来时从句动词、过去时报道动词和真正现在时。仅在汉译英语中以下时态搭配类型中出现，包括：过去时报道动词和真正现在完成时从句动词（3例）和过去时报道动词和真正将来时从句动词搭配（1例）。

[同意][承认]指表征说话人开放对话空间，承认多种声音的存在，而报道者对命题内容公开表明同意态度。仅在汉译英语中以下时态搭配类型中出现，包括：过去时报道动词和意图现在完成时从句动词（2例）、过去时报道动词和意图将来时从句动词（4例）。

[否定][承认]指表征说话人开放对话空间，承认多种声音的存在，而报道者把肯定命题引入语篇，然后以权威声音否定其可能性，是报道者表明自己立场的手段。其中汉译英语显著多于英语母语的情况存在于以下时态搭配类型中，包括：过去时报道动词与意图绝对现在时从句动词。仅在汉译英语以下时态搭配类型中出现的包括：过去时报道动词和意图绝对过去时从句动词（6例）、过去时报道动词和意图将来时从句动词（1例）、现在完成时与意图过去完成时的介入（1例）和现在完成时报道动词与意图相对过去时从句动词（1例）。

[反对][承认]指表征说话人开放对话空间，承认多种声音的存在，而报道者引入并承认一种肯定的观点，但在语篇发展的过程中提出合理的理据否认命题。仅在汉译英语以下时态搭配类型中出现，包括：过去时报道动词与意图绝对现在时从句动词（1例）和过去时报道动词和意图将来时从句动词（1例）。仅在英语母语以下时态搭配类型中出现，包括：过去时报道动词和意图绝对过去时从句动词（1例）。

[背书][接纳]指表征说话人承认命题仅是一种可能性，还存在其他声音，持中立态度，而报道者通过外部声音说明所述命题正确、理据充分或不容置疑。其中汉译英语显著多于英语母语的情况出现在以下时态搭配类型中，包括：过去时报道动词与意图绝对现在时从句动词。仅出现在汉译英语中以下时态搭配类型，包括：过去时报道动词和意图现在完成时从句动词（1例）和过去时报道动词和意图将来时从句动词（1例）。仅出现在英语母语中以下时态搭配类型中，包括：现在时报道动词与意图相对过去时从句动词（1例）。

[同意][接纳]指表征说话人承认命题仅是一种可能性，还存在其他声音，持中立态度，而报道者对命题内容公开表明同意态度。仅出现在汉译英语中以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词（1例）。

[背书][疏远]指表征说话人与命题之间的距离，与外部声音不一致，拒绝承担命题责任（岳颖 2011：34），而报道者通过外部声音说明所述命题正确、理据充分或不容置疑。仅出现在汉译英语中以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词（1例）、过去完成时报道动词和意图绝对过去时从句动词（1例）。

[否定][疏远]指表征说话人与命题之间的距离，与外部声音不一致，拒绝承担命题责任，而报道者把肯定命题引入语篇，然后以权威声音否定其可能性，是报道者表明自己立场的手段。仅出现在汉译英语中以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词（1例）。

4.1.4 [对话收缩][对话收缩]

汉译英语和英语母语[对话收缩][对话收缩]介入搭配共有5种类型：[背书][背书]、[否定][背书]、[背书][同意]、[背书][否定]和[否定][否定]。

[背书][背书]指表征说话人和报道者通过外部声音说明所述命题正确、理据充分或不容置疑。其中在汉译英语中显著多于英语母语的情况出现在以下时态搭配类型中，包括：过去时报道动词与意图绝对过去时从句动词。仅出现在汉译英语中以下时态搭配类型，包括：过去时报道动词与意图绝对现在时从句动词（4例）、过去时报道动词和意图现在完成时从句动词（2例）。

[否定][背书]指表征说话人通过外部声音说明所述命题正确、理据充分或不容置疑，而报道者把肯定命题引入语篇，然后以权威声音否定其可能性，是报道者表明自己立场的手段。仅出现在英语母语以下时态搭配中，为过去时报道动词和意图绝对过去时从句动词（1例）。

[背书][同意]指表征说话人对命题内容公开表明同意态度，而报道者通过外部声音说明所述命题正确、理据充分或不容置疑。仅出现在汉译英语以下时态搭配中，为过去时报道动词和意图现在完成时从句动词（3例）。

[背书][否定]指表征说话人把肯定命题引入语篇，然后以权威声音否定其可能性，是报道者表明自己立场的手段，而报道者通过外部声音说明所述命题正确、理据充分或不容置疑。仅出现在汉译英语以下时态搭配中，包括：过去时报道动词与意图绝对现在时从句动词（1例）、过去时报道动词和意图绝对过去时从句动词（1例）、过去时报道动词和意图将来时从句动词（1例）。

[否定][否定]指表征说话人和报道者把肯定命题引入语篇，然后以权威声音否定其可能性，是表明自己立场的手段。仅出现在英语母语中以下时态搭配类型中，为过去时报道动词与意图绝对现在时从句动词（2例）。

5. 结语

本研究以汉译英翻译体新闻语篇中间接引语的时态不一致现象为研究对象，聚焦时态不一致现象的人际功能，应用评价理论中的介入系统对报道者和被报道者的立场进行分析。对于时态不一致的介入资源，我们发现如下规律：

（1）通过分析汉译英语和英语母语时态搭配类型与[对话扩展][对话扩展]的对应关系得出结论，汉译英语在以下时态不一致类型的介入资源搭配中显著多于英语母语，包括：过去时报道动词与意图绝对时态搭配的[承认][承认]、[承认][接纳]、[承认][疏远]、[接纳][承认]、[接纳][接纳]、[疏远][承认]、[疏远][接纳]。英语母语在以下时态不一致类型的介入资源搭配中显著多于汉译英语，包括：现在时、现在完成时报道动词与意图相对时态搭配的[承认][承认]、[承认][接纳]、[疏远][承认]。另外，[接纳][接纳]、[疏远][疏远]在汉译英语的现在时报道动词与意图相对时态搭配中也存在显著优势。[承认][承认]在英语母语过去将来时报道动词与意图绝对时态的搭配中也存在显著优势。以上表明对于汉译英语主要依靠过去时报道动词与意图绝对时态来表明[对话扩展][对话扩展]的立场，而英语母语主要依靠现在时、现在完成时报道动词与意图相对时态搭配来表明[对话扩展][对话扩展]的立场。

（2）通过分析汉译英语和英语母语时态搭配类型与[对话扩展][对话收缩]的对应关系得出结论，汉译英语在以下时态的介入资源搭配中占有优势：[疏远][反对]、[疏远][背书]、[承认][背书]、[接纳][背书]、[疏远][否定]、[承认][否定]。仅出现在英语母语的搭配包括：[承认][反对]、[接纳][否定]。因此汉译英语更频繁地持有中性及消极报道立场，拒绝为表征说话人消极或积极立场承担责任。而英语母语报道者则倾向于采用中性报道立场，承认表征说话人的消极立场是可能的声音之一，并且不反对其他声音介入。

（3）通过分析汉译英语和英语母语时态搭配类型与[对话收缩][对话扩展]的对应关系得出结论，汉译英语较英语母语有绝对优势，其中[背书][承认]、[同意]

[承认]、[否定][承认]、[同意][接纳]、[背书][疏远]、[否定][疏远]仅出现在汉译英语中，[反对][承认]、[背书][接纳]也仅在过去时报道动词和意图绝对过去时从句动词以及现在时报道动词与意图相对过去时从句动词搭配中各出现1例。因此，汉译英语报道者更善于运用时态不一致搭配这一语言策略采取对话收缩立场，挑战、限制其他声音和立场，进而压缩对话空间，将报道者的态度引入对话空间。

（4）通过分析汉译英语和英语母语时态搭配类型与[对话收缩][对话收缩]的对应关系得出结论，汉译英语较英语母语有绝对优势的介入资源搭配包括：[背书][背书]、[背书][同意]和[背书][否定]，而本族语特有的时态不一致介入资源搭配包括：[否定][背书]和[否定][否定]。这说明汉译英语报道者对于表征说话人明确表明立场、压缩对话空间的策略会利用外部声音采取公开支持的态度和立场，而本族语者则会引用权威声音否定先前引入的肯定命题，同时彰显了自身鲜明的消极立场。

Martin & White（2005：94）的评价理论里指出介入系统指“为作者置身于与当前交际语境中的其他声音和立场比肩并‘介入’其内而提供手段的所有话语总和”。因此时态也是一种评价资源，用以承载态度或立场，但在评价理论里，用墨不多。从人际功能中的介入功能探索报道动词和从句动词时态不一致所带有的报道者立场，得出翻译体英语和本族语英语中，除意图绝对过去时、意图将来时及本族语中的意图绝对现在时持报道者消极立场外，其他都主要承担表明标明报道者积极或中性立场责任的结论，扩展和加深了评价资源的应用。

参考文献

Davidse, K. & L. Vandelanotte. 2011. Tense use in direct and indirect speech in English [J]. *Journal of Pragmatics* 43(1): 236-250.

Declerck, R. & K. Tanaka. 1996. Constraints on tense choice in reported speech [J]. *Studia Linguistica* 50(3): 283-301.

Leech, G. & M. Short. 1981. *Style in Fiction: A Linguistic Introduction to English Fictional Prose* [M]. London: Longman.

Ma, X. & S. Strassel. Gale phase 1 Chinese broadcast news parallel text - part 1 ldc2007t23 2007 [OL]. https://catalog.ldc.upenn.edu/LDC2007T23.（2019-05-15读取）

Ma, X. & S. Strassel. Gale phase 1 Chinese broadcast news parallel text - part 2 ldc2008t08 [OL]. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T08.（2019-05-15读取）

Ma, X. & S. Strassel. Gale phase 1 Chinese broadcast news parallel text - part 3 ldc2008t08 [OL]. https://catalog.ldc.upenn.edu/LDC2008T18.（2019-05-15读取）

Martin, J. & P. White. 2005. *The Language of Evaluation: Appraisal in English* [M]. Houndmills, Basingstoke: Palgrave Macmillan.

Munday, J. 2012. *Evaluation in Translatino: Critical Points of Translator Decision-making* [M]. London: Routledge.

Vandelanotte, L. & K. Davidse. 2009. The emergence and structure of *be like* and related quotatives: A constructional account [J]. *Cognitive Linguistics* 20(4): 777-807.

赖　彦，2014，新闻报道语篇的时态序列变异及认知阐释 [J],《浙江传媒学院学报》(6)：121-126。

赖　彦，2015，新闻语篇间接转述言语的时体变异 [J],《外语与外语教学》(1)：19-25。

马景秀，2008，英语新闻语篇间接引语时态非连续性现象刍议 [J],《西安外国语大学学报》(4)：18-21。

牛新生，1985，新闻报道中间接引语时态变化不规则现象初探 [J],《外国语文教学》(3)：136-138。

王　伟，1992，英语新闻广播中的时态不一致现象 [J],《解放军外语学院学报》(6)：9-10。

辛　斌，2011，间接引语指示中心的统一和分离：认知符号学的视角 [J],《外语研究》(3)：7-11。

岳　颖，2011，学术语篇的介入资源与人际意义构建——给予学习者语篇的个案研究 [J],《当代外语研究》(7)：30-35。

通信地址： 310023 浙江省杭州市浙江外国语学院英语语言文化学院

基于语料库的作家作品词汇风格分析——以茅盾、巴金、老舍为例

上海交通大学 陈好修

提要：本文以现代文学第二个十年（1928—1937）中的三位代表作家——茅盾、巴金、老舍的作品为研究对象，通过对三个20万字作品文本库的统计分析，得出三位作家文本中词长、型例比、单现词、指示代词、关键词等维度的差异。在数据分析结果的基础上，尝试原因分析，探究造成三位作家词汇使用差异的时代文化因素。

关键词：风格、计算风格学、语料库、分词、词频统计

1. 引言

语言风格是运用语言表达形成的综合效果，包括语言的民族风格、时代风格、流派风格等（黎运汉 1990）。风格在语言层面表现为各种语言成分的使用频次。基于语料库并运用统计的方法进行语言计量特征分析是语言风格研究的重要方法。

茅盾、巴金、老舍这三位作家是现代文学史第二个十年的代表性人物，他们的作品一直深受读者喜爱。语料库的兴起为汉语风格研究提供了强有力的手段。运用定量分析将语言结构成分量化，能为感性的阅读体验提供一种理性、直观的解释，并发现不同作品之间新的、细微的差异。将语料库语言学应用到更多与语言相关的交叉学科中，是风格研究领域一块待耕耘的领域。

2. 本文研究思路及方法

本研究分五步：

第一步，构建作家作品文本库。

选取同时代的三位作家茅盾、巴金、老舍不同时代的小说各20万字，构建共计60万字的平衡语料库。我们综合考虑了作家的生活时代以及作品的体裁。老

舍、茅盾、巴金三位作家均出生于1900年前后，在现代文学分期中，他们都属于第二个十年（1928—1937）的代表性人物。在作品的体裁方面，选取的都是小说。

每位作家的作品语料选取情况如下：

1. 茅盾（1896—1981）

茅盾原名沈德鸿，字雁冰，浙江桐乡乌镇人。茅盾是现代文学第二个十年间极具代表性的作家。其凭借反映社会现实的“社会剖析小说”在20世纪30年代开创了新的文学范式。

文本库选取了茅盾30年代创作的小说《子夜》前10万字，40年代的小说《腐蚀》10万字。

2. 巴金（1904—2005）

原名李尧棠，字芾甘，四川成都人。他被誉为五四运动以来最有影响力的作家之一。

巴金的创作可以分为前后两个时期。1928年—抗日战争，他的作品被称为“青春的赞歌”，贯穿着无政府主义思想。抗日战争之后，巴金的作品体现出了深沉的悲剧艺术，并开始带有现实主义特色。

文本库选取了巴金20世纪30年代的小说《家》前10万字，40年代的小说《寒夜》10万字。

3. 老舍（1899—1966）。

原名舒庆春，字舍予，满族正红旗，北京人，被誉为“人民艺术家”。老舍的创作分期比较明显，可以分为20年代，30年代，40年代，50、60年代四个时期。文本库选取了老舍20年代的小说《老张的哲学》5万字，30年代的小说《骆驼祥子》5万字，40年代长篇《四世同堂》5万字，50、60年代的小说《正红旗下》5万字。

第二步，自动分词及词性标注。

用ICTCLAS工具处理所得语料，获得自动分词及词性标注结果，以词/词性的格式，给每个词打上标签。处理后的文本如下文所示：

> 我们/r 所/u 要/v 介绍/v 的/u 是/v 祥子/n ，/w 不/d 是/v 骆驼/n ，/w 因为/c“/w 骆驼/n ”/w 只/d 是/v 个/q 外号/n ；/w 那么/c ，/w 我们/r 就/d 先/d 说/v 祥子/n ，/w 随手/d 儿/k 把/p 骆驼/n 与/c 祥子/n 那/r 点/m 关系/n 说/v 过去/v ，/w 就/d 算/v 了/y 。/w ——老舍《骆驼祥子》

用ICTCLAS工具（张华平、商建云 2019）处理后得到三个文件：1. maodun.seg.txt；2. bajin.seg.txt；3. laoshe.seg.txt。

第三步，获取词频。

运用python分别统计各作家作品的词语列表，包括每个词语的词性及各自的使用频率，按照从大到小的顺序排列，得到三个文件：1. maodunpilv.txt；2. bajinpinlv.txt；3. laoshepinlv.txt。

第四步，数据计算。

运用Excel应用程序进行相关数据的筛选、计算、统计。

第五步，数据分析。

流程图示如下：

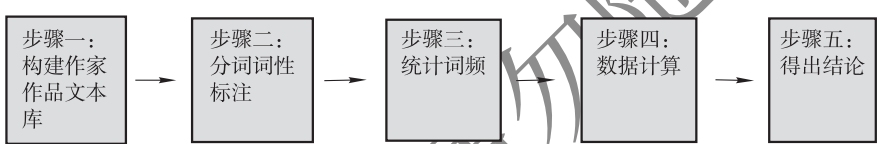


图1　文章思路及研究方法图解

3. 茅盾、巴金、老舍作品风格对比分析

我们选取了词长、型例比、单现词、人称代词、指示代词、关键词这六个语言维度，分别对老舍、茅盾、巴金三位作家的作品进行了比较分析。

各语言维度的意义、数据获得方法以及计算公式[1]：

1. 词长

词长是文本长度和词数的比值，其大小可以反映出语言使用者用词的长度。该数据通过计算公式计算获得。（计算公式详见下文，下同。）

2. 型例比

型例比即形符和类符比值，形符（tokens）就是词数，类符（types）就是指语料库中每一词型（word form）的数量，又叫作词型数（比如，“要多读别人读过的好书”这句话，词数是9，词型数是8，那么型例比就是9/8）。型例比是反映文本词汇密度的数据，型例比越小，说明该作家使用的词汇越丰富。该数据可根据计算公式计算而得。

3. 单现词比例

单现词就是文本中只出现一次的词的数量，这个参数也可以反映出词汇的丰

富程度。单现词比例越大，说明词汇运用越丰富。该数据从统计此表中词频为1的词的数量，除以总词数获得。

4. 关键词

关键词就是文本中出现频率较高的词。有的学者将其称为高频词。本文选取了在语料中出现最高频的20个名词、20个动词、20个形容词。在词表中根据词频排序后，筛选频率最高的前20个名词、动词、形容词所得。

各语言结构类型计算公式（方法）：

1. 词长=字数/词数

词数：在Excel中数据筛选，过滤掉了标点，将得到的每一词的词频求和。

2. 型例比=词数/词型数

词型数：文本中出现的词的种类，在词表中，不论该词出现频率是多少，都记作1。在Excel中筛选出除了标点之外的词的种类所得。

3. 单现词比例=单现词数量/词数

4. 人称代词比例=各人称代词数量/人称代词总数量

表1　词长、型例比、单现词数量表[2]

作家	词数	词型数	单现词	词长	型例比	单现词比例
茅盾	127,576	14,491	8,230/4,977	1.5677	8.8038	0.0645
巴金	129,704	11,704	6,623/3,769	1.5420	11.0820	0.0510
老舍	129,414	14,425	7,966/5,796	1.5454	8.9715	0.0616

3.1　词长分析

语料库中每个作家的文本除掉标点均是20万字（考虑到句子完整以及表达需要，语料的规模有小幅出入，但不超过200字。具体数目是茅盾：200,117字；巴金：200,256字；老舍；200,203字）。从上表可以看出，巴金和老舍的词长相差不大，都在1.54左右，茅盾的词长是1.5677，比巴金高出0.0257，比老舍高出0.0223，这说明茅盾运用的多音节词，如复合词、成语、惯用语等多音节词的数量多一些。

3.1.1　音节数量分析

为了继续深入全面考察词的长短情况，本文进一步统计出此表中单音节、双音节、三音节、四音节词的数量。因为大于四音节的词多是切词留下的一些句式

标记，所以没有分析在内。切分结果如下表：

表2　音节数目比较表

作者（词数） 类型	单音节	双音节	三音节	四音节	大于四音节	小计
茅盾（127,576）	81,103	39,236	3,045	883	3,293	127,560
老舍（129,414）	86,228	38,016	2,093	853	2,209	129,399
巴金（129,704）	83,493	41,240	1,645	400	2,916	129,703

（注：括号中是三位作家的总词数，为了防止处理出现错误，经过验算，切分后各个音节数目相加基本等于总词数，误差接近于零）

为了使结果更为直观，用柱状图表示如下：

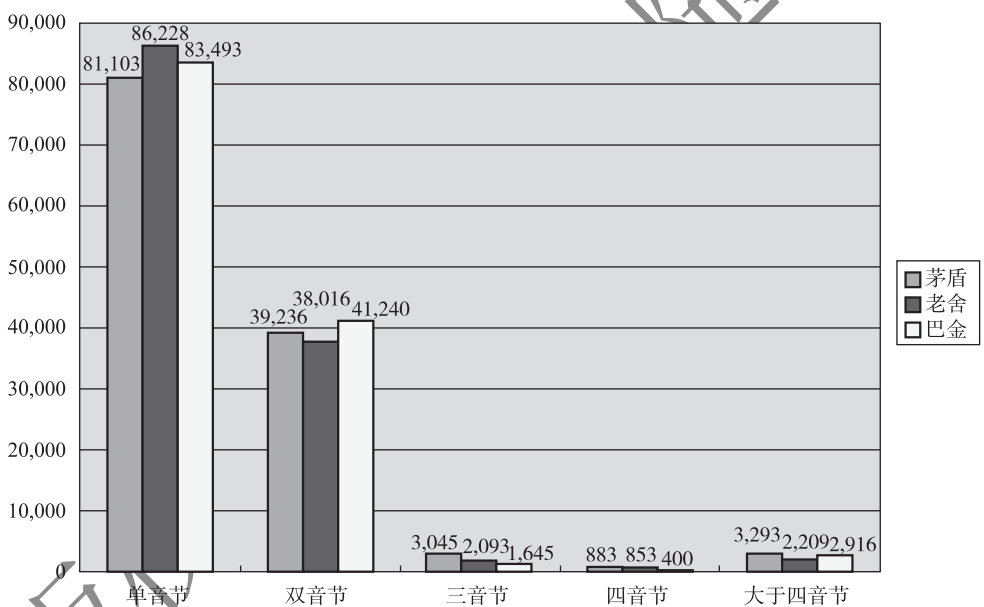


图2　音节数目比较图

可以看出，三位作家对单音节词的使用相差不大，老舍用得更多一些；双音节词的使用情况也相似，巴金用得更多一些。三音节词中，茅盾使用得最多，老舍次之，巴金最少。四音节词中，茅盾使用得最多，老舍次之，巴金最少。

3.1.2　三字格分析

通过表2，我们知道茅盾使用的三音节词是最多的，三音节词在现代汉语里主要是一些人名、地名、俗语。茅盾之所以使用如此多的三音节词，与以下因素

有关。

首先，与茅盾小说中的人物有关。茅盾擅长都市题材的写作，无论是《子夜》还是《腐蚀》，其背景都是上海、重庆一类的都市。在人物选取方面，茅盾描写的更多的是社会上流人物，如买办资本家吴孙莆、杜竹斋，银行资本家赵伯韬，经济学家李玉亭。而且，在结构方面，茅盾擅长鸿篇巨制，其作品人物众多，关系复杂。这些因素使得其作品中出现的人名、地名众多。而在一些表达方式上，上流社会团体也与下层民众不同，而更“讲究”一些。比如，仅就成年已婚女性类的称呼，茅盾作品中就出现了少奶奶（90次）、姨太太（3次）、密司*（*表示姓氏）（1次）、*老太（1次）、小老婆（1次）等几种。老舍作品中用的是孙媳妇（12次）、儿媳妇（9次）、老太太（27次）、姑奶奶（6次）、老妈妈（2次）等一系列具有北京特色的传统称呼，而巴金作品中基本没有使用这些词。

其次，与俗语的使用有关。现代汉语中的俗语一般都是三字格的，如“戴高帽”“侃大山”“跑火车”。茅盾作品中使用了大量的俗语，如“兜圈子”（3次）“闹乱子”（2次）“吊膀子”（2次）“献殷勤”（2次）“出风头”（1次）“开天窗”（1次）。老舍作品中的三字格俗语有“西北风”（2次）“阎王账”（1次）“绕弯子”（2次）“出风头”（1次）。巴金作品中的三字格俗语有“老好人”（4次）“贱骨头”（1次）。上述因素导致茅盾和老舍作品中三字格使用较多。

3.1.3　四字格分析

四字格是现代汉语中一类特殊格式，四字格中还经常伴有押韵、双声、叠韵、音节的重复等情况出现。比如，锒铛入狱（一、二音节叠韵）、大刀阔斧（一、二音节双声）、浑浑噩噩（前后两个音节重复）、斤斤计较（前两个音节重复，四个音节声母相同）。

经分析，茅盾使用的四字格有575种883个，如：志同道合、得意忘形、关门大吉、心口如一、牛头马面、朝三暮四、搜索枯肠、满腹经纶。老舍使用的四字格有274种853个，如：虎头虎脑、眉清目秀、有板有眼、发号施令、营私舞弊、满不在乎、不卑不亢、不辞而别等。巴金使用的四字格有256种400个，如：没精打采、不以为然、自言自语、有气无力、不由自主、不好意思、毫不迟疑、隐隐约约、大吃一惊、四面八方、有说有笑。

擅长运用四字格是语言富有音乐美的表现，独特的四字格式，使得语言富有韵律和节奏美，朗朗上口，铿锵有力，宛如音乐在耳畔流淌。

3.2　型例比分析

在三位作家中，巴金和老舍作品的型例比相差无几。茅盾作品的型例比是8.2514，巴金的作品是11.0820，老舍的作品是8.9715。柱状图示意如下：

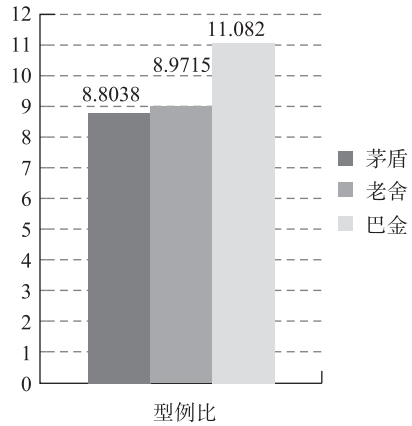


图3　型例比柱状图

从上图可以看出，茅盾作品的型例比最小，也就是说，其词汇丰富程度更强。我们认为，造成茅盾词汇丰富的原因有如下几个方面：

第一，独特的小说范式。这是从制导因素中的客观方面，也就是表达对象上看的。茅盾的创作是“社会分析型小说”，比起其他作家的小说，要求作者有更为开阔的视野和多触角。比如《子夜》中描写的上海经济界，《腐蚀》中小资产阶级对于革命的态度的描写。在小说结构上，为了提高小说反映生活和人的心灵深度的可能性，茅盾作品常常采取鸿篇巨制，并将小说向中长篇扩展。不仅如此，茅盾还建立起了全新的革命现实主义文学模式，注重大规模、全景式地反映社会现实。

第二，丰富的人生阅历。这是从表达主体分析的。茅盾用词丰富程度极高，这不仅因为其作品的内容广博，而且与其个人极高的文学修养也不无关联。茅盾是中国现代著名作家，又是社会活动家、文学评论家、文化活动家，是五四新文化运动先驱者之一，也是我国革命文艺奠基人之一。

茅盾生于浙江乌镇，这里毗邻上海，人文荟萃，素有鱼米之乡之称。没落的农业发展和独特的地理人文环境，成就了茅盾勇于面向世界的、开放的文化心态以及精致入微的笔触。他在年纪极小时便读过私塾、家塾，八岁入小学并熟读四大名著和古典文学，对算数、绘画也感兴趣。后来，茅盾考入了北京大学预科，毕业后先后做过商务印书馆和《小说月报》编译。五卅运动爆发之后，茅盾积极投身群众革命运动。先后在上海、武汉、重庆、长沙、广州、新疆、香港地区任职，并有客居日本、访学苏联的经历。丰富的人生阅历、广泛的兴趣爱好、多领域的建树，这些使得茅盾的词汇储备量比起常人更加丰富，成为茅盾取之不尽的语言宝藏。

不妨再拿例子说明。

在《子夜》开头的一场景物描写中，茅盾运用了“软风”“浊水”“暮霭”“薄

雾”等词，是一个形语素+一个名语素，而不是“风”“水”“晚上的云彩”；颜色词里，茅盾用了“冥色”“赤光”“绿焰”“青燐”，而不是简单的“红光”“绿色的火焰”“金色”“绿色”这些一般词汇。力避重复、力求贴切、力创新意，是茅盾创作表达的一大特色。

在《家》开头中的景物描写中，巴金运用了“风”“雪片”“墙角”“水泥”等事物名词，颜色词运用了“白色”“白茫茫”，都属于常用词汇。

《骆驼祥子》中的一处景物描写，老舍运用了“水”“暮色”“河”，颜色词用的是“绿”“深绿”“长绿”，虽然也颇多别称，力避重复，但是因为这些都是词组，所以计算机都自动切分开来了。这也显示出了作为人民艺术家的老舍更多地运用浅显易懂的词语来描写景物，追求词汇表达的平实、贴切与平民化。

其实，词汇的丰富程度与作品面向的读者群体、作品的表达对象也是密切相关的，在语言学中，这与不同的社会言语社团相关。如果表达的对象是下层民众，那么就要运用一些通俗易懂的词汇。不能单凭这一项来判断某位作家文学修养的高下。

在同等语料规模中，茅盾作品的单现词出现比例最大，老舍次之，巴金最小。单现词也是反映词汇丰富程度的一个参数。这里反映的情况与型例比相一致。

3.3 人称代词分析

三位作家使用的人称代词也存在较明显的不同，详见下表：

表3 人称代词表

作家	第一人称				第二人称		第三人称						总计
人称	我	我们	咱	咱们	你	你们	他	她	它	他们	她们	它们	
茅盾	3,827	264	0	35	1,310	92	1,990	652	24	425	41	3	8,663
小计	4,126				1,402		3,135						8,663
比例	0.4763				0.1618		0.3619						1
巴金	2,167	339	0	0	1,796	177	4,541	2,628	107	442	84	39	12,320
小计	2,506				1,973		7,841						12,320
比例	0.2034				0.1602		0.6264						1
老舍	1,341	167	33	135	970	86	2,869	978	45	384	19	26	7,053
小计	1,676				1,056		4,321						7,053
比例	0.2376				0.1497		0.6127						1

从上表可以看出，三位作家作品自身比较中，茅盾运用第一、三人称更多一些；巴金和老舍都倾向于运用第三人称。在作者之间进行比较，茅盾更倾向于运

用第一人称叙事，其比例约是巴金、老舍的两倍；极少运用第三人称，只有巴金、老舍的二分之一；老舍较少运用第二人称叙事。

在写作中，不同人称的运用产生不同的表达效果。

1. 第一人称叙事直接拉近了作者和读者的距离，给人身临其境之感，另外，运用第一人称还有助于抒发个人情感，而且更适合进行心理描写。

2. 第二人称的作用主要是拉近叙述者和读者之间的距离，便于情感交流，产生亲切感。

3. 第三人称叙事更加冷静客观，便于叙事，而且不受时间和空间的限制。

三位作家使用的人称比例图示如下：

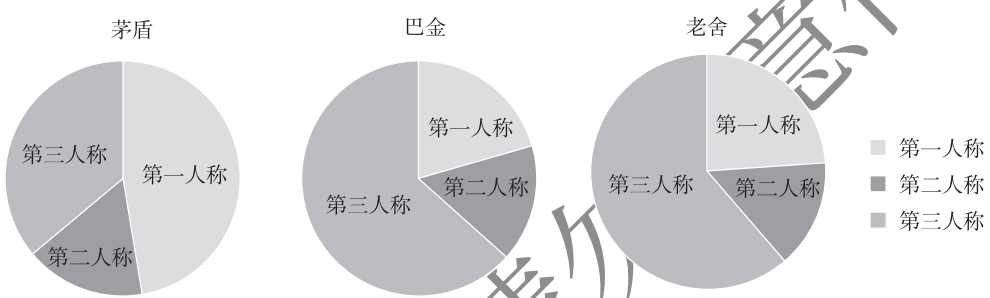


图4　人称代词比例图表

众所周知，茅盾是个多产作家，各种类型的作品都尝试过。在叙事上，他也不仅仅局限于第三人称，后期的作品《腐蚀》就是一个明显的例子。这部小说采取的是第一人称日记体的形式，细致描写了女主人公赵惠明的心路历程，婉转凄恻，动人心扉。在这部作品中，大段的心理独白随处可见。例如，

> 近来感觉到最大的痛苦，是没有地方可以说话。我心里的话太多了，可是找不到一个人可以让我痛痛快快对他说一场。
>
> 近来使我十二万分痛苦的，便是我还有记忆，不能把过去的事，完全忘记。这些“回忆”的毒蛇，吮吸我的血液，把我弄成神经衰弱。
>
> 近来我更加看不起我自己，因为我还有所谓“希望”。有时我甚至于有梦想。我做了不少的白日梦：我又有知心的朋友了，又可以心口如一，真心的笑了，而且，天翻地覆一个大变动，把过去的我深深埋葬，一个新生的我在光天化日之下有说有笑——并且也有适宜于我的工作。
>
> 我万分不解，为什么我还敢有这样非分之想，还敢有这样不怕羞

的想望。难道我还能打破重重魔障，挽救自己么？

——茅盾《腐蚀——九月十五日》

这是《腐蚀》开头第一篇日记，直接运用“我”叙事，真实地表达了赵慧明此刻摇摆不定、孤独矛盾的心理。

第二、三人称的使用比例较高的是巴金，大量的第三人称写作使得巴金在人物刻画以及场景描述上游刃有余，常常使人置若其描绘的场景之中。巴金第三人称的叙事拉近了作者与读者之间的距离，更加便于抒情。

3.4 指示代词分析

《现代汉语》（增订四版）中概括出了指示代词“这”和“那”的三种主要用法：

第一，指示。如：“这孩子调皮”；“那放羊的刚过来”。

第二，代替。如：“这可是最好的瞄准手”；“那里寸草不生”。

第三种是虚指用法，主要用于对举中。如“咱不图这，不图那，就图那娃思想好”。

人们在使用语言进行表达时总是存在着一种矛盾，那便是表达上要求尽可能完满、详尽和省力的矛盾，这也是表达和人的生理的一种矛盾。语言表达追求经济性与效率性，而人们生理上和精神上存在自然的惰性，于是就倾向于少量的、更具有普遍性的语言单位，指示代词便是一类这样的单位。指示代词的运用，让我们不必对话语交际中已经反复出现的概念重复阐述，而是以一种更为简洁的方式代替。

茅盾使用指示代词“这”1,251次，频率0.0098，“那”758次，频率0.0059。巴金使用“这”705次，频率0.0054，“那”368次，频率0.0028。老舍使用“这”500次，频率0.0039，“那”352次，频率0.0027。可见，茅盾的指示代词使用频率明显高于巴金和老舍。

茅盾使用“这”“那”的频率明显高于巴金和老舍，说明其叙事时善于运用上文事物指代下文。在表达上，“这”“那”的使用不仅使行文更加紧凑、流畅，且可以营造一种画面感，培养读者的想象力。

看下例：

怎么/r 不/d 捉/v！ /w 可是/c 捉/v 不/d 完/v 。/w 啊/e 哟/e！ /w 真/d 不/d 知道/v 哪里/r 来/v 的/u 这/r 许多/m 不/d 要/v 性命/n 的/u 人/n！ /w ——/w 可/v 是/v ，/w 四/m 妹/n ，/w 你/r 这/r 一/m 身/q 衣服/n 实/a 在/p 看/v 了/u 叫/v 人/n 笑/v 。/w 这/r 还/d 是/v 十/m 年/q 前/f 的/

u 装束/n！ /w 明天/t 赶快/d 换/v 一/m 身/n 罢/v！ /w ”/w

——茅盾《子夜》

这是《子夜》中二小姐芙芳和四小姐蕙芳的对话，二人谈论起上海遍地开花的共产党员以及宣传标语时，使用了三个“这”。其中，第一、二个都是指示作用，第三个是替代作用。茅盾运用指示代词“这”，把人物、事物定位，给人以明晰的画面感，叙事上也更为清晰。

3.5 关键词分析

截取三位作家作品中出现频率最高的十个名词、十个动词、十个形容词进行分析。关键词列表如下：

表 4 关键词比较表

茅盾			巴金			老舍		
词语	词性	词频	词语	词性	词频	词语	词性	词频
人	n	667	人	n	771	人	n	671
话	n	296	母亲	n	510	钱	n	480
吴	n	257	话	n	397	车	n	362
事	n	248	头	n	385	象	n	342
手	n	198	觉慧	n	272	事	n	293
脸	n	191	声音	n	244	老张	n	270
头	n	186	时候	n	216	时候	n	247
小昭	n	183	琴	n	199	祥子	n	238
荪	n	168	事	n	198	儿	n	228
时候	n	161	脸	n	193	先生	n	227
是	v	2,573	说	v	1,648	是	v	2,228
有	v	1,120	是	v	1,493	有	v	1,058
说	v	1,031	去	v	920	说	v	917
来	v	657	走	v	823	去	v	833
要	v	568	有	v	719	看	v	510
去	v	516	来	v	611	要	v	486
笑	v	463	看	v	599	来	v	472
看	v	433	要	v	545	能	v	442
想	v	334	想	v	476	出	v	426
出	v	318	会	v	376	走	v	417

（待续）

（续表）

茅盾			巴金			老舍		
词语	词性	词频	词语	词性	词频	词语	词性	词频
好	a	305	好	a	566	好	a	558
大	a	275	大	a	253	大	a	522
小	a	252	新	a	193	小	a	472
突然	a	94	小	a	115	老	a	288
老	a	88	痛苦	a	103	快	a	102
胖	a	70	早	a	87	长	a	91
新	a	56	冷	a	69	红	a	91
红	a	53	突然	a	67	高	a	88
昭	a	46	高兴	a	61	少	a	72
满	a	46	红	a	59	黑	a	70

通过上表可以看出，三位作家作品中高频词的使用上重合率很多，大都属于基本词汇。仔细观察可以看出，三者存在如下差异：

3.5.1　名词

除去表示人名的词和重合率较高的名词“人”，我们发现，茅盾、巴金作品中的十个出现频率最高的名词中，都出现了2—3个表示人体器官的词语，如手、脸、头，而老舍先生的作品中没有出现。

在小说中，表示人体器官的词语多用来进行人的肖像描写。出现这样的差异不由得引导我们思考老舍在刻画人物相貌方面有着和巴金、茅盾不一样的手法。

老舍描写人物肖像时，往往抓住其最主要的特点，然后配合人物的行动、言谈来描写，正像老舍自己所说的那样，“先有个轮廓，而后顺手以种种行动来使外貌活动起来，是一种很有效的方法。”比如，《四世同堂》中“大赤包”的描写：“是个快五十岁了还专爱穿大红衣服的胖妇人。之所以叫她大赤包，因为赤包儿经儿童揉弄以后，皮儿便皱起来，露出里面的黑种子。冠太太的脸上也有不少的皱纹，而且鼻子上有许多雀斑，尽管她还擦粉抹红，也掩饰不了脸上的褶子与黑点。”这一段肖像描写抓住了人物的突出特点，配合人物行动、言谈表现，展示出了老舍先生在肖像描写上的独到之处。

一个有趣的现象是，在进行关键词分析时，我们发现“眼睛（眼）”一词虽然没有出现在前十个高频词之中，但是在三位作家的作品中都出现了较高的频率。与其相关的词语如“眼眶、眼前、眼色”更是占了很大的比重，这也显示出“眼睛是心灵窗户”的现象。“中国现代文学之父”——鲁迅先生在人物肖像刻画最让

人惊叹有两种手法。一种是白描，还有就是“画眼”。如《祝福》中祥林嫂的“眼珠间或一轮”令人印象深刻。

在表示生产生活资料的名词中，老舍先生的作品中出现概率很大。比如“车、钱”，这也显示出老舍先生关心下层民众生活、同情小人物疾苦的普世情怀。

3.5.2　动词

三位作家作品的高频动词中均出现了联系动词“是”“有”，能愿动词“要”，表示基本动作的“走”“来”“去”等。茅盾和巴金的作品中出现了表示心理活动的动词“想”，茅盾的作品中还出现了刻画人物情貌的“笑”。这说明茅盾、巴金对于人物的心理描写细致入微，而茅盾更是注重人物情貌的刻画。

3.5.3　形容词

相较其他两位作家，巴金的作品在高频形容词中出现了两个表示情感的性质形容词，即“痛苦、高兴”，这体现出了巴金在创作中对人物情感的独到把握。狄德罗曾经宣称“没有感情这个品质，任何笔调都不可能打动人心”。巴金素有“抒情”大师的称号，他是一个感情充沛并且饱蘸着热情写作的作家，他的抒情真挚而强烈。“巴金明明是在创作小说，却每每像一位热情的鼓动诗人，哭歌忘形，炽烈的情绪了无遮掩。无论是痛苦、焦灼、失望、悲愤、希冀、期待，全都倾注在字里行间。”“他的文笔是用心血和胆汁浸润过的。”“一烧烧到白热度，便毫不隐瞒，毫不修饰，照情感的原热度，崩裂到字里行间。”高频词中“痛苦、高兴”的出现验证了我们这种阅读体验。

4．结语

茅盾、巴金、老舍三位作家的作品在词汇风格上有着鲜明特色。

三字格、四字格的运用使茅盾的作品更富有音乐美，丰富的词汇为其广阔的社会视野奠定了坚实的语言基础，大量的语气词运用彰显出其细腻、高超的心理描写技巧。高频指示代词让茅盾的小说在鸿篇巨制之中避免了场景杂乱和人物混淆的弊病，使得他的小说场景生动、叙事脉络清晰，人物关系明了。第一、三人称的叙事模式拉近与读者距离的同时不乏理性，使其作品既饱含着生命的哭诉，又具有冷静、深刻的社会剖析。

抒情大师巴金尽情书写人物的离合悲欢。他饱蘸着一腔热情，忠诚于读者，用高频度的感情形容词抓住读者的心灵。大量的谓语修饰语使他的人物生动而鲜活，动作形象而逼真，如确乎存在于现实生活中一般活灵活现。第二人称叙事拉近了与读者的距离，心理动词的运用使得其心理描写更加细致入微。

老舍先生素有“人民艺术家”和“语言大师”的称号。一直以来，平民化、生活化、幽默、京味儿是人们阅读老舍作品的直观感受。数据表明，老舍先生作

品中表示生产生活资料的名词大量涌现，表现了其关心下层民众疾苦的济世情怀。助词“的”是老舍先生在五四运动时期受到西方文学的影响，对于欧化句的大胆尝试。文言词的运用让老舍先生的作品具有古雅的韵味。

注　释

1. 计算方法见黄伟、刘海涛（2009）。
2. 本文的计算结果均保留到小数点后四位。

参考文献

黄　伟、刘海涛，2009，汉语语体计量特征在文本聚类中的应用 [J]，《计算机工程与应用》（29）：25-27。

黎运汉，1990，《汉语风格探索》[M]。北京：商务印书馆。

彭政策，1987，作品语言风格的数量研究 [J]，《当代修辞学》（2）：9-11。

王道英，2005，《“这”“那”的指示功能研究》[M]。上海：学林出版社。

徐炳昌，1988，关于作家语言风格研究方法的思索 [J]，《扬州大学学报（人文社会科学版）》（1）：109-112。

俞士汶，2003，《计算语言学概论》[M]。北京：商务印书馆。

曾毅平、朱晓文，2006，计算方法在汉语风格学研究中的应用 [J]，《福建师范大学学报（哲学社会科学版）》（1）：14-17。

张华平、商建云，2019，NLPIR-Parser：大数据语义智能分析平台 [J]。《语料库语言学》（1）：87-104。

通信地址： 200240　上海市上海交通大学人文学院

林纾翻译语料库的创建与研究*

福建工程学院 戴光荣

提要：林纾小说翻译作品研究是一项浩大的文化工程，学界还没有对其进行系统梳理与探讨，大多数研究是基于个人内省式分析，单纯依赖其一部甚或某些章节词句进行探讨，结果会偏离具体的翻译实践与当时的社会文化背景。本文旨在介绍林纾翻译语料库创建过程中所关注的问题，挖掘林纾翻译口译合作者及其合作翻译信息、源语文本为英文的相关书目信息、语料库分词与翻译策略标注等。林纾译文语料库的创建，将为翻译学界提供一种“解读林纾”的新方法，为全方位认识与评价林纾提供客观、科学的数据。

关键词：林纾翻译语料库、语料库创建、林纾口译合作者、解读林纾

1. 引言

在外汉翻译文学史上，林纾是绕不开的关键人物之一。回顾学界对翻译文学的梳理，大多集中在“史”的叙述，诸如中国出版的第一部翻译文学史著作《中国翻译文学史稿》（陈玉刚 1989），以及后来的《中国翻译文学史》（孟昭毅、李载道 2005），《中国现代翻译文学史》（谢天振、查明建 2004），《中国20世纪外国文学翻译史》（查明建、谢天振 2007），《20世纪中国翻译文学史》（连燕堂 2009）等，这类著作大多关注两个方面：一是翻译文学作品产生的时代背景；二是译家的生平、译作及翻译思想（方法），很少关注译作本身，缺乏从语言层面将译作进行细致的文本分析与批评。以陈玉刚的《中国翻译文学史稿》为例，该书介绍了梁启超、严复、林纾、鲁迅、茅盾、郭沫若、巴金、瞿秋白等著名译家，将他们的生平事迹、翻译方法与策略等作了详细介绍，大多是从文化层面进行解读，没有对这些译家的任何一部译作进行文本分析，存在着“译文学”意识严重缺乏、“译文不在场”的情况（王向远 2015）。

纵观多年来学界对林纾的研究，特别是对其译作的研究，大多基于个人内省式分析。学界探讨林纾的研究范围有待拓宽，研究方法也有待改进。林纾小说翻译作品本身就是一项工程浩大的文化事业，单纯依赖其一部或某些章节词句进行探讨，结果会偏离具体的翻译实践。到目前为止，学界还没有对林纾的翻译作品进行系统性的探讨。

* 本文系福建省社会科学研究基地重大项目（FJ2015JDZ037）；地方文献整理研究中心重点项目（DFWX2015-A01）阶段性成果。

为了更好地对林纾翻译的作品进行科学客观的描述与分析，需要我们借助当代先进的计算机技术与语言学研究的最新成果。随着学科的发展，尤其是近年来计算语言学、语料库语言学、语料库翻译学、翻译风格学等的发展（Malmkjær 2004；Boase-Beier 2006；Studer 2008；Ji 2009；Ho 2011；Huang 2015），为我们更好地开展林纾研究提供了新的视角（戴光荣 2015，2018a，2018b，2018c）。语料库翻译学以真实的双语语料或翻译语料为研究对象，以数据统计和理论分析为研究方法，采用语内与语际对比相结合，依据语言学、文学和文化理论等理论，对翻译现象进行历时或共时的描写和解释，探索翻译过程与翻译本质等内容（胡开宝 2011；王克非 2012）。本文在语料库翻译学视角下，探讨林纾翻译语料库创建及其研究。

2. 林纾翻译语料库创建可行性

林纾一生翻译了大量作品，尤其在文学作品翻译方面数量众多。其中许多译作堪称精品。然而世人很少对其翻译的作品进行穷尽性挖掘，很多研究也只是针对其某些或某一部译作进行探讨，没能从整体上对林纾的翻译风格进行分析与把握，从而作出更为科学、客观而详尽的分析。进行语料收集与语料库创建，对于个人研究者来说，是一个浩大的研究工程。这就向我们提出了创建林纾译文语料库的必要性问题。

创建林纾译文语料库的主要任务是语料收集与语料库创建，包括以下阶段：（1）语料选取；（2）语料样本除噪；（3）样本标注；（4）样本入库。本人主持的研究团队利用学校图书馆馆藏的林纾研究专题资源库[1]，外加商务印书馆的鼎力帮助，对林纾翻译小说文本进行了详细整理。所搜集的林纾翻译小说都是繁体竖排格式，部分文字识别与输入存在很多困难，需要进行大量的人工校对。研究团队从2011年10月开始展开了这项工作，投入了大量的人力与物力。目前林纾译文的总量基本保持不变（不排除其未曾刊发的译作被发现的可能性），随着资料的充实，林纾翻译语料库将逐步完善。

3. 林纾翻译语料库书目选择

林纾翻译西方文学作品（其中以小说为主）数量惊人，这是众所周知的。据统计，林译作品有246种，已发表222种，未刊作品24种。他的翻译事业始于1897年，终于1921年，前后近四分之一个世纪。翻译作品来自英国的最多，占半数以上，共106种；其次是法国，共30种；美国有26种，俄国有12种，此外希腊、德国、日本、比利时、瑞士、挪威、西班牙各一种（马泰来 1981，1982，2013；陈平原 1989；林薇 1990；张俊才 2007）。

林纾翻译作品，除了在商务印书馆等知名出版社出版之外，还有很多在杂志上连载发表。如从1906年9月起林纾的译作《空谷佳人》《荒唐言》《罗刹因果录》《鱼雁抉微》《桃大王因果录》《赂史》《戎马书生》等先后在《东方杂志》上连载。《东方杂志》1919年12月改版，林纾译品退出《东方杂志》。在这些年总共145期之中，有58期上刊有林纾的译作，占总数的40%。林译小说支撑起了《东方杂志》（1904—1919）的小说栏，特别是从1909年4月至1919年12月杜亚泉编辑《东方杂志》的10年中，更是对林纾的作品青睐有加，使林译小说成为早期《东方杂志》文学生命的重要支柱（王勇 2009）。

考虑到本研究后续将深入进行林纾翻译源语文本的探讨，因此我们重点收集源语文本为英文的相关书目信息，并列表如下。

表1　林纾翻译作品一览（原著为英文）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
1. 英国哈葛德 Sir Henry Rider Haggard（1856—1925）	1	*Eric Brighteyes*（1891）	《埃司兰情侠传》	魏易	1904	广智书局
	2	*Nada the Lily*（1892）	《鬼山狼侠传》	曾宗巩	1905	商务印书馆
	3	*Joan Haste*（1895）	《迦茵小传》	魏易	1905	商务印书馆
	4	*Cleopatra*（1889）	《埃及金塔剖尸记》	曾宗巩	1905	商务印书馆
	5	*Montezuma's Daughter*（1893）	《英孝子火山报仇录》	魏易	1905	商务印书馆
	6	*Allan Quatermain*（1887）	《斐洲烟水愁城录》	曾宗巩	1905	商务印书馆
	7	*Mr. Meeson's Will*	《玉雪留痕》	魏易	1905	商务印书馆
	8	*Colonel Quaritch V.C.*（1888）	《洪罕女郎传》	魏易	1906	商务印书馆
	9	*Black Heart and White Heart and Other Stories*（1900）	《蛮荒志异》	曾宗巩	1906	商务印书馆
	10	*Beatrice*（1890）	《红礁画桨录》	魏易	1906	商务印书馆
	11	*Dawn*（1884）	《橡湖仙影》	魏易	1906	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
1. 英国 哈葛德 Sir Henry Rider Haggard （1856—1925）	12	*People of the Mist*（1894）	《雾中人》	曾宗巩	1906	商务印书馆
	13	*King Solomon's Mines*（1885）	《钟乳髑髅》	曾宗巩	1908	商务印书馆
	14	*Jess*（1887）	《玑司刺虎记》	陈家麟	1909	商务印书馆
	15	*Fair Margaret*（1907）	《双雄较剑录》	陈家麟	1915	商务印书馆
	16	*She*（1886）	《三千年艳尸记》	曾宗巩	1910	商务印书馆
	17	*Benita*（1906）	《古鬼遗金记》	陈家麟	1912-13	商务印书馆
	18	*The Ghost Kings*（1908）	《天女离魂记》	陈家麟	1917	商务印书馆
	19	*The Brethren*（1904）	《烟火马》	陈家麟	1917	商务印书馆
	20	*The Witch's Head*（1887）	《铁匣头颅，又续篇》	陈家麟	1919	商务印书馆
	21	*Maiwa's Revenge*（1888）	《豪士述猎》	陈家麟	1919	商务印书馆
	22	*The World's Desire*（1890）	《金梭神女再生缘》	陈家麟	1920	商务印书馆
	23	*Queen Sheba's Ring*（1910）	《炸鬼记》	陈家麟	1921	商务印书馆
2. 英国 科南利达 Arthur Conan Doyle （1859—1930）	24	*Micah Clarke*（1889）	《金风铁雨录》	魏易	1907	商务印书馆
	25	*A Study in Scarlet*（1887）	《歇洛克奇案开场》	魏易	1908	商务印书馆
	26	*The Refugees*（1893）	《恨绮愁罗记》	魏易	1908	商务印书馆
	27	*Uncle Bernac*（1897）	《髯刺客传》	魏易	1908	商务印书馆
	28	*The Doings of Raffles Haw*（1892）	《电影楼台》	魏易	1908	商务印书馆
	29	*Beyond the City*（1892）	《蛇女士传》	魏易	1908	商务印书馆
	30	*The White Company*（1891）	《黑太子南征录》	魏易	1909	商务印书馆
3. 英国 却而司迭更司Charles Dickens (1812—1870）	31	*Nicholas Nickleby*（1839）	《滑稽外史》	魏易	1907	商务印书馆
	32	*The Old Curiosity Shop*（1841）	《孝女耐儿传》	魏易	1907	商务印书馆
	33	*David Copperfield*（1850）	《块肉余生述，又续编》	魏易	1908	商务印书馆
	34	*Oliver Twist*（1838）	《贼史》	魏易	1908	商务印书馆
	35	*Dombey and Son*（1848）	《冰雪因缘》	魏易	1909	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
4. 英国 莎士比 William Shakespeare (1564—1616)	36	*Richard II*（1597）	《雷差得纪》	陈家麟	1916	《小说月报》卷 7 第 1 号（V7 N1）
	37	*Henry IV*（1598—1600）	《亨利第四纪》	陈家麟	1916	《小说月报》V7 N2-4
	38	*Henry VI*（1594—1623）	《亨利第六遗事》	陈家麟	1916	商务印书馆
	39	*Julius Caesar*（1623）	《凯彻遗事》	陈家麟	1916	《小说月报》V7 N5-7
	40	*Henry V*（1600）	《亨利第五纪》	陈家麟	1925	《小说世界》V12 N9-10
5. 英国 司各德 Walter Scott （1771—1832）	41	*Ivanhoe*（1820）	《撒克逊劫后英雄略》	魏易	1905	商务印书馆
	42	*The Talisman*（1825）	《十字军英雄记》	魏易	1907	商务印书馆
	43	*The Betrothed*（1825）	《剑底鸳鸯》	魏易	1907	商务印书馆
6. 美国 华盛顿欧文 Washington Irving （1783—1859）	44	*The Sketch Book of Geoffrey Crayon, Gent.*（1820）	《拊掌录》	魏易	1907	商务印书馆
	45	*The Alhambra*（1832）	《大食故宫馀载》	魏易	1907	商务印书馆
	46	*Tales of a Traveler*（1824）	《旅行述异》	魏易	1907	商务印书馆
7. 英国 达孚 Daniel Defoe （1660—1731）	47	*Life and Strange Surprising Adventures of Robinson Crusoe* 1719	《鲁滨孙漂流记》	曾宗巩	1905	商务印书馆
	48	*Farther Adventures of Robinson Crusoe*（1719）	《鲁滨孙漂流续记》	曾宗巩	1906	商务印书馆
8. 英国 马支孟德/马尺芒忒 Arthur W. Marchmont （1852—1923）	49	*For Love or Crown*（1901）	《西利亚郡主别传》	魏易	1908	商务印书馆
	50	*The Man Who Was Dead*（1907）	《黑楼情孽》	陈家麟	1914	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
9. 美国 锁司倭司女士/沙司卫甫夫人 Emma D. E. N. Southworth（1819—1899）	51	*The Changed Brides*（1869）	《薄倖郎》	陈家麟	1915	商务印书馆
	52	*The Bride of Llewellyn*（1864）	《以德报怨》	毛文钟	1922	商务印书馆
10. 英国 杨支 Charlotte Mary Yonge（1823—1901）	53	*The Dove in the Eagle's Nest*（1866）	《鹰梯小豪杰》	陈家麟	1916	商务印书馆
	54	*The Lances of Lynwood*（1855）	《戎马书生》	陈家麟	1920	商务印书馆
11. 美国 克雷夫人 （Bertha M. Clay）	55	英文书名待考	《想夫怜》	毛文钟	1920	《小说月报》V11 N9-12
	56	英文书名待考	《僵桃记》	毛文钟	1921	商务印书馆
12. 美国 斯土活 Harriet Beecher Stowe（1811—1896）	57	*Uncle Tom's Cabin*（1852）	《黑奴吁天录》	魏易	1901	武林魏氏刊本
13. 英国 阿纳乐德 Thomas Arnold 1795—1842	58	*The Second Punic War*（1886）	《布匿第二次战纪》	魏易	1903	京师大学堂官书局
14. 英国 兰姆姐弟 Charles Lamb（1775—1834）；Mary Lamb（1764—1847）	59	*Tales from Shakespeare*（1807）	《吟边燕语》	魏易	1904	商务印书馆
15. 美国 阿丁 William L. Alden 1837—1908	60	*Jimmy Brown Trying to Find Europe*（1889）	《美洲童子万里寻亲记》	曾宗巩	1904	商务印书馆
16. 英国 洛加德 John Gibson Lockhart 1794—1854	61	*History of Napoleon Buonaparte*（1829）	《拿破仑本纪》	魏易	1905	商务印书馆
17. 英国 斯威佛特 Jonathan Swift 1667—1745	62	*Gulliver's Travels*（1726）	《海外轩渠录》	魏易	1906	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
18. 英国 阿瑟毛利森 Arthur Morrison 1863—1945	63	*Chronicles of Martin Heweitt*（1895）	《神枢鬼藏录》	魏易	1907	商务印书馆
19. 英国 几拉德	64	英文书名待考	《花因》	魏易	1907	中外日报馆
20. 英国 大限克力司蒂穆雷 David Christie Murray 1847—1907	65	*The Martyred Fool*（1895）	《双孝子噀血酬恩记》	魏易	1907	商务印书馆
21. 英国 路易司地文，佛尼司地文 Robert Louis Stevenson 1850—1894；Fanny Van de Graft Stevenson 1840—1914	66	*More New Arabian Nights: The Dynamiter*（1885）	《新天方夜谭》	曾宗巩	1908	商务印书馆
22. 英国 麦里郝斯 Sophia H. Maclehose ?—1912	67	*Tales from Spenser, Chosen from the Faerie Queene*（1890）	《荒唐言》	曾宗巩	1914	商务印书馆
23. 英国 约翰沃克森罕	68	英文书名待考	《天囚忏悔录》	魏易	1908	商务印书馆
24. 英国 男爵夫人阿克西 Baroness Emma Orczy 1865—1947	69	*The Scarlet Pimpernel*（1905）	《大侠红蘩蕗传》	魏易	1908	商务印书馆
25. 英国 却洛得倭康，诺埃克尔司 （作者英文名待考）	70	英文书名待考	《彗星夺婿录》	魏易	1909	商务印书馆
26. 英国 蚩立伯倭本翰 E. Phillips Oppenheim 1866—1946	71	*The Secret*（1907）	《藕孔避兵录》	魏易	1909	商务印书馆
27. 安东尼贺迫 Anthony Hope 1863—1933	72	*A Man of Mark*（1890）	《西奴林娜小传》	魏易	1909	商务印书馆
28. 英国 司丢阿忒	73	英文书名待考	《脂粉议员》	魏易	1909	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
29. 英国 色东麦里曼 Henry Seton Merriman 1862—1903	74	*From One Generation to Another*（1892）	《芦花馀孽》	魏易	1909	商务印书馆
30. 英国 马克丹诺保德庆 M. McDonnel Bodkin 1850—1933	75	*The Quests of Paul Beck*（1908）	《贝克侦探谈，又续编》	陈家麟	1909	商务印书馆
31. 英国 测次希洛	76	英文书名待考	《残蝉曳声录》	陈家麟	1914	商务印书馆
32. 英国 希洛	77	英文书名待考	《罗刹雌风》	力树萱	1915	商务印书馆
33. 英国 倭尔吞	78	英文书名待考	《深谷美人》	陈器同	1914	宣元阁
34. 英国 马格内	79	英文书名待考	《石麟移月记》	陈家麟	1915	商务印书馆
35. 美国 包鲁乌因 James Baldwin 1841—1925	80	*Thirty More Famous Stories Retold*（1905）	《秋灯谭屑》	陈家麟	1916	商务印书馆
36. 英国 希登希路	81	英文书名待考	《红篋记》	陈家麟	1916	《小说月报》V7 N3-10
37. 英国 威利孙	82	英文书名待考	《情窝》	力树萱	1916	商务印书馆
38. 英国 克拉克 Mary Cowden Clarke 1809—1898	83	*The Thane's Daughter*（1850）	《奇女格露枝小传》	陈家麟	1916	商务印书馆
39. 英国 鹘刚伟 （作者英文名待考）	84	英文书名待考	《云破月来缘》	胡朝梁	1916	商务印书馆
40. 美国巴苏谨 （作者英文名待考）	85	英文书名待考	《橄榄仙》	陈家麟	1916	商务印书馆
41. 英国 倩伯司 W. & R. Chambers, Ltd.	86	*Chambers's Complete Tales for Infants*	《诗人解颐语》	陈家麟	1916	商务印书馆
42. 英国 Charles Cowden Clarke 1787—1877	87	*Tales from Chaucer in Prose*（2nd ed. 1870）	《坎特伯雷故事》	陈家麟	1916—1917	《小说月报》V7 N12/V8 N2-10
43. 英国 利华奴 （作者英文名待考）	88	英文书名待考	《柔乡述险》	陈家麟	1917	《小说月报》V8 N1-6
44. 英国 布司白 Guy Boothby 1867—1905	89	*A Brighton Tragedy*（1905）	《女师饮剑记》	陈家麟	1917	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
45. 英国 参恩女士（作者英文名待考）	90	英文书名待考	《桃大王因果录》	陈家麟	1918	商务印书馆
46. 英国 陈施利（作者英文名待考）	91	英文书名待考	《牝贼情丝记》	陈家麟	1917	商务印书馆
47. 英国 赖其镗女士（作者英文名待考）	92	英文书名待考	《痴郎幻影》	陈器同	1918	商务印书馆
48. 英国 巴克雷 Florence L. Barclay 1862—1921	93	*The Rosary*（1909）	《玫瑰花，又续编》	陈家麟	1918	商务印书馆
49. 英国 亚波倭得 Allan Upward 1863—1926	94	*The Phantom Torpedo-boats*（1905）	《赂史》	陈家麟	1920	商务印书馆
50. 美国尼可拉司 Nicholas Carter	95	英文书名待考	《焦头烂额》	陈家麟	1920	商务印书馆
51. 英国 美森（作者英文名待考）	96	英文书名待考	《妄言妄听》	陈家麟	1920	商务印书馆
52. 英国 约克魁迭斯（作者英文名待考）	97	英文书名待考	《西楼鬼语》	陈家麟	1919	商务印书馆
53. 英国 武英尼（作者英文名待考）	98	英文书名待考	《鬼窟藏娇》	陈家麟	1919	商务印书馆
54. 美国卡扣登 Charles Major 1856—1913	99	*When Knighthood Was In Flower*（1898）	《莲心藕缕缘》	陈家麟	1919	商务印书馆
55. 美国 堪伯路	100	英文书名待考	《还珠艳史》	陈家麟	1920	商务印书馆
56. 英国 达威生 Gladys Davidson	101	*Stories from the Operas*（1914）	《泰西古剧》	陈家麟	1920	商务印书馆
57. 英国 高桑斯（作者英文名待考）	102	英文书名待考	《欧战春闺梦，又续编》	陈家麟	1920	商务印书馆
58. 英国 伯鲁夫因支 Thomas Bulfinch, 1796—1867	103	英文书名待考	《怪董》	陈家麟	1921	商务印书馆

（待续）

（续表）

作者信息	序号	英文书名及出版时间	中文书名	口译者	出版年	出版社/出处
59. 英国 斐鲁丁 Henry Fielding 1707—1754	104	*A Journey from this World to the Next*（1743）	《洞冥记》	陈家麟	1921	商务印书馆
60. 英国 安司倭司 William Harrison Ainsworth 1805—1882	105	*Windsor Castle*（1843）	《厉鬼犯跸记》	毛文钟	1921	商务印书馆
61. 英国 威而司 （作者英文名待考）	106	英文书名待考	《鬼悟》	毛文钟	1921	商务印书馆
62. 英国 高尔忒 （作者英文名待考）	107	英文书名待考	《马妒》	毛文钟	1921	商务印书馆
63. 英国 卡文 （作者英文名待考）	108	英文书名待考	《沧波淹谍记》	毛文钟	1921	商务印书馆
64. 英国 路易 （作者英文名待考）	109	英文书名待考	《埃及异闻录》	毛文钟	1921	商务印书馆
65. 英国 伯明罕 George A. Birmingham 1865—1950	110	*The Island Mystery*（1918）	《沙利沙女王小纪》	毛文钟	1921	商务印书馆
66. 英国 道因 （作者英文名待考）	111	英文书名待考	《情海疑波》	林凯	1921	商务印书馆
67. 英国 泊恩 （作者英文名待考）	112	英文书名待考	《瞩目英雄》	毛文钟	1922	商务印书馆
68. 美国 鲁兰司 （作者英文名待考）	113	英文书名待考	《情翳》	毛文钟	1922	商务印书馆
69. 英国 克林登女士 （作者英文名待考）	114	英文书名待考	《情天补恨录》	毛文钟	1924	商务印书馆
70. 英国 巴文 Marjorie Bowen	115	*Carnival of Florence*（1915）	《妖髡缨首记》	毛文钟	1923	《小说世界》V2 N8/V3 N9
71. 美国 亨利 O. Henry 1862—1910	116	*The Gentle Grafter*（1904）	《善良的骗子》	不详	1925	《小说世界》V9 N1-13

从表1可以看出，林纾及其口译合作者所翻译的英语国家作家多达71位，其中包括了许多知名作家，如美国的Harriet Beecher Stowe（斯托夫人）、O. Henry（欧·亨利）、Washington Irving（华盛顿·欧文）、英国Sir Henry Rider Haggard（哈葛德）、Jonathan Swift（斯威佛特）、Charles Dickens（查尔斯·狄更斯）、Arthur Conan Doyle（科南·道尔）、William Shakespeare（威廉·莎士比亚）、Walter Scott（司各德）、Daniel Defoe（丹尼尔·迪福）。

4. 林纾翻译口译合作者

对林纾的翻译而言，其文风的吸引力，至少要从三个因素加以考虑：首先是合作者（不同合作者都会影响林纾对小说的选择和翻译方式）；其次是其本人作为翻译家有一个发展阶段；最后是小说原著的性质（Hanan 2004）。作为一位不懂外文的古文家，林纾的翻译完全依赖于精通不同语言的口译合作者，其口译合作者由于工作环境、文学修养等差异，翻译效果也大不同。在长达20多年的翻译活动中，林纾的口译合作者多达20多位，其中合作翻译数量多、影响广的口译者是魏易、陈家麟、曾宗巩、王庆通、毛文钟等人。这几个人当中，魏易（1880—1932）跟林纾的合作居功至伟，他的合作对于林纾小说乃至对林纾一生事业均有着重要影响。1981年商务印书馆纪念建馆85周年，重版了“林译小说”十本，其中有七本为魏易口述。

下面表格收集了林纾及其口译合作者的相关信息：

表2　与魏易合作翻译一览表

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
1	1. 英国哈葛德 Sir Henry Rider Haggard（1856—1925）	*Eric Brighteyes*（1891）	《埃司兰情侠传》	1904	广智书局
2		*Joan Haste*（1895）	《迦茵小传》	1905	商务印书馆
3		*Montezuma's Daughter*（1893）	《英孝子火山报仇录》	1905	商务印书馆
4		*Mr. Meeson's Will*	《玉雪留痕》	1905	商务印书馆
5		*Colonel Quaritch V.C.*（1888）	《洪罕女郎传》	1906	商务印书馆
6		*Beatrice*（1890）	《红礁画桨录》	1906	商务印书馆
7		*Dawn*（1884）	《橡湖仙影》	1906	商务印书馆

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
8	2. 英国 科南利达 Arthur Conan Doyle （1859—1930）	*Micah Clarke*（1889）	《金风铁雨录》	1907	商务印书馆
9		*A Study in Scarlet*（1887）	《歇洛克奇案开场》	1908	商务印书馆
10		*The Refugees*（1893）	《恨绮愁罗记》	1908	商务印书馆
11		*Uncle Bernac*（1897）	《髯刺客传》	1908	商务印书馆
12		*The Doings of Raffles Haw*（1892）	《电影楼台》	1908	商务印书馆
13		*Beyond the City*（1892）	《蛇女士传》	1908	商务印书馆
14		*The White Company*（1891）	《黑太子南征录》	1909	商务印书馆
15	3. 英国 却而司迭更司 Charles Dickens （1812—1870）	*Nicholas Nickleby*（1839）	《滑稽外史》	1907	商务印书馆
16		*The Old Curiosity Shop*（1841）	《孝女耐儿传》	1907	商务印书馆
17		*David Copperfield*（1850）	《块肉余生述，又续编》	1908	商务印书馆
18		*Oliver Twist*（1838）	《贼史》	1908	商务印书馆
19		*Dombey and Son*（1848）	《冰雪因缘》	1909	商务印书馆
20	4. 英国 司各德 Walter Scott （1771—1832）	*Ivanhoe*（1820）	《撒克逊劫后英雄略》	1905	商务印书馆
21		*The Talisman*（1825）	《十字军英雄记》	1907	商务印书馆
22		*The Betrothed*（1825）	《剑底鸳鸯》	1907	商务印书馆
23	5. 美国 华盛顿欧文 Washington Irving （1783—1859）	*The Sketch Book of Geoffrey Crayon, Gent.*（1820）	《拊掌录》	1907	商务印书馆
24		*The Alhambra*（1832）	《大食故宫馀载》	1907	商务印书馆
25		*Tales of a Traveller*（1824）	《旅行述异》	1907	商务印书馆
26	6. 英国 马支孟德 / 马尺芒忒 Arthur W. Marchmont （1852—1923）	*For Love or Crown*（1901）	《西利亚郡主别传》	1908	商务印书馆
27	7. 阿纳乐德 Thomas Arnold 1795—1842	*The Second Punic War*（1886）	《布匿第二次战纪》	1903	京师大学堂官书局

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
28	8. 英国 兰姆姐弟 Charles Lamb（1775—1834）；Mary Lamb（1764—1847）	*Tales from Shakespeare*（1807）	《吟边燕语》	1904	商务印书馆
29	9. 英国 洛加德 John Gibson Lockhart（1794—1854）	*History of Napoleon Buonaparte*（1829）	《拿破仑本纪》	1905	商务印书馆
30	10. 英国斯威佛特 Jonathan Swift（1667—1745）	*Gulliver's Travels*（1726）	《海外轩渠录》	1906	商务印书馆
31	11. 英国阿瑟毛利森 Arthur Morrison（1863—1945）	*Chronicles of Martin Heweitt*（1895）	《神枢鬼藏录》	1907	商务印书馆
32	12. 英国几拉德（作者英文名待考）	英文书名待考	《花因》	1907	中外日报馆
33	13. 大限克力司蒂穆雷 David Christie Murray（1847—1907）	*The Martyred Fool*（1895）	《双孝子噀血酬恩记》	1907	商务印书馆
34	14. 约翰沃克森罕（作者英文名待考）	英文书名待考	《天囚忏悔录》	1908	商务印书馆
35	15. 男爵夫人阿克西 Baroness Emma Orczy（1865—1947）	*The Scarlet Pimpernel*（1905）	《大侠红蘩蕗传》	1908	商务印书馆
36	16. 却洛得倭康，诺埃克尔司（作者英文名待考）	英文书名待考	《彗星夺婿录》	1909	商务印书馆
37	17. 蜚立伯倭本翰 E. Phillips Oppenheim（1866—1946）	*The Secret*（1907）	《藕孔避兵录》	1909	商务印书馆
38	18. 安东尼霍迫 Anthony Hope（1863—1933）	*A Man of Mark*（1890）	《西奴林娜小传》	1909	商务印书馆
39	19. 司丢阿忒（作者英文名待考）	英文书名待考	《脂粉议员》	1909	商务印书馆
40	20. 色东麦里曼 Henry Seton Merriman（1862—1903）	*From One Generation to Another*（1892）	《芦花馀孽》	1909	商务印书馆
41	21. 德国哈伯兰 Michael Haberlandt（1860—1940）	*Volkerkunde*（1898）	《民种学》	1903	京师大学堂官书局
42	22. 日本德富健次郎（1868-1927）	英文书名待考	《不如归》	1908	商务印书馆

表3　与陈家麟合作翻译一览表

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
1	1. 英国哈葛德 Sir Henry Rider Haggard 1856—1925	*Jess*（1887）	《玑司刺虎记》	1909	商务印书馆
2		*Fair Margaret*（1907）	《双雄较剑录》	1915	商务印书馆
3		*Benita*（1906）	《古鬼遗金记》	1912—13	商务印书馆
4		*The Ghost Kings*（1908）	《天女离魂记》	1917	商务印书馆
5		*The Brethren*（1904）	《烟火马》	1917	商务印书馆
6		*The Witch's Head*（1887）	《铁匣头颅，又续篇》	1919	商务印书馆
7		*Maiwa's Revenge*（1888）	《豪士述猎》	1919	商务印书馆
8		*The World's Desire*（1890）	《金梭神女再生缘》	1920	商务印书馆
9		*Queen Sheba's Ring*（1910）	《炸鬼记》	1921	商务印书馆
10	2. 英国莎士比 William Shakespeare 1564—1616	*Richard II*（1597）	《雷差得纪》	1916	《小说月报》V7 N1
11		*Henry IV*（1598—1600）	《亨利第四纪》	1916	《小说月报》V7 N2-4
12		*Henry VI*（1594—1623）	《亨利第六遗事》	1916	商务印书馆
13		*Julius Caesar*（1623）	《凯彻遗事》	1916	《小说月报》V7 N5-7
14		*Henry V*（1600）	《亨利第五纪》	1925	《小说世界》V12 N9-10
15	3. 英国马支孟德/马尺芒忒 Arthur W. Marchmont 1852—1923	*The Man Who Was Dead*（1907）	《黑楼情孽》	1914	商务印书馆

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
16	4. 美国锁司倭司女士/沙司卫甫夫人 Emma D. E. N. Southworth 1819—1899	*The Changed Brides*（1869）	《薄倖郎》	1915	商务印书馆
17	5. 英国杨支 Charlotte Mary Yonge 1823—1901	*The Dove in the Eagle's Nest*（1866）	《鹰梯小豪杰》	1916	商务印书馆
18	中英文作者信息待考	*The Lances of Lynwood*（1855）	《戎马书生》	1920	商务印书馆
19	6. 英国马克丹诺保德庆 M. McDonnel Bodkin 1850—1933	*The Quests of Paul Beck*（1908）	《贝克侦探谈，又续编》	1909	商务印书馆
20	7. 英国测次希洛 （作者英文名待考）	英文书名待考	《残蝉曳声录》	1914	商务印书馆
21	8. 英国马格内 （作者英文名待考）	英文书名待考	《石麟移月记》	1915	商务印书馆
22	9. 美国包鲁乌因 James Baldwin 1841—1925	*Thirty More Famous Stories Retold*（1905）	《秋灯谭屑》	1916	商务印书馆
23	10. 希登希路 （作者英文名待考）	英文书名待考	《红篋记》	1916	《小说月报》V7 N3-10
24	11. 克拉克 Mary Cowden Clarke 1809—1898	*The Thane's Daughter*（1850）	《奇女格露枝小传》	1916	商务印书馆
25	12. 美国巴苏谨 （作者英文名待考）	英文书名待考	《橄榄仙》	1916	商务印书馆
26	13. 倩伯司 W. & R. Chambers, Ltd.	*Chambers's Complete Tales for Infants*	《诗人解颐语》	1916	商务印书馆
27	14. 克拉克 Charles Cowden Clarke 1787—1877	*Tales from Chaucer in Prose*（2nd ed. 1870）	《坎特伯雷故事》	1916—1917	《小说月报》V7 N12/V8 N2-10

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
28	15. 利华奴 （作者英文名待考）	英文书名待考	《柔乡述险》	1917	《小说月报》V8 N1-6
29	16. 布司白 Guy Boothby 1867—1905	*A Brighton Tragedy*（1905）	《女师饮剑记》	1917	商务印书馆
30	17. 参恩女士 （作者英文名待考）	英文书名待考	《桃大王因果录》	1918	商务印书馆
31	18. 陈施利 （作者英文名待考）	英文书名待考	《牝贼情丝记》	1917	商务印书馆
32	19. 赖其镗女士 （作者英文名待考）	英文书名待考	《痴郎幻影》	1918	商务印书馆
33	20. 巴克雷 Florence L. Barclay 1862—1921	*The Rosary*（1909）	《玫瑰花，又续编》	1918	商务印书馆
34	21. 亚波倭得 Allan Upward 1863—1926	*The Phantom Torpedo-boats*（1905）	《赂史》	1920	商务印书馆
35	22. 美国尼可拉司 Nicholas Carter	英文书名待考	《焦头烂额》	1920	商务印书馆
36	23. 美森 （作者英文名待考）	英文书名待考	《妄言妄听》	1920	商务印书馆
37	24. 约克魁迭斯 （作者英文名待考）	英文书名待考	《西楼鬼语》	1919	商务印书馆
38	25. 武英尼 （作者英文名待考）	英文书名待考	《鬼窟藏娇》	1919	商务印书馆
39	26. 美国卡扣登 Charles Major 1856—1913	*When Knighthood Was In Flower*（1898）	《莲心藕缕缘》	1919	商务印书馆
40	27.（美）堪伯路 （作者英文名待考）	英文书名待考	《还珠艳史》	1920	商务印书馆
41	28. 达威生 Gladys Davidson	*Stories from the Operas*（1914）	《泰西古剧》	1920	商务印书馆
42	29. 高桑斯 （作者英文名待考）	英文书名待考	《欧战春闺梦，又续编》	1920	商务印书馆

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
43	30. 伯鲁夫因支 Thomas Bulfinch 1796—1867	外文书名待考	《怪董》	1921	商务印书馆
44	31. 斐鲁丁 Henry Fielding 1707—1754	*A Journey from this World to the Next*（1743）	《洞冥记》	1921	商务印书馆
45	32. 周鲁倭 （作者外文名待考）	外文书名待考	《情天异彩》	1919	商务印书馆
46	33. 魁特 （作者外文名待考）	外文书名待考	《俄宫秘史》	1921	商务印书馆
47	34. 俄国托尔斯泰 Лев Толстой 1828—1910	外文书名待考	《罗刹因果录》	1915	商务印书馆
48		外文书名待考	《社会声影录》	1917	商务印书馆
49		*Люцери*（1857）	《路西恩/琉森》	1917	《小说月报》V8 N5
50		*Смерть Ивана Ильича*（1886）	《人鬼关头/伊凡·伊里奇之死》	1917	《小说月报》V8 N7-10
51		外文书名待考	《恨缕情丝》	1918	《小说月报》V9 N1-11
52		*Детство*（1852），*Отрочество*（1854），*Юность*（1857）	《现身说法/幼年·少年·青年》	1918	商务印书馆
53		*Записки маркера*（1855）	《球房纪事》	1920	《小说月报》V11 N3
54		*Альберт*（1857）	《乐师雅路白忒遗事》	1920	《小说月报》V11 N4
55		*Кавказкий пленник*（1872）	《高加索之囚》	1920	《小说月报》V11 N5
56		*Три смери*（1858）	《三种死法》	1924	《小说世界》V5 N1

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
57	35. 瑞士大卫威司 Johann David Wyss 1743—1818	*Der schweizerische Robinson*（1813）	《鹧巢记，又续编》	1920	商务印书馆
58	36. 西班牙西万提司 Miguel de Cervantes Saavedra 1547—1616	*Don Quixote de Ia Mancha, 1*（1605）	《魔侠传/堂吉诃德》	1922	商务印书馆
59	37. 大威森（作者外文名待考）	外文书名待考	《拿云手》	1917	《小说海》V8 N1-8

表4　与曾宗巩合作翻译一览表

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
1	1. 英国哈葛德 Sir Henry Rider Haggard 1856—1925	*Nada the Lily*（1892）	《鬼山狼侠传》	1905	商务印书馆
2		*Cleopatra*（1889）	《埃及金塔剖尸记》	1905	商务印书馆
3		*Allan Quatermain*（1887）	《斐洲烟水愁城录》	1905	商务印书馆
4		*Black Heart and White Heart and Other Stories*（1900）	《蛮荒志异》	1906	商务印书馆
5		*People of the Mist*（1894）	《雾中人》	1906	商务印书馆
6		*King Solomon's Mines*（1885）	《钟乳髑髅》	1908	商务印书馆
7		*She*（1886）	《三千年艳尸记》	1910	商务印书馆

（待续）

（续表）

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
8	2. 英国达孚 Daniel Defoe 1660—1731 两种	*Life and Strange Surprising Adventures of Robinson Crusoe*（1719）	《鲁滨孙漂流记》	1905	商务印书馆
9		*Farther Advantures of Robinson Crusoe*（1719）	《鲁滨孙漂流续记》	1906	商务印书馆
10	3. 美国阿丁 William L. Alden 1837—1908	*Jimmy Brown Trying to Find Europe*（1889）	《美洲童子万里寻亲记》	1904	商务印书馆
11	4. 路易司地文，佛尼司地文 Robert Louis Stevenson 1850—1894；Fanny Van de Graft Stevenson 1840—1914	*More New Arabian Nights: The Dynamiter*（1885）	《新天方夜谭》	1908	商务印书馆
12	5. 麦里郝斯 Sophia H. Maclehose ?—1912	*Tales from Spenser, Chosen from the Faerie Queene*（1890）	《荒唐言》	1914	商务印书馆
13	6. 阿猛查登 Erckmann-chatrian	*Histoire d'un conscrit de*（1813）	《利俾瑟战血余腥记》	1904	文明书局
14	作者中外文名待考	*Waterloo: suite de Conscrit de 1813*（1865）	《滑铁庐战血余腥记》	1904	文明书局

表5　与王庆通合译一览表

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
1	1. 小仲马 Alexandre Dumas，fils，1824—1895	*Antonine*（1849）	《香钩情眼》	1916	商务印书馆
2		*L'Affaire C1émenceau*（1866）	《血华鸳鸯枕》	1916	《小说月报》V7 N8-12
3		*Aventures de quatre femmes et d'un perroquet*（1846—47）	《鹦鹉缘，又续编、三编》	1918	商务印书馆
4		*La boite d'argent*（1855）	《伊罗埋心记》	1920	《小说月报》V11 N1-2
5		*Le docteur servans*	《无原可作》	1919	《妇女杂志》V8 No1-12
6	2. 大仲马（Alexandre Dumas，pèfe，1802—1870）	*Une fille du régent*（1845）	《蟹莲郡主传》	1915	商务印书馆
7	3. 孟德斯鸠（Charles Louis de Secondat Montesquieu，1689—1755）	*Lettres Persanes*（1721）	《鱼雁抉微》	1915—1917	《东方杂志》V9-14
8	4. 辟厄喀珉（Pierre Loti，1850—1923）	*Pécheur d'Islande*（1886）	《鱼海泪波》	1915	商务印书馆
9	5. 爽梭阿过伯（Francois Coppée，1842—1908）	*Le coupable*（1897）	《溷中花》	1915	商务印书馆
10	6. 海斯班（Jean Richepin，1849—1926）	*Monsieur Destrémeaux, roman psychoIogique*（1882）	《白夫人感旧录》	1917	《小说月报》V8 N11-12
11	7. 丹米尔、（俄）华伊尔		《金台春梦录》	1918	商务印书馆
12	8. 比利时恩海贡斯翁士（Hendrick Conscience，1812—1883）	*De arme edelman*（1851）	《孝友镜》	1918	商务印书馆

表6　与毛文钟合作翻译一览表

序号	作者信息	原著书名及出版时间	中文书名	出版年	出版社/出处
1	1. 美国克雷夫人 Bertha M. Clay	书名外文名待考	《想夫怜》	1920	《小说月报》V11 N9-12
2		书名外文名待考	《僵桃记》	1921	商务印书馆
3	2. 安司倭司 William Harrison Ainsworth 1805—1882	*Windsor Castle*（1843）	《厉鬼犯跸记》	1921	商务印书馆
4	3. 威而司（作者外文名待考）	书名外文名待考	《鬼悟》	1921	商务印书馆
5	4. 高尔忒（作者外文名待考）	书名外文名待考	《马妒》	1921	商务印书馆
6	5. 卡文（作者外文名待考）	书名外文名待考	《沧波淹谍记》	1921	商务印书馆
7	6. 路易（作者外文名待考）	书名外文名待考	《埃及异闻录》	1921	商务印书馆
8	7. 伯明罕 George A. Birmingham（1865—1950）	*The Island Mystery*（1918）	《沙利沙女王小纪》	1921	商务印书馆
9	8. 泊恩（作者外文名待考）	书名外文名待考	《瞿目英雄》	1922	商务印书馆
10	9.（美）鲁兰司（作者外文名待考）	书名外文名待考	《情翳》	1922	商务印书馆
11	10. 克林登女士（作者外文名待考）	书名外文名待考	《情天补恨录》	1924	商务印书馆
12	11. 巴文 Marjorie Bowen	*Carnival of Florence*（1915）	《妖髡缨首记》	1923	《小说世界》V2 N8/V3 N9
13	12. 预勾（Victor Hugo，1802—1885）	*Quatre-vingt-treize*（1874）	《双雄义死录》	1921	商务印书馆
14	13. 挪威伊卜森（Henrik Ibsen，1828—1906）	*Gengangere*（1881）	《梅孽/群鬼》	1921	商务印书馆

5. 语料库分词与翻译策略标注

一个语料库的价值除了所收集的样本之外，还体现在对所收集的语料样本进

行标注。这是对语料库样本进行新的增值。语料库信息标注能为研究者提供丰富的信息。这类信息通过语料库检索软件，能够迅速进行提取，展开相应的统计分析，这是基于语料库基础上进行研究的便捷之处。

林纾的译作，繁体、竖排、无标点，很多印刷质量不高，部分文字无法辨识，这是前期加工过程中的难点，耗时较多。语料库建成之后的目的很明确：帮助我们快捷检索与开展文本分析。根据现有的语料库分析工具，我们在后期加工过程中，对语料库进行分词与标注（重点标注林纾翻译过程中采取的各类添加或者删减标记，而不是简单的词性标注）。

语料库标注层次的细化与采用科学的标注方法，将对今后在此平台上展开的林纾翻译研究提供非常重要的帮助。林纾翻译语料库与一般的翻译语料库相比，有其特殊性，原因在于林纾不通外语，其“翻译”出来的译文是依赖口译者对源语的翻译，在翻译过程中留下的“译者”痕迹是很明显的。另外，林纾本身所具有的桐城派文笔，对其翻译也产生了重要的影响。

基于以上分析，我们对林纾翻译语料库展开的信息标注，包含以下几个方面：

1）元信息的标注：包括源语标题、源语作者（含国别）、翻译合作者名字、翻译时间、出版时间（含出版社等相关信息）；

2）译者按语的标注；

3）语法信息标注；

4）翻译策略标注；

5）特殊信息如特定时代各种语言信息标注。

其他各类标注，将随着语料库样本收集的扩展而逐步展开。

6. 结论

在林纾翻译语料库基础上，后续研究还可以进行多译本对照研究。以林纾译本为依托，其他现代汉语译本为对照，将英语原文、林纾译本、现代汉语译本进行对比研究，这样就可实现译者对比以及文言与白话对比，还可以进行英语、汉语文言、汉语白话对比。

林纾译文语料库的创建，将为翻译学界提供一种“解读林纾”的新手段与方法，也将为翻译学界提供全新的研究成果，为全方位认识与评价林纾提供客观、科学的数据。这将在描写翻译学研究领域内提供一个可供参照的研究范式，为翻译名家研究、翻译风格研究、翻译语言特征分析、翻译策略研究等多方面提供新的视角。

注　释

1. 福建工程学院肇始于1896年创办的苍霞精舍。据《福建通志》记载，1896年，清邮传部尚书陈璧返闽主讲凤池书院时，与举人林纾以及商部郎中力钧、补用知府孙葆晋等人，在苍霞洲林纾旧居创设苍霞精舍，这是“福州有学校之始”。对于创始人之一的林纾，福建工程学院投入了大量的人力物力进行挖掘，在社会各界的帮助下，馆藏林纾专题资源库涵盖了林纾古诗文及小说创作、文学翻译、教育读本、字画集等。

参考文献

Boase-Beier, J. 2006. *Stylistic Approaches to Translation* [M]. Manchester: St. Jerome Publishing.

Hanan, P. 2004. *Chinese Fiction of the Nineteenth and Early Twentieth Centuries* [M]. New York: Columbia University Press.

Ho, Y. 2011. *Corpus Stylistics in Principles and Practice: A Stylistic Exploration of John Fowles' the Magus* [M]. London: Continuum.

Huang, L. 2015. *Style in Translation: A Corpus-based Perspective* [M]. Heidelberg: Springer.

Ji, M. 2009. Corpus stylistics in translation studies: two modern Chinese translations of *Don Quijote* [J]. *Language and Literature* 18(1): 61-73.

Malmkjær, K. 2004. Translational stylistics: Dulcken's translations of Hans Christian Andersen [J]. *Language and Literature* 13(1): 13-24.

Studer, P. 2008. *Historical Corpus Stylistics: Media, Technology and Change* [M]. London: Continuum.

陈平原，1989，《二十世纪中国小说史·第一卷（1897—1916年）》[M]。北京：北京大学出版社。

陈玉刚，1989，《中国翻译文学史稿》[M]。北京：中国对外翻译出版公司。

戴光荣，2015，描写翻译学视阈下林纾译文语料库的创建 [J]，《福建工程学院学报》（5）：419-422。

戴光荣，2018a，从语料库视角看中国文学作品“走出去”[N]，《中国社会科学报》，10月19日第四版。

戴光荣，2018b，民族危机下爱国情怀的抒写：林纾翻译语料库的序跋词表分析 [J]，《福州大学学报（哲学社会科学版）》（6）：86-90。

戴光荣，2018c，清末民初翻译对现代性的促进——从林纾翻译序跋谈 [J]，《中国科技翻译》（4）：66-69。

胡开宝，2011，《语料库翻译学概论》[M]。上海：上海交通大学出版社。

连燕堂，2009，《二十世纪中国翻译文学史：近代卷》[M]。天津：百花文艺出版社。林薇，1990，《百年沉浮——林纾研究综述》[M]。天津：天津教育出版社。

马泰来，1981，林纾翻译作品全目 [A]，载《林纾的翻译》[C]。北京：商务印书馆，60-103。

马泰来，1982，关于《林纾翻译作品全目》[J]，《读书》（10）：140-142。

马泰来，2013，罗香林教授和我的林纾翻译研究 [J]，《天禄论丛——中国研究图书馆员学会学刊》（3）：53-59。

孟昭毅、李载道，2005，《中国翻译文学史》[M]。北京：北京大学出版社。

王克非，2012，《语料库翻译学探索》[M]。上海：上海交通大学出版社。

王向远，2015，“译文不在场”的翻译文学史——“译文学”意识的缺失与中国翻译文学史著作的缺憾 [J],《文学评论》(3)：65-71。

王　勇，2009，林纾与《东方杂志》[J],《福建工程学院学报》(5)：425-429。

谢天振、查明建，2004，《中国现代翻译文学史（1898-1949）》[M]。上海：上海外语教育出版社。

查明建、谢天振，2007，《中国20世纪外国文学翻译史（上下卷）》[M]。武汉：湖北教育出版社。

张俊才，2007，《林纾评传》[M]。北京：中华书局。

通信地址：350118 福建省福州市 福建工程学院人文学院

Python词向量训练与应用技术解析*

北京外国语大学/赣南师范大学 邓海龙

提要：词向量具有优越的语义表示性能。在大数据时代，词向量技术在语言研究中蕴含着广阔的应用前景。本文面向无编程经验的外语教学科研人员，简要介绍自然语言处理中的词向量技术的操作流程，具体包括Python环境配置、Gensim模块安装、词向量训练、保存、加载、应用及其可视化等步骤。本文提供注释完整、可直接运行的Python源代码，读者可根据需要设置相关参数，用于个人的词向量应用研究。

关键词：Python、词向量、训练、应用

1. 引言

词向量（word embedding，又称词嵌入）是自然语言处理中用于词汇表示的技术。2013年，谷歌（Google）公司研发了Word2vec工具[1]，用于从大规模语料中训练词向量（Mikolov *et al.* 2013a，2013b，2013c）。评测数据显示，由Word2vec训练得到的词向量在词义相似度计算与类比推理方面性能优异。词向量语义特征是基于上下文语境共现词汇信息经过神经网络计算而得，其语言学理论基础是分布式语义观（Harris 1954），与语料库语言学具有较好的契合度。词向量技术甚至被认为是揭开神经机器翻译“黑箱”的钥匙（冯志伟 2019）。最近，有研究人员基于词向量开展词汇语义演变研究（Hamilton *et al.* 2016）；也有研究者揭示出词向量中隐含性别、职业等歧视内容（Caliskan *et al.* 2017）；还有学者利用词向量进行话语分析（Liu & Lei 2018）。可见，词向量在语言学研究中的应用方兴未艾，潜力无限。

词向量训练与使用一般要求研究人员具有一定的编程经验。Word2vec原作者公布的源码由C语言编制而成，对非专业程序员而言，学习难度非常大。幸运的是，Gensim包提供了易于操控、效率较高的Python词向量训练与使用模块[2]。Python语法简洁易懂，是当前人工智能及深度学习常用的编程语言，拥有强大的交流社区和丰富的开放资源。国内已有专门书籍介绍Python在语料库研究中的应

* 本文系教育部人文社会科学重点研究基地重大项目子课题“大数据视野下的外语及外语学习研究”（17JJD740003）的阶段性成果。

用（管新潮 2018）。本文不要求读者具有任何编程经验，从零开始介绍Python词向量训练与使用，以期有助于词向量技术在语言研究中得到更多应用。

本文主要介绍Python环境配置、Gensim包安装、词向量训练与使用。笔者已经提供了编写完整，便于配置，可以直接运行的配套Python源代码文件，读者可从北外语料库语言学网站或者Github源代码托管平台下载[3]。由于Python是跨平台脚本语言，本文提供的源码也可以在Linux、Mac OS等不同系统环境中直接运行。不过，需要注意的是，文中所有操作说明都是基于Windows 10操作系统，使用其他系统的读者需要酌情区别处理。另外，因为教程各个环节存在先后依赖关系，对没有任何编程经验的读者，最好按照教程顺序依次完成，以免出错。

2. Python环境配置及编程基础

2.1 环境配置

我们通过安装Anaconda包管理软件快速配置Python编程环境。Anaconda软件的下载链接是：https://www.anaconda.com/distribution/。打开网页，根据系统情况，选择合适的版本进行下载（见图1）。在安装过程中，程序会有各种提示，一般接受默认设置即可，但要注意勾选图2中的两个勾选项，以设置环境变量。

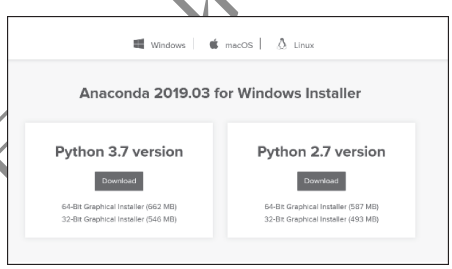


图1 Anaconda下载界面

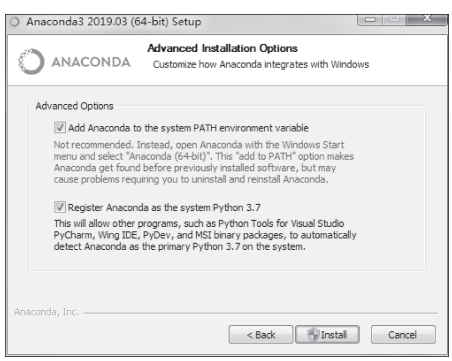


图2 Anaconda安装选项

2.2 Jupyter Notebook

安装好Python开发环境之后，开发人员可以使用简单的记事本编写代码，然后在命令行下调用。但是，为了提高开发效率，我们通常使用集成开发环境（Integrated Development Environment，IDE），将代码编写、运行和调试等功能整合在一个编程工具中实现。Python编程开发工具很多，如常用的PyCharm[4]，Python安装包自带的IDLE[5]，以及集成在Anaconda包中的Spyder[6]等等。本文使用Jupyter Notebook[7]进行Python编程。Jupyter Notebook是一个以网页为界面的交互式编程工具，它将每一个源代码文件以笔记本（notebook）形式显示与保存。在Jupyter Notebook编程界面，用户能够动态修改程序代码，输出各种图表，插入丰富的文档注释，非常适合科研人员进行探索性数据分析。

2.2.1 启动程序

按常规软件使用方式，我们可以通过打开“开始”菜单，从弹出的列表中找到Anaconda下面的Jupyter Notebook快捷键进行启动。但是，这种启动方式不便选择当前文件夹路径，本文推荐使用下面的启动方式。

（1）打开自己准备存放代码的文件夹。例如，将本文提供的“Python词向量学习1”文件夹放在E盘根目录下面（当然也可以选择其他任意地址），见图3。

图3 “Python词向量学习1”文件夹目录

（2）单击所在文件夹的地址栏，地址栏变为高亮可编辑状态（如下图4所示）。

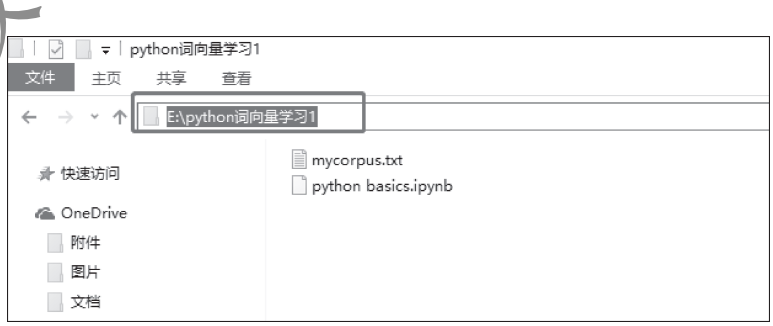


图4 单击高亮地址栏

（3）在地址栏输入“cmd”（图5），回车后系统弹出命令行窗口（图6）。

图5　在地址栏输入“cmd”命令

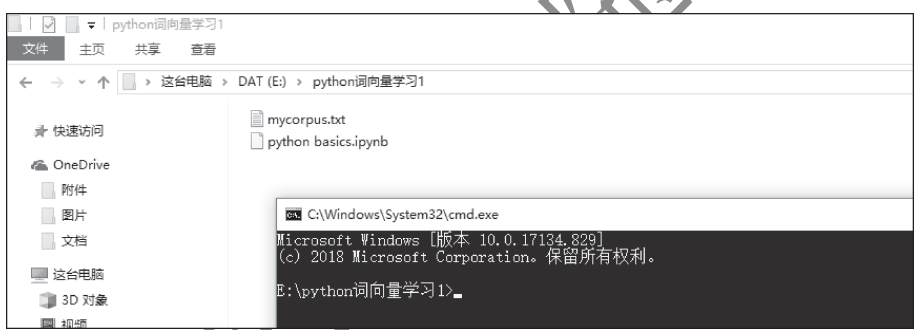


图6　调出命令行窗口

（4）在弹出的命令行窗口中输入“jupyter notebook”（图7），回车，稍等片刻，系统默认浏览器就会打开Jupyter Notebook以Home为标题的网络界面（图8）。因为文件夹里有两个文件，所以Home网页的Files标签分页中显示了两个文件名。

图7　在命令行窗口输入“jupyter notebook”命令

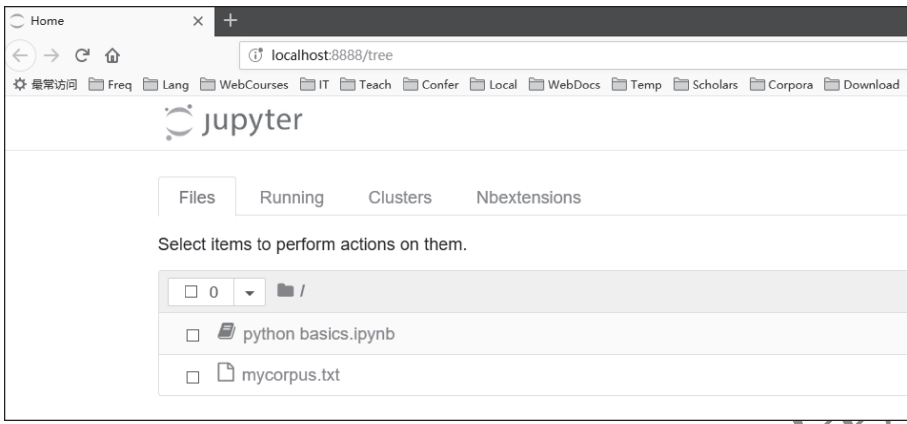


图8　Jupyter notebook的Home主页

2.2.2　Notebook使用说明

（1）点击Home页面右边的New按钮，弹出下拉菜单，单击Python 3（图9），程序弹出ipython notebook新建页面（图10）。

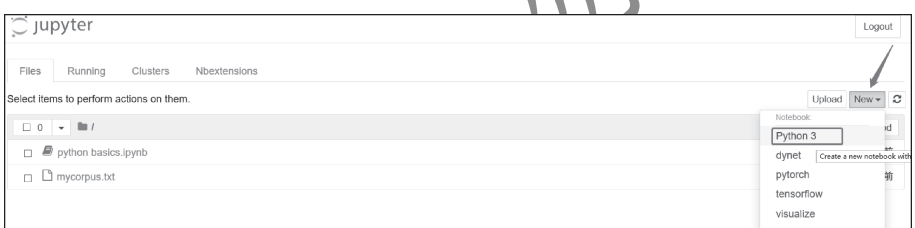


图9　新建notebook页面

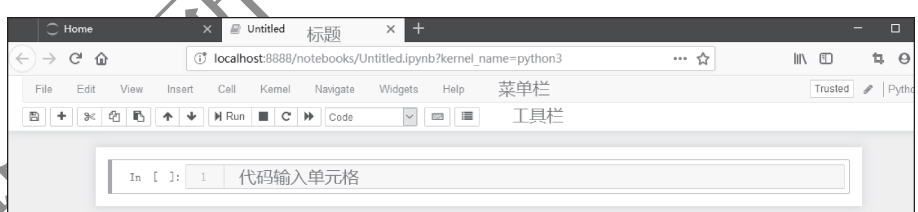


图10　Notebook新建页面及各个功能区

（2）该新建页面是代码编写、运行与调试的主要区域。如图10所示，页面中包括标题、菜单栏、工具栏与代码输入单元格等四个主要部分。

标题是指该ipython notebook文件的名称。新建笔记本默认标题为“Untitled”。此时，若切换回Home页面，我们可以看到一个名为“Untitled.ipynb”的文件显示在文件列表中；也可以在“Python词向量学习1”文件夹中发现一个名为“Untitled.ipynb”的文件。

菜单栏里有很多菜单按钮，分别提供不同功能。这些菜单名称常常见名知义，读者可自行搜索相关介绍并在反复练习中不断熟悉。这里主要介绍三个下拉菜单中我们后面会用到的一些按钮。

首先是“File”下拉菜单栏中的“Rename”按钮（图11），点击后能够修改当前笔记本的名称。我们将当前笔记本更名为“test”，该笔记本的标题、Home页面中的文件列表以及文件夹中的名称全部改为“test.ipynb”，见图12。

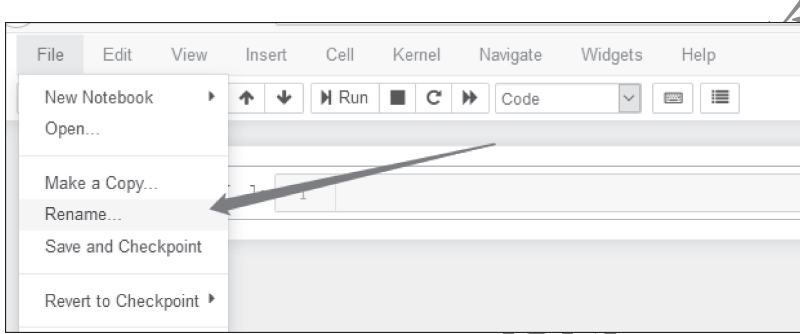


图11　Notebook重命名

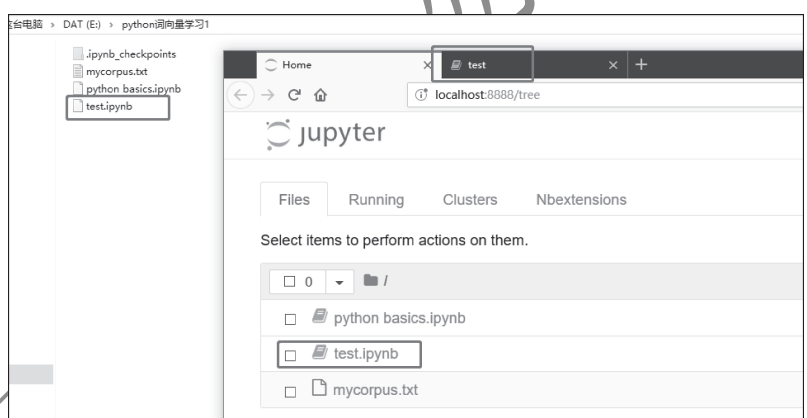


图12　Notebook标题名的三处显示

其次是“Insert”下拉菜单栏中的“Insert Cell Above”以及“Insert Cell Below”两个按钮（图13），可以在指定代码单元格前面（或者后面）插入新的空白单元格。

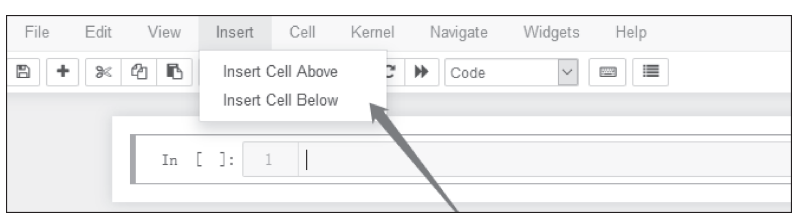


图13　插入单元格操作菜单

最后是“Cell”下拉菜单中的“Run All”等按钮（如图14所示），可以批量运行单元格代码。

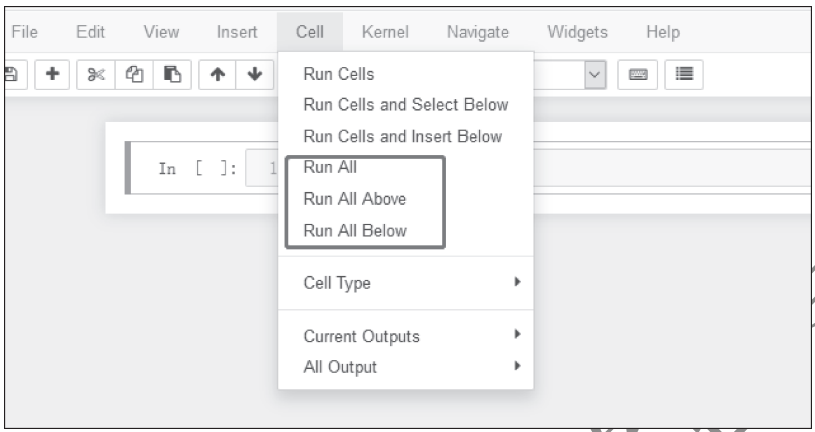


图14　运行单元格菜单

工具栏提供了一些快捷功能。只要将光标移至这些工具的图标上面停留片刻，页面就会弹出显示该工具的相关功能说明。其中常用的一个工具是“Run”按钮，点击后运行当前光标所在单元格代码。

代码输入单元格是程序编写的主要区域。

2.3　Python编程简介

这里提供Python编程入门知识，以方便没有任何Python编程经验的读者更好地理解和使用后面的程序代码。

2.3.1　运行与输出

Jupyter Notebook以单元格为基础运行代码。运行方式如下：单击单元格代码，使该单元格处于活动状态，然后点击工具栏中的“Run”按钮，若无出错中断，系统将运行该单元格内所有代码。运行单元格代码的快捷键方法是：点击单元格，然后按住shift键同时回车。在Python中，输出函数是“print”。例如，若要程序输出“hello world！”字符串，我们只需在前面示例打开的“test.ipynb”笔记本第一个单元格中输入“print（'hello world!'）”，然后按前述方式运行该单元格，系统将输出“hello word!”，见图15。

图15　Python运行与输出

2.3.2 字符串处理

切换到Home主页，双击打开“python basics.ipynb”，我们以其中的代码为例说明Python中的字符串处理。字符串处理对于语料自动分析非常重要。Python字符串以单引号或者双引号进行创建。例如，“a = 'I love python 3'”表示将字符串“I love python 3”赋值给变量“a”。在运行完这行代码后，a的值就是“I love python 3”这个字符串。如果在新的单元格中输入“a”，然后运行，系统就会打印出变量“a”里面的内容“I love python 3”，见图16。

```
In [1]: 1 a = 'I love python 3'

In [2]: 1 a
Out[2]: 'I love python 3'
```

图16 Python字符串赋值

Python字符串切片处理非常简便，既可以指定切片，如取字符串从第2个字符到第6个字符（从0开始计数，不包含最后一个），也可以根据空格等特定字符进行分割，见图17。

```
In [3]: 1 a[2:6]
Out[3]: 'love'

In [4]: 1 a.split()
Out[4]: ['I', 'love', 'python', '3']
```

图17 Python字符串切片与分割

Python字符串合并更加直观，只需用“+”号将两个字符串连接起来即可，如图18。

```
In [5]: 1 b ='very good'

In [6]: 1 a+b
Out[6]: 'I love python 3very good'

In [7]: 1 a + ', and you love python too'
Out[7]: 'I love python 3, and you love python too'
```

图18 Python字符串合并

Python字符串变量包含了内置函数，可以直接进行大小写操作，如下图19。

```
In  [8]:  1  a.lower()
Out[8]:  'i love python 3'

In  [9]:  1  a.upper()
Out[9]:  'I LOVE PYTHON 3'
```

图19　Python字符串内置操作函数

2.3.3　文本文件读取

文本文件读取在计算机语言分析中不可或缺。Python文件读取用一行代码就可以实现，如下图20所示。若要读取其他文件，只要将英文括号内的“mycorpus.txt”换成要读取文件的路径即可。

```
In  [10]:  1  open('mycorpus.txt').readlines()
Out[10]:  ['Human machine interface for lab abc computer applications\n',
           'A survey of user opinion of computer system response time\n',
           'The EPS user interface management system\n',
           'System and human system engineering testing of EPS\n',
           'Relation of user perceived response time to error measurement\n',
           'The generation of random binary unordered trees\n',
           'The intersection graph of paths in trees\n',
           'Graph minors IV Widths of trees and well quasi ordering\n',
           'Graph minors A survey\n']
```

图20　Python文本文件读取

如果要对读取内容按行进行一些特定处理，我们可以使用for循环逐行操作。图21示例为将文本中所有单词都小写。

```
In  [11]:  1  for line in open('mycorpus.txt'):
           2      line = line.strip()
           3      line = line.lower()
           4      print(line)

human machine interface for lab abc computer applications
a survey of user opinion of computer system response time
the eps user interface management system
system and human system engineering testing of eps
relation of user perceived response time to error measurement
the generation of random binary unordered trees
the intersection graph of paths in trees
graph minors iv widths of trees and well quasi ordering
graph minors a survey
```

图21　Python文本文件逐行读取与操作

3. 安装Gensim包

Gensim[8]是由Python语言编写的开源语义计算工具包，集成了主题建模和词向量等文本语义分析工具。和其他Python包一样，Gensim的安装方法很简单。我们只要打开任意命令行窗口，键入“pip install gensim”，然后回车，系统就会自动从网上下载相关文件进行安装（图22），根据网络情况，所耗时间有所差异。

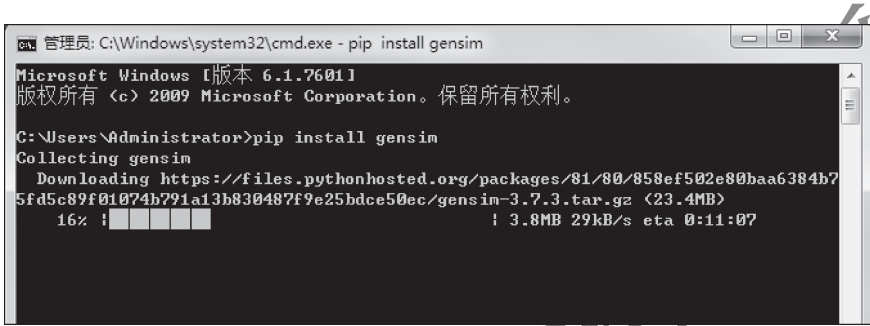


图22　用pip安装Gensim包

4. 词向量训练

使用词向量之前，我们需要大量语料对词向量进行训练。本部分将介绍语料预处理，训练参数设置以及启动词向量训练程序等相关操作。

4.1　启动程序

（1）打开本文提供的“python词向量学习2”文件夹。此时文件夹内只有“genism_word2vec_training.ipynb”文件。

（2）按照前面推荐的Jupyter Notebook启动方法，在文件夹路径上输入“cmd”调出命令行，并在弹出的命令行终端键入“jupyter notebook”启动Jupyter Notebook。系统在默认浏览器中打开Jupyter Notebook的Home主页，文件列表里显示有“genism_word2vec_training.ipynb”文件，如下图23所示。

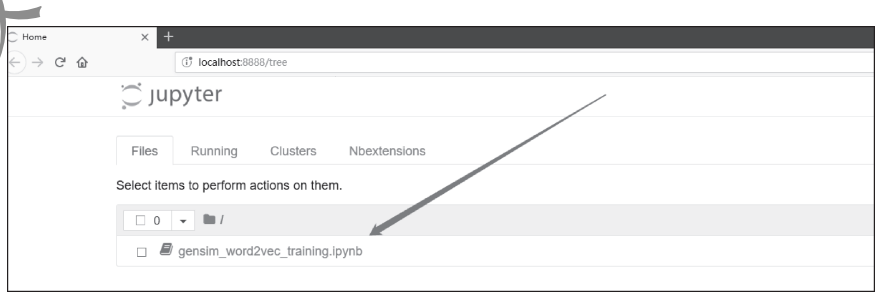


图23　在“python词向量学习2”文件夹内启动Jupyter Notebook后的主页

（3）双击打开“genism_word2vec_training.ipynb”文件，如下图24所示。

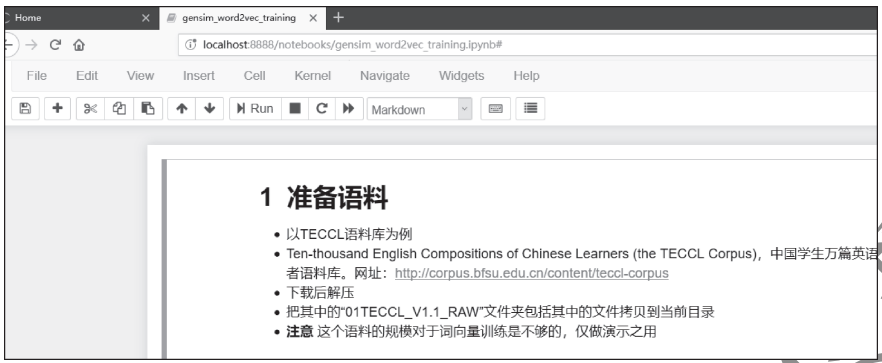


图24 “genism_word2vec_training.ipynb”文件显示界面

4.2 准备语料

本文使用中国学生万篇英语作文语料库V1.1（Ten-thousand English Compositions of Chinese Learners，简称TECCL corpus）进行词向量训练操作演示。TECCL语料库是由北外许家金教授创建的开放获取学习者语料库（许家金 2016），可从北外语料库网站免费下载[9]。我们将TECCL语料库下载后解压，并把其中的“01TECCL_V1.1_RAW”文件夹（包括里面所有文件）整个复制到“python词向量学习2”文件夹内（如图25所示）。需要指出的是，该语料规模并不足以充分训练词向量，此处仅作快速演示之用。

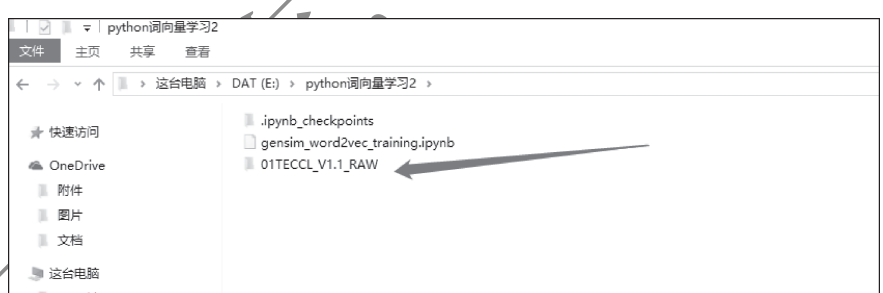


图25 复制TECCL语料库

4.3 参数设置

一般情况下，用户只需在这部分设置词向量训练参数、语料地址、模型保存地址以及语料库预处理要求等内容。

4.3.1 词向量参数

词向量参数包括五项内容（图26）。修改时只要改动其中的数值，注意保留后面的英文逗号。

（1）向量维度（size）。该参数指词向量的维度。若设置为100，说明每个词的词汇语义特征用100个数来表示。

（2）窗口大小（window）。该参数指训练时使用的上下文窗口大小。Gensim官方文档的注释是：当前词与预测词之间在同一句子中的最大距离（Maximum distance between the current and predicted word within a sentence）。一般窗口越小，词向量捕捉到更多词法及句法特征；窗口越大，则包含更多主题相关信息。默认窗口大小设置为5。

（3）最小词频（min_count）。语料中低于此频数的单词不列入词向量训练。默认最小词频为5。

（4）迭代次数（iter）。该参数用于设置词向量训练时的迭代次数。每迭代一次，程序对整个语料训练完成一次。一般情况下，若语料数据较小，为提升训练效果，可适当增加迭代次数。默认迭代次数为5。

（5）训练算法（sg）。词向量训练主要有两种训练算法，一种是跳字模型（skip-gram），以中心词预测上下文单词；另一种是连续词袋模型（continuous bag of words，简称CBOW），以上下文单词预测中心词（Mikolov *et al*. 2013a：5）。一般地，连续词袋模型的训练速度快，但跳字模型的训练效果较好，在语料数量较小的情况下尤其明显。该参数若设置为1，训练算法为跳字模型，0则为连续词袋模型。考虑到语言研究中的语料库通常规模较小，默认训练算法设置为1，即跳字模型。

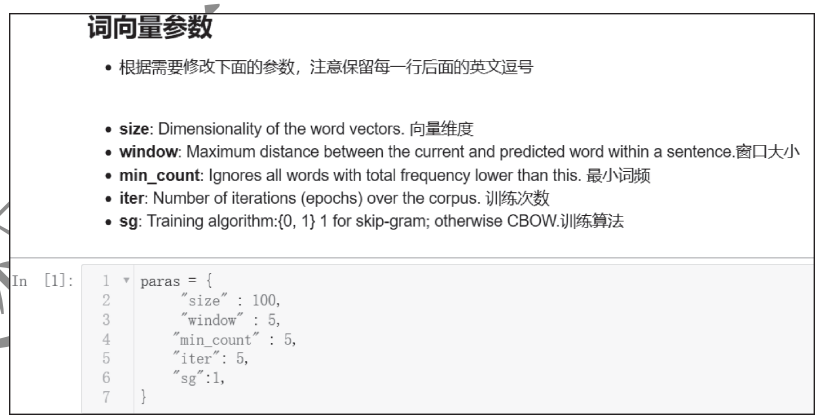


图26　词向量参数设置

4.3.2　语料库地址

修改填写“corpus_path = ”后面英文引号里面的地址内容便可以设置语料库地址。语料库地址是指用于训练词向量的语料所在目录地址。该目录下面的语料

由一个或多个以“.txt”为后缀的文本文件组成。语料地址可以是相对路径（以“genism_word2vec_training.ipynb”文件的路径为参照），也可以是绝对路径（从盘符开始的完整路径），但要注意，Python的文件路径分割符号是“/”，不同于Windows系统中的“\”，正确写法样例如“E:/word2vec/corpus”。

因为用于演示的TECCL语料库放在程序源代码同一目录，所以只需要填写语料库目录名称即可，如图27所示。

```
In [2]: 1  corpus_path = "01TECCL_V1.1_RAW"
```

图27　语料库地址

4.3.3　词向量保存地址

词向量训练一般耗时较长，通常将训练结果保存为一个文件，以方便后续使用。与语料库地址类似，词向量地址可以使用相对路径与绝对路径。需要注意的是，语料库地址是一个目录地址，但词向量保存地址是一个文件名地址。词向量保存文件名若以“.txt”与“.bin”为后缀，程序将以谷歌Word2vec工具中的词向量格式进行保存，否则以Gensim默认格式保存。以谷歌Word2vec格式保存的词向量文件通常便于其他词向量工具加载和调用。

词向量保存地址是在“saved_path = ”后面的英文引号里面修改填写。我们现在将保存地址设置为当前目录下命名为“teccl.txt”的文件，如图28所示。

```
In [3]: 1  saved_path = "teccl.txt"
```

图28　词向量保存地址

4.3.4　语料预处理

词向量训练是以词为单位，以句子为组别进行的神经网络语言模型训练，即单词上下文语境只考虑同一句子中给定窗口大小内的其他单词。因此，如何界定单词和分割句子将会影响词向量训练结果。具体而言，语料预处理主要涉及是否分句、分词、统一大小写或者去除停用词等设置。去除停用词通常是为了去除冠词、介词等虚词，保留实义词，以突显词向量的语义特征维度。

这个部分主要根据情况填写“True”或“False”（注意首字母大写）。其中，分句与分词功能都是调用NLTK工具包实现的。去除停用词若填写为“True”，可以选择自己定义停用词表，也可以调用默认的NLTK停用词表。若调用默认的NLTK停用词表，stopwords赋值为None，否则将自己定义的停用词表以Python列表的形式赋值给stopwords，例如填写为stopwords = [‘is’，‘a’，‘the’，‘an’]。

4.4 运行训练

因为笔者已经对Gensim词向量训练源码进行了封装，读者不用关心代码实现细节，只要在前一环节结合研究需要完成参数设置就可直接运行。

运行方法如下：单击菜单栏上的Cell按钮，然后在弹出的下拉菜单里点击Run All。程序将从整个代码笔记本的第一行单元格开始，依次执行代码。最后一行代码是词向量训练的主函数，耗时较长，在行首显示为“*”星号，表明程序正在运行中（如下图29）。当训练顺利完成，星号变成数字之后，我们可以在当前目录（“python词向量学习2”文件夹下）找到已经训练好的“teccl.txt”文件。

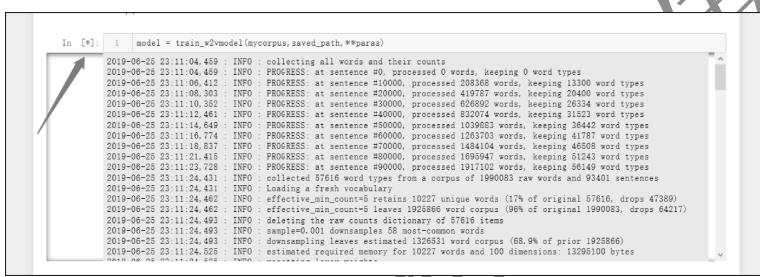


图29 词向量训练

4.5 常见问题

4.5.1 NLTK数据缺失

如果在运行过程中没有出现上图29所示输出，而是输出错误提示（图30），显示数据查找出错（Lookup Error），并且在错误提示末尾给出建议解决办法（图31），说明存在NLTK数据缺失问题。词向量训练默认调用NLTK的分词分句工具以及停用词表，需要NLTK数据。

图30 NLTK数据缺失错误提示

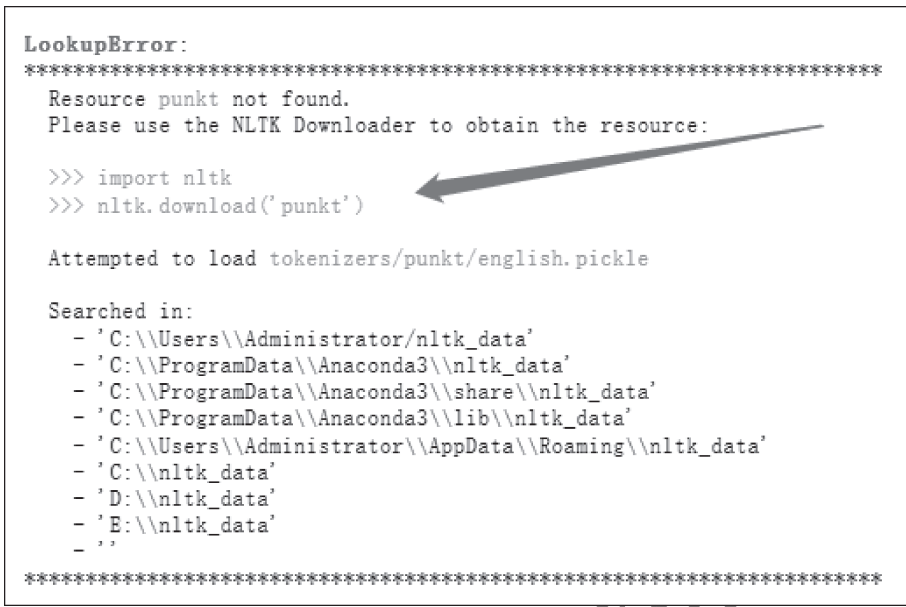


图31　NLTK数据缺失建议解决办法

解决问题的办法有两种。

（1）下载NLTK数据

根据上图31提示，打开命令行窗口，输入“python”，调出Python程序，然后输入“import nltk”，回车；接着输入“nltk.download（“punkt”）”，再回车。稍等片刻，下载完成后，重新运行即可。

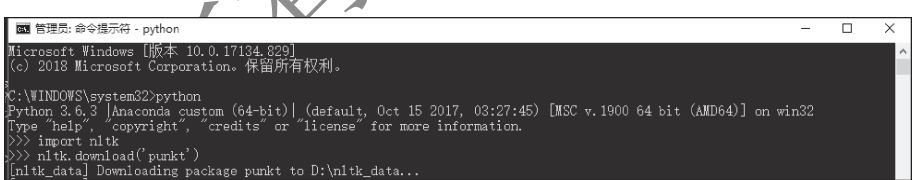


图32　NLTK数据个别下载

前面所示只是单个数据的下载办法。因为NLTK的数据比较多，NLTK提供了一个图形界面工具，可以下载全部数据。启动该工具的方法如下：打开命令行窗口，输入“python”，调出Python程序，然后输入“import nltk”，回车；接着输入“nltk.download（）”，再回车。稍等片刻，系统弹出NLTK数据下载工具（如下图33所示）。我们可以部分或者全部选取其中的数据，然后点击左下角的“Download”按钮进行下载。不过，由于NLTK数据在国外服务器上，下载速度比较缓慢，而且不稳定，如果全选，可能耗时较长，建议根据使用需要，分次下载。

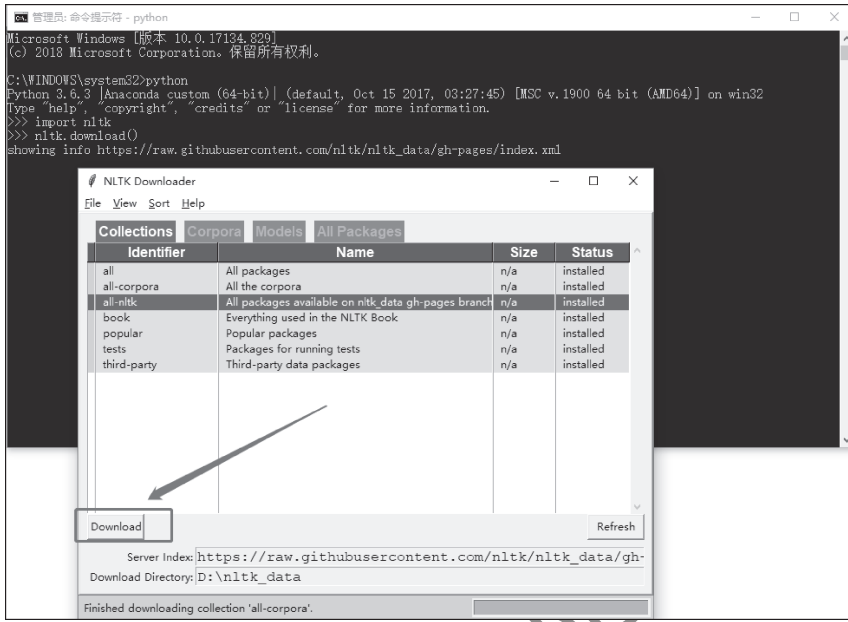


图33　NLTK数据下载图形界面

（2）复制NLTK数据到指定目录

由于NLTK数据下载很不稳定，本文提供了词向量训练可能会用到的部分NLTK数据，放在“nltk_data.rar”文件中。如果按照前面的方法没能下载成功nltk数据，可以将“nltk_data.rar”文件解压，然后放在C、D、E中任一盘的根目录下面。图34是以E盘为例的nltk_data路径。有些解压缩软件在解压缩后，会产生两层nltk_data目录，需要删除一层。数据复制到指定目录后，重新运行笔记本代码，问题便可得到解决。

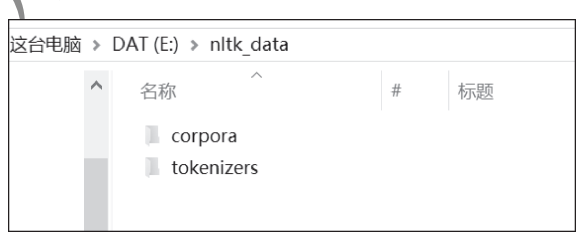


图34　NLTK数据放置目录

4.5.2　训练速度太慢

正常情况下，按当前笔记本主流配置，以默认设置使用TECCL语料库训练词向量，耗时应该在5分钟以内。如果远远超过这个时间（在20分钟以上），则

有可能需要安装C编译器。若C编译器缺失，运行训练时会出现如图35中提示："Install a C compiler and reinstall genism for fast training"。

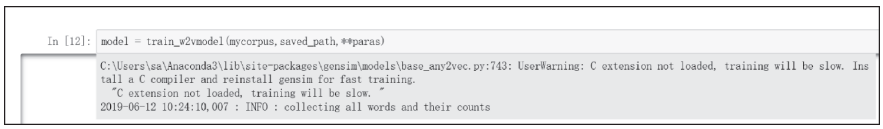


图35　C编译器缺失提示

解决该问题的办法如下：打开命令行窗口，依次输入下列命令。依次输入的意思是，等当前命令运行顺利完成后输入下一个命令，若因为网络原因导致失败，还需要重新输入当前命令，直到运行成功后再输入下一个命令。若在安装过程中提示"是否继续"（"proceed?（[y]/n）"），请输入"y"确认。

（1）conda install mingw libpython

（2）pip uninstall gensim

（3）conda install genism

（4）pip install scipy

5. 词向量探索

经过大规模语料训练的词向量可以用于词语相似度查询、计算和可视化分析。本部分以Gensim工具包提供的词向量操作接口为基础进行词向量探索分析操作演示。为方便操练，本文提供基于英语国家语料库（British National Corpus，简称BNC）[10]训练的词向量模型进行演示练习。

5.1　启动程序

启动程序步骤与词向量训练类似。

（1）打开本文提供下载的"python词向量学习3"文件夹。如果文件夹内只有"gensim_word2vec_exploring.ipynb"文件，请按引言部分说明的网址下载"bnc_lower.bin"文件，并将它放入文件夹内，如图36所示。

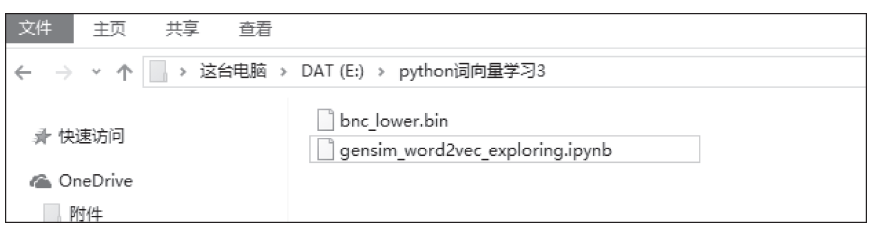


图36　"python词向量学习3"文件夹

（2）在文件夹路径上输入“cmd”调出命令行，并在弹出的命令行终端键入“jupyter notebook”启动Jupyter Notebook。系统在默认浏览器中打开Jupyter Notebook的Home主页，在文件列表中可以看到“gensim_word2vec_exploring.ipynb”。双击打开该文件，如图37所示。

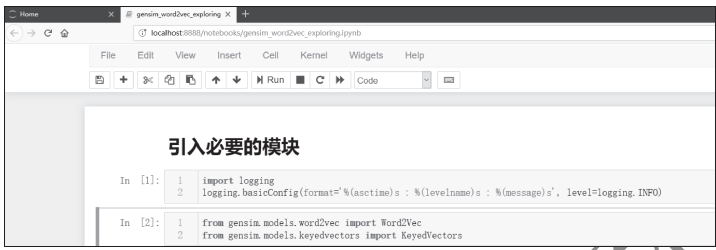


图37 “genism_word2vec_exploring.ipynb”文件显示

5.2 引入模块

图37所示两行单元格代码的主要作用是为了引入必要的模块，以便后面调用。我们只需单击第1行单元格代码，接着点击工具栏中的“Run”按钮运行；然后，如法炮制，运行第2行单元格，则完成了引入模块操作。

5.3 加载模型

加载模型部分有两行单元格代码，如下图38所示。第一行单元格代码是模型加载函数，只需按照单元格代码运行方法，直接运行其中代码即可。第二行单元格里面有执行模型加载代码，需要根据实际情况填写模型所在路径及文件名，图中箭头所示是填写好的BNC词向量模型地址。这里的地址写法与前面词向量训练中的词向量保存地址要求一样，可以填写相对路径与绝对路径，但要注意路径分割符号是反斜杠。填写好后，运行该行代码。如果加载顺利，出现如图38所示输出信息，则模型已经存放到变量model中，可以在后续代码中调用。

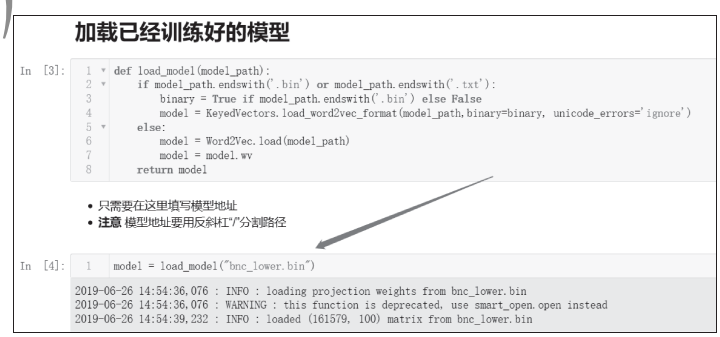


图38 加载词向量

5.4 词向量查询

词向量查询包括相似词检索、相似度计算与类比推理。

5.4.1 相似词检索

相似词检索函数是“model.most_similar”。图39显示了如何查询“apple”在BNC模型中的相似词。英文引号里是要查询的词，可以修改为其他词。“topn=”后面的数字可设置显示最相似词的数量。输出结果是一个列表，每行引号里的是相似词，而后面跟随的是余弦相似度数值。

```
In  [5]:  1  model.most_similar('apple',topn=10)

          2019-06-26 14:54:39,248 : INFO : precomputing L2-norms of w

Out[5]:  [('apricot', 0.6790697574615479),
          ('ice-cream', 0.6653609275817871),
          ('icecream', 0.6515501141548157),
          ('pear', 0.6417576670646667),
          ('onion', 0.6204477548599243),
```

图39 相似词检索

需要注意的是，只有模型中包含的单词才能查询得到结果。例如，所演示BNC模型已经将所有单词都小写了，若输入包含大写字母的单词，就会出错（如图40所示）。另外，词向量模型一般也只有单词，无法查询词组。同理可知，后面相似度计算、类比推理等相关操作也都只能输入模型中已有单词。

```
  1  model.most_similar('Apple',topn=10)

KeyError                                  Traceback (most recent call last)
<ipython-input-16-d2d641f58285> in <module>()
----> 1 model.most_similar('Apple',topn=10)

c:\programdata\anaconda3\lib\site-packages\gensim\models\keyedvectors.py in most_s
indexer)
    550                 mean.append(weight * word)
    551             else:
--> 552                 mean.append(weight * self.word_vec(word, use_norm=True))
    553                 if word in self.vocab:
    554                     all_words.add(self.vocab[word].index)

c:\programdata\anaconda3\lib\site-packages\gensim\models\keyedvectors.py in word_v
    465             return result
    466         else:
--> 467             raise KeyError("word '%s' not in vocabulary" % word)
    468
    469     def get_vector(self, word):

KeyError: "word 'Apple' not in vocabulary"
```

图40 查询非模型内单词出错信息

5.4.2 相似度计算

两个单词的相似度使用“model.similarity”函数进行计算（如图41所示）。若

要计算其他单词之间的相似度，只需将英文引号之内的单词修改为目标单词即可。

```
In [7]:  1  model.similarity('apple','banana')
Out[7]: 0.5903186
```

图41　相似度计算

5.4.3　类比推理

词向量最为人们所称道的就是它在类比推理（analogical reasoning）上表现出众，其中引用最多的例子是，词向量模型能够计算“king–man+woman≈queen”（Mikolov *et al.* 2013a，2013c）。换言之，针对“男人之于国王相当于女人之于什么”这样的问题，词向量模型能回答出“女王”。

词向量类比推理使用的函数与相似词检索一样，都是“model.most_similar”，但它输入参数的格式有所不同。“positive=”后面输入的是需要进行“加法”计算的单词，以字符串列表形式表示；“negative=”后面输入的则是需要进行“减法”计算的单词，也是以字符串列表形式表示。“king–man+woman≈queen”的类比推理示例见图42，表明词向量能捕捉到性别信息。同样，词向量也能够捕捉到城市与国家之间的关系（如首都），计算“beijing-china+japan≈tokyo”，见图43。

```
In [8]:  1  model.most_similar(positive=['king','woman'], negative=['man'],topn=10)
Out[8]: [('queen', 0.8614708185195923),
         ('prince', 0.7710027694702148),
         ('emperor', 0.7509008646011353),
         ('empress', 0.7503201961517334),
         ('pope', 0.749457836151123),
         ('duke', 0.7162444591522217),
         ('princess', 0.7060992121696472),
         ('throne', 0.7047961354255676),
         ('countess', 0.6967053413391113),
         ('lij', 0.6810743808746338)]
```

图42　类比推理示例1

```
In [9]:  1  model.most_similar(positive=['beijing','japan'], negative=['china'],topn=10)
Out[9]: [('tokyo', 0.7525430917739868),
         ('washington', 0.6978838443756104),
         ('stockholm', 0.6947522163391113),
         ('budapest', 0.6755738258361816),
         ('frankfurt', 0.6744974851608276),
         ('ottawa', 0.6646943688392639),
         ('moscow', 0.6634535789489746),
         ('tehran', 0.6627970933914185),
         ('seoul', 0.6627578139305115),
         ('madrid', 0.6606320142745972)]
```

图43　类比推理示例2

5.5　可视化呈现

可视化呈现部分中的前三个单元格代码都是为可视化作准备，包括模块导入、

子功能函数等，只需要依次运行即可。运行完后，在下图44所示的单元格中输入需要可视化的单词。每个单词以空格分开，放在英文引号里面。

图44　输入要可视化的单词

按照格式填写好需要可视化的单词之后，运行该单元格以及后续单元格，就可以看到词向量可视化平面图。

5.6　其他工具简介

由于词向量具有一次训练，可以反复使用、到处应用的特点，网上已有不少现成的词向量模型及其查询与可视化工具。对有些研究，我们并不需要编写甚至运行任何代码，可以使用现成模型与工具进行各种词向量操作。例如，Sketch Engine的词向量相似词在线查询工具包含多个语料库词向量模型可供开放检索。

6. 结语

本文以Gensim包为基础，对Python词向量训练与应用技术进行了简要解析。由于笔者已经对词向量源代码已经作了封装和简化，显著降低了词向量应用的技术门槛。读者不必关心词向量训练、查询与可视化的具体实现细节，可以专注于自身的研究问题，从而加速词向量技术应用。另一方面，对Python基础较好、希望操控更多实现细节的研究人员而言，源代码中包含了许多基础函数可供直接调用或者稍加修改使用，有助于大幅减轻后期编码工作量。

最近，以深度学习为引擎驱动，自然语言处理研究迎来了新的高潮，也给语言研究带来了重要机遇。近年来，词向量的发展可谓日新月异，从谷歌的Word2vec开始，又涌现出Glove、FastText、ELMo、BERT以及XLNet等不同词向量训练模型[12]，语言表示性能不断攀升，让人目不暇接。显然，本文只是提供了词向量的基础性介绍，仅作抛砖引玉之用，要充分利用大数据及人工智能技术助力语言研究，尚需学界同仁共同努力。

注　释

1. 详见网址：https://code.google.com/p/word2vec/。
2. 详见网址：https://radimrehurek.com/gensim/models/word2vec.html。

3. 北外语料库语言学网站的下载地址为http://corpus.bfsu.edu.cn/tools。GitHub下载地址为：https://github.com/aarondeng/python_word2vec_tutorials，打开页面后，点击“Clone or download”，选择“Download ZIP”则可打包下载全部源代码；数据文件按照说明下载。
4. 详见http://www.jetbrains.com/pycharm/。
5. 详见文档https://docs.python.org/3/library/idle.html。
6. 详见https://www.spyder-ide.org/。
7. 详见https://jupyter.org/。
8. 详见网址https://radimrehurek.com/gensim/。
9. 下载地址为http://corpus.bfsu.edu.cn/content/teccl-corpus。
10. BNC免费下载地址http://ota.ox.ac.uk/desc/2554。
11. 网络地址：https://embeddings.sketchengine.co.uk/。

此处仅列举了部分重要词向量模型。它们都开放源代码，可供复现结果或直接应用。

参考文献

Caliskan, A., J. Bryson & A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases [J]. *Science* 356(6334): 183-186.

Hamilton, W., J. Leskovec & D. Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change [A]. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)* [C]. 1849-1501.

Harris, Z. 1954. Distributional Structure [J]. *Word* 10(2-3): 146-162.

Liu, D. & L. Lei. 2018. The appeal to political sentiment: An analysis of Donald Trump’s and Hillary Clinton’s speech themes and discourse strategies in the 2016 US presidential election [J]. *Discourse, Context & Media* 25: 143-152.

Mikolov, T., K. Chen, G. Corrado & J. Dean. 2013a. Efficient estimation of word representations in vector space [R]. Paper presented at *The Workshop at International Conference on Learning Representations (ICLR)*. Scottsdale, Arizona.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado & J. Dean. 2013b. Distributed representations of words and phrases and their compositionality [R]. Paper presented at *The International Conference on Neural Information Processing Systems (NIPS)*. Lake Tahoe, Nevada.

Mikolov, T., W. Yih & G. Zweig. 2013c. Linguistic regularities in continuous space word representations [R]. Paper presented at *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. Atlanta, Georgia.

冯志伟，2019，词向量及其在自然语言处理中的应用 [J],《外语电化教学》(1)：3-11。

管新潮，2018,《语料库与Python应用》[M]。上海：上海交通大学出版社。

许家金，2016，“中国学生万篇英语作文语料库”介绍 [J],《语料库语言学》(2)：108-112。

通信地址： 100089 北京市北京外国语大学中国外语与教育研究中心/赣南师范大学

《语料库口译研究的出路》述评

四川大学 刘雨凤

Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.). 2018. *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer Nature Singapore Pte Ltd. xvi+215pp.

1. 引言

基于语料库的口译研究一直落后于相应的笔译研究。尽管Shlesinger（1998）和Hu（2016）等学者早已有过相关探讨，但在整个语料库翻译学中，笔译仍处于优势地位，研究者多将目光集中于研究笔译中的翻译共性（Baker 1996；Laviosa 1998）、译者风格（Baker 2000；Olohan 2004）、翻译规范（孙艺风 2003；李德超、邓静 2004）等。然而，口译和笔译应该是语料库翻译学下两个享有平等地位的分支（Shlesinger 1998）。本文介绍的《语料库口译研究的出路》一书（以下简称《出路》）共十一章。本文首先将全书各章进行分类和内容概述，然后进行简要评析。

2. 内容概述

全书可分为六个部分：回顾与展望（第一章）、语料库的设计与检索（第二章）、基于语料库的认知负荷探索（第三章、第八章）、口译语言的特点（第四章、第五章、第六章、第七章）、特殊语境下的口译策略（第九章、第十章）、机器翻译技术（第十一章）。

第一部分，即第一章，是对语料库口译研究的过去、现状和未来的概括。本章总结了语料库口译研究的历程和所涉及的语种，提出由于口译是特定模式的翻译，其发生场合与笔译有所不同，语料库的建设面临巨大挑战。但同时，Web 2.0时代的到来，也为大型口译语料库的建设提供了更多的支持，使语料库口译研究脱离司徒罗斌（Robin Setton)（2011）所说的“小作坊”（cottage industry）（p. 34）的尴尬境地。作者在结尾处呼吁学界关注口译语料库和笔译语料库的不同，强调开辟对齐和标注的新路径，以及语料库对译员培训的重要性（p. 8）。

第二部分，即第二章，进入操作层面，即如何建设口译语料库以及如何有

效检索口译语料。本章强调，研究者在建库前应该明确研究目标，以便针对性建库。建库过程中的一个难点在于转写，口语的语音特征以及译者、听众、演讲者之间的互动较难完整保留。此外，本章探讨了多个转写软件的优缺点，包括EXMARaLDA、Handbrake、VLC和Reaper，以便研究者择优选用。同时，本章指出，元信息（metadata）的记录不可或缺，必须做到真实、准确。对于口译语料的对齐，作者还介绍了相应的软件：文本之间的对齐可使用Intertext Editor软件；而文本与音频之间的对齐可使用SpeechIndexer或Transana软件；文本和视频之间的对齐可采用Subtitle Workshop软件。最后，本章指出不同的语料库软件各有优缺点，因此每个语料库都应有与之对应的检索方法，如AntConc、TextStat、Coma、NoSkE等，但作者并未详述。

第三部分，即第三章和第八章，主要对Gile（1997）的认知负荷模型进行验证，即探索译员的“黑匣子”。第三章对比了作为译出语的荷兰语与母语者的荷兰语，主要考察有声停顿的频率，如“嗯”“呃”等，探索认知负荷与有声停顿频率是否呈负相关关系。由于语料库收集的真实语料十分庞大，得出的结果也更加令人信服，是佐证前人研究成果的得力工具。然而，第三章的作者也意识到，尽管通过语料库可以得出大量的数据，但产生该数据的原因到底是译员的认知负荷过大还是译员有意采取的翻译策略很难判定。第八章聚焦隐喻翻译中的认知负荷，提出隐喻翻译的四种策略——直译、替换、意译、省略。通过研究表明，翻译创造力越低的隐喻所需的认知负荷越少。此外，本章作者还强调，语料库的口译研究旨在探究已经发生的翻译事实，而非规定应该如何翻译。

第四部分是本书的核心内容，涉及口译语言的特点，包括第四章、第五章、第六章和第七章。第四章关注英语和汉语在定语修饰语顺序上的不同，旨在探究语种对口译语特征的影响。本章作者将目光从印欧语转移到非印欧语。同时，本章采用了三角验证法（triangulation），借用了基于可读性公式（Flesch 1948）的软件Readability Analyzer，测量不同语言体系下的口译语言特征。第五章指出，口译员往往依赖程式语言，因为程式语能够减轻译者的认知负荷。程式语通常取决于其使用频率和特定的形式、语义、语用功能或惯用性。并且，以英语为母语的译者平均掌握的程式语比单字词汇多。作者提到，目前利用语料库建成的写作和笔译程式语已经成型，但演讲和口译的相关语料库所见不多。因此，对译员的培训还应着重培养他们对程式语的敏感度，而利用如WordSmith的词丛功能，译员能够更加高效地学习和掌握常用的程式语，提高口译技能。第六章以小见大，从英语导句词that出发，证明口译语言的存在——即口译语言有别于笔译和自然语言。该章佐证了Olohan & Baker（2000）对于英语翻译中that的探索。最后，本章作者得出结论，that在英语口译语和英语作为第二或第三语言中使用更多，在英语母语中使用更少，并且显化和简化特征在英语口译语中较为明显。第七章考虑了性别

在英语—西班牙语—意大利语口译中的影响。本章主要采用的是定量分析，用t检验和协方差分析对男女译员在口译速度、语言特征等方面进行比较。本章的结果略显复杂，且未通过定性分析进行解释。

第五部分，包括第九章和第十章，探讨了特殊语境下的口译策略。第九章关注的是欧盟领导人辩论场合中的口译，第十章关注的是足球新闻发布会上的口译。尽管两者的口译场合截然不同，但其研究问题类似——Q & A环节中口译员的特殊策略，且两者都是电视口译（screen interpreting）。借助大规模语料库，研究者们发现口译员在提问环节中更多地充当调和者的角色，有意弱化尖锐的问题或答案。

最后一部分，即第十一章借用多个高级评估模型（BLEU模型、RIBES模型和Translation Edit Rate模型）比较职业译员、机器口译、字幕翻译（TED演讲）和自然语言之间的质量和特点。结果发现机器口译仍有可取之处，呼吁研究者们对机器口译的关注，尤其需要关注自动语音识别技术在机器口译中的作用。

3. 简评

总体而言，学界关于口译语料库的研究不如笔译语料库的研究蓬勃。尽管学者们对语料库在口译研究中的作用寄予厚望，但实践与愿望仍然差距不小。鉴于此，《出路》一书也许可以给相关学者从事语料库口译研究提供借鉴和希望。

第一，建立口译语料库可借鉴相关的软件和工具，以达到减少建库周期的目的。市场上已经出现了多款转写软件和对齐软件，减少了人工投入力度，可将精力投入到对语料的分析上。

第二，目前已有的语料库软件各有利弊，研究者在研究前可先熟悉各款语料库软件，如常用的AntConc、ParaConc等，也可跨学科合作，为特定的研究目的设计自己的语料库软件，如国内梁茂成等（2010）设计了独特的建库软件。

第三，学者们应该认识到，语料库只是得出数据的便捷工具，对数据的解释还应该借助其他的理论模型和框架，如Gile（1997）的认知负荷模型、Flesch（1948）可读性公式、BLEU/RIBES/Translation Edit Rate评估模型等；或者统计学方法，如t检验和协方差分析等。

第四，对比仍然是得出有力结论的方法。如将自建语料库与BNC等已建成语料库进行比较，将口译语料库与笔译语料库、自然语言语料库、机器翻译语料进行比较。

第五，语料库虽然庞大，数据虽然集中，但微小的切入点也是得出合理结论的必要前提，如从that出发，分析英语作为译语和自然语言的结构特点。国内不少学者也常采用类似的切入点，如研究莎剧中的good一词（胡开宝、马秀奇2015）等。但关于口译的相关研究仍处于初级阶段。

第六，机器翻译虽然取得了一定发展，但仍有诸多局限，尤其是即时特点强的口译。研究者们不能一味否定机器翻译，相反，可以利用机器翻译的相关成果，将语音自动识别技术运用到职业译员的口译过程当中，也可围绕机器口译作相关的研究。

然而，本书仍存在以下不足：

第一，结构上，本书稍显冗杂和零散，没有设立专题讨论专门的口译研究，如理论探索、技术要求、口译策略、译员风格、译员培训等。

第二，内容上，本书仅仅收录了两篇涉及亚洲语言的文章——汉语和日语，其余章节均围绕欧洲各国的语言。实际上，欧洲各国语言和汉语、日语存在着巨大差异，基于前者得出的结论很难直接运用于后者。

第三，研究方法上，本书收录的文章仍然缺乏定量和定性研究的结合。三角验证法将会是未来翻译学研究的一大趋势。定量和定性的结合实际囊括了实证研究（包括语料库翻译学）的两个重要方面：1）得出有效数据；2）作出合理解释。

第四，内容深度上，容易浅尝辄止。尽管在第一章末，作者呼吁学界关注语料库在译员培养中的作用，但本书并未收录专门利用语料库培训优秀译员的文章。而第二章对于技术问题的探讨，也主要集中在理论层面，对实际的操作借鉴意义不大。

总体而言，本书能为对采用语料库研究专门口译问题感兴趣的学者提供支持和帮助，是语料库口译研究的一本重要入门书籍。

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: Studies in Language Engineering, in Honour of Juan C. Sage* [C]. Amsterdam: John Benjamins. 175-186.

Baker, M. 2000. Towards a methodology for investigating the style of a literary translator [J]. *Target* 12(2): 241-266.

Flesch, R. 1948. A new readability yardstick [J]. *Journal of Applied Psychology* 32(3): 221-233.

Gile, D. 1997. Conference interpreting as a cognitive management problem [A]. In J. H. Danks, G. M. Shreve, S. B. Fountain & M. Mcbeath (eds.). *Cognitive Processes in Translation and Interpreting* [C]. Thousand Oaks, CA: SAGE Publications. 196-214.

Hu, K. 2016. *Introducing Corpus-based Translation Studies* [M]. Shanghai: Shanghai Jiao Tong University Press.

Laviosa, S. 1998. The corpus-based approach: A new paradigm in translation studies [J]. *Meta* 43(4): 474-479.

Olohan, M. 2004. *Introducing Corpora in Translation Studies* [M]. London: Routledge.

Olohan, M. & M. Baker. 2000. Reporting *that* in translated English: Evidence for subconscious

processes of explicitation? [J]. *Across Languages and Cultures* 1(2): 141-158.

Setton, R. 2011. Corpus-based interpreting studies (CIS): Overview and prospects [A]. In A. Kruger, K. Wallmach & J. Munday (eds.). *Corpus-based Translation Studies: Research and Applications* [C]. London: Continuum. 33-75.

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

胡开宝、马秀奇，2015，莎士比亚戏剧汉译本中“good”评价意义的再现研究 [J],《当代外语研究》(3)：7-13。

李德超、邓　静，2004，传统翻译观念的逾越：彻斯特曼的翻译规范论 [J],《外国语》(4)：68-75。

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。

孙艺风，2003，翻译规范与主体意识 [J],《中国翻译》(3)：3-9。

通信地址：610064 四川省成都市四川大学外国语学院

English abstracts

A multi-dimensional inquiry into the syntactic complexity of Chinese English learners' oral English

Based on a review of measures investigated in English syntactic complexity studies, this article establishes, via factor analysis, a new syntactic complexity model of four dimensions, i.e. subordinate ratio, unit length, coordinate phrases and nominal structures. The study also examines the effects of grade level and genre on college students' syntactic complexity in oral English in terms of four dimensions. The ANOVA results show that all indices except C/T, the one representing subordinate ratio which maintains stable across grade levels, are positively correlated with grade level and sensitive to genre. Students tend to produce more complex sentences in argumentation than on narration at all four dimensions. Tentative explanations are provided for findings.

Engagement resources in inconsistent tense uses in translational English news discourse from Chinese

This study examines engagement resources in inconsistent tense uses in translational English news from Chinese. Based on Martin & White's engagement system of the Appraisal Theory, we classify and investigate reporters' and reported speaker stance. It is found that of 17 types of inconsistent tense collocations, translational English mainly uses past tense reporting verbs and intentional absolute tenses to show [Dialogistic expansion] [Dialogic contraction] stance, while native English mainly relies on present tense, present perfect reporting verbs collocated with intentional relative tenses to show [Dialogistic expansion] [Dialogic expansion] stance. Translational English more frequently holds neutral and negative stance, refusing to take responsibility for represented speakers' negative or positive stance. On the other hand, reporters in native English tend to take neutral reporting stance, and do not refuse other voice interference, acknowledging that represented speakers' negative stance is one of possible voices. Reporters in translational English are more adept in employing the linguistic strategy of inconsistent tense collocations, taking dialogic contraction stance to challenge and restrict other voices and stance, thereby contracting dialogic space and introducing reporters' attitudes into dialogic space. Reporters in translational English often employ external voices to openly demonstrate supportive attitudes and stance to represented speakers' strategy of showing clear stance and contracting dialogic space, while reporters in native English often quote external authoritative voices to

counter previously introduced affirmative propositions and show clear negative stance at the same time.

A corpus-based analysis of the lexical styles of Mao Dun, Ba Jin and Lao She

This paper focuses on three representative writers, namely Mao Dun, Ba Jin and Lao She in the second decade (1928—1937) of Chinese modern literature. Samples of 200,000 Chinese characters from the three writers respectively were collected for lexical profile analysis. Their differences in terms of mean word length, type-token ratio, hapax, demonstratives and key content words. Underlying reasons for the lexical differences were discussed in light of their individual literary characteristics.

The building of Lin Shu's translation corpus and its research

Lin Shu's translations of novels are a huge cultural project, which has not been systematically discussed. Most of the studies have based on introspective analysis, relying solely on one book or even some chapters or phrases. As a result, they will deviate from the contextualized translation practice and the social and cultural situation at that time. The paper introduces the issues concerned in the process of building Lin Shu's translation corpus, trying to explore Lin Shu's collaborators and their translations, bibliographic information of the source text in English, word segmentation and annotation of translation strategy, etc. The Lin Shu's translation corpus will provide a new method for interpreting Lin Shu in the field of translation studies. It will also provide new research results for the field of translation studies and offer objective and scientific data for fuller understanding and evaluation of Lin Shu.

Training and exploring word embeddings in Python

Word embedding models perform excellently in semantic representation. The age of big data sees huge potential of word embeddings' applications to language studies. This article attempts to demonstrate how to train and explore word embeddings in Python with Gensim for language researchers who are not familiar with computer programming. The illustrated tutorial covers the set-up of a Python development environment, the installation of the Genism package, and the training, saving, loading, exploration and visualization of word embeddings. Along with

this tutorial, fully commented, executable and configurable Python source code files in IPython Notebook format are provided to smooth the learning curve, and thus lower the threshold for application.