

语言资源与智能

2018 年第 4 期 (总第 9 期)



语言资源高精尖创新中心 主编

2018 年 10 月

目 录

智库研究

- 国家出行，语言先行（李宇明教授）

中心动态

- 李宇明教授出席第十一届中国社会语言学国际学术研讨会
- 世界知识的中文表达
——“2018 海内外中国语言学者联谊会第九届学术论坛”综述
- 世界语言资源保护大会隆重开幕，中心积极参与
- 李宇明教授当选为中国辞书学会会长
- 中心两位项目负责人获 2018 年政府特殊津贴

项目进展

- “汉语中介语语料库建设创新工程”项目进展报告会顺利召开
- “新选中国名诗 1000 首：当代诗学名家经典选释系列”项目进展交流会顺利召开
- “语言识别理论及语言数量统计的方法论研究”项目进展交流会顺利召开
- “面向冬奥会的机器翻译资源建设与技术研究”项目进展检查会顺利召开
- “基于知识图谱的北京冬奥项目智能问答系统”项目进展交流会顺利召开

领域资讯

- 国际计算语言学协会亚太地区分会（AAACL）成立等资讯 3 则
- 语言资源与智能领域 ESI 高被引论文 4 篇、热点论文 3 篇

➤ 智库研究

国家出行，语言先行

李宇明教授

“一带一路”需要语言铺路。语言之路铺设如何，直接关系到政策能否沟通、道路能否联通、贸易能否畅通、货币能否流通、民心能否相通？关系到“一带一路”能有多宽广，能行多快，能行多远？



“一带一路”倡议以来的五年，也是努力铺设语言道路的五年。例如：汉语在国际上的传播速度持续增快，现在 146 个国家和地区都有汉语教学活动，近 70 个国家把汉语列入国民教育体系，孔子学院和孔子课堂的布局自觉向“一带一路”国家倾斜。在外语语种的教育方面，中小学在已经开设英语、俄语、日语的基础上，又开设了法语、西班牙语、德语；大学外语专业的语种不断增加，现在已有近 100 种，目标是凡有国语身份的语言，都将进入大学的外语教育。“一带一路”语言服务的会议不断召开，理念在更新，行动在跟进，语言人才市场在逐渐形成，产出了一批理论成果和可观的社会效益。中华概念、术语外译工作持续进行，这是向世界“讲好中国故事”的基础准备。信息自动检索、语言自动翻译、机器自动写作等现代语言技术又有新进步，据调查，我国有 60 余款线上机器翻译系统，其中有 8 款系统能够支持 90 种语言的自动

翻译，有三款系统可以支持百种以上语言的自动翻译。语言智能推进“一带一路”进程取得新成效。

在铺设语言道路的进程中，大家更加明白：语言是国家的先遣队，国家要去何方，语言就应先到何方。中国正在由“本土型国家”向“国际型国家”转变，国家的语言能力也应由“本土型国家”向“国际型国家”发展。

“国际型国家”的语言能力，就是能够满足国家处理好海内外事务的语言能力。当代跨国跨地区的活跃语言，包括英语、汉语在内约有 20 来种。这 20 来种语言基本上是当代人类知识的主要负载者和交流者，因此，国家必须“熟练掌握”这些语言。与个人熟练掌握一种语言不同，国家对一种语言的熟练掌握起码应达到如下三点：其一，能够用它来进行口语和书面语的交际；其二，能够掌握它的语言知识和语言技术，能够进行语言翻译和知识交换；其三，有较为完整的语言教育体系。

除了这 20 来种语言之外，世界上还有一些较为重要的语言或方言，它们是某国的国语或官方语言，或者是一些国家内部的大语言或大方言。比如：老挝语，缅甸语，柬埔寨的高棉语，菲律宾的他加禄语，印度尼西亚的爪哇语，尼泊尔语，不丹的宗卡语，斯里兰卡的僧伽罗语，南亚的泰米尔语、旁遮普语，印度的泰卢固语、马拉地语、古吉拉特语等，土库曼语，塔吉克语，阿塞拜疆语，爱沙尼亚语，拉脱维亚语，立陶宛语，中东地区的库尔德语，西班牙的加泰罗尼亚语，法国的科西嘉语，中国的藏语、维吾尔语、彝语、傣语和汉语的粤方言、闽南方言、客家方言等。这些语言或方言大约有 200 种，使用这些语言或方言的民族或族群都比较大，都有一定的影响力。要使“一带一路”倡议得到广泛响

应，要建立人类命运共同体，做到民心相通，也应当掌握好这些语言，具有用这些语言进行交流的能力。

要满足“一带一路”倡议的需求，具有“国际型国家”的语言能力，较为理想的状态就是具有 200 来种语言的能力，亦即：能够从 20 来种跨国跨地区的活跃语言中获取当代知识，能够用 200 来种语言进行沟通，向外讲好中国故事。与此相比，我国的语言能力还有很大的提升空间。

达到“国际型国家”的语言能力，首在提高语言意识，真正认识到语言是“一带一路”发展、是国家走向世界的“先行官”。语言文字是人类最为重要的沟通工具，但也不能仅仅只看到语言的工具性；语言也是打开“心结”、消解误会的心灵锁钥，也是文化的阐释者和建构者。建立人类命运共同体，必须全方位发挥语言的作用。

汉语在国际语言生活中所发挥的作用，是衡量“国际型国家”语言能力的一个重要指标。促进汉语国际传播，发展汉语的“第二语言”人口，汉语国际教育的重要途径。在全力推进汉语国际教育之时，还要特别关注学过汉语者的汉语使用，关注国际组织的汉语使用，关注汉语文本的国际影响力，特别是汉语文本在国际新闻舆论、国际学术研究、国际商贸活动中的影响力。这种影响力，基本决定了汉语在国际语言生活的应用价值和学习价值。

国家的外语能力是国家语言能力的重要组成部分。在关注大学外语教育的同时，要特别关注中小学的外语教育。第二语言学习的黄金时期在小学和中学。发展中小学外语教育的措施之一，是根据我国各地外语使用的特点来部署外语教育语种，比如东北地区便于学习俄语、日语，西南地区便于学习东盟的语言，西北地区便于学习南亚、中亚的语言等。此外，我国的跨境语言有近 30 种，如蒙古语、朝鲜语、赫哲语、鄂伦

春语、鄂温克语、维吾尔语、哈萨克语、乌兹别克语、柯尔克孜语、锡伯语、图瓦语、藏语、苗语、瑶语、壮语、傣语、景颇语、怒语、哈尼语、独龙语、佤语等。这些跨境语言在对外交往和文化传播中，往往可以发挥外语的作用。研究跨境语言，掌握好跨境语言，发挥好跨境语言的作用，不仅可以迅速增长国家的语言能力，而且可以沿边境建立多功能的语言带，增强跨境的语言文化沟通，也增加边境的语言文化安全。

现代语言技术仍然是提升国家语言能力的关键，提升机器自动翻译的水平乃当务之急。现在的多语自动翻译系统，多是以英语为“轴心语言”的。要争取建立以汉语为轴心语言的多语种自动翻译系统，使汉语能够逐渐成为机器翻译的“路径依赖语言”。同时，这一自动翻译系统要足够强大，能以之向世界最重要的 20 来种语言获取信息，向世界 200 来种语言输出信息。研发这一翻译系统还有更为重大的文化意义，因为现在世界的学术语言已经形成了英语一家独大的局面，近十年来，在 SCI 数据库中英语文献占了将近 97%，在 SSCI 数据库中英语文献占了将近 95%。如此局面不改，国人必须通过翻译来获取科学文献，汉语将很难成为世界的学术语言。开发这一系统，可以帮助解决国人获取世界科学文献的问题，帮助解决“世界知识的中文表达”问题。

语言学是研究语言及其相关问题的科学，是培养各种语言人才、发展国家语言能力的学术基础。中国是世界语言研究的三大源头之一，古代的文字、音韵、训诂之学被称为“小学”，其研究主要是为读经服务的，是经学的翼羽。1898 年《马氏文通》的出版，标志着中国的语言研究进入到现代时期，语言成为专门的学科研究对象。但是直到今天，我国在人才培养的学科体系中，语言学却没有独立，仍是“小学”的衣钵，文学的附庸。中国的语言研究与语言学教育，以语种为藩篱，分割在中

国语言文学、外国语言文学等学科中，外国语言文学、少数民族语言文学下面仍然被语种分割开。而国际上，多数大学都有语言学系，或独立设置，或在哲学、法学、医学、心理学、脑科学、计算机科学等学院里。正如数学是自然科学的工具学科一样，语言学也具有人文社会科学的工具学科性质。在当代的科学发展中，语言学已经超出了人文社会科学的范畴，进入到科学技术领域，而且与许多学科正在快速地交叉互融，形成一批批交叉学科，因此，语言学也是许多学科的学生必修课。

当前，我国正在开展的“世界一流大学和一流学科”建设，与语言相关的一流学科有“中国语言文学、外国语言文学、语言学、现代语言学”。这四个学科名目逻辑上显然不能自洽，反映了我国语言学学科设置亟需完善，也反映了用旧有的学科观念包容不下今日之语言学。完善语言学的学科设置，是发展语言学事业、培养语言人才的需要，更是为“一带一路”铺路、为国家走向世界做好“先行”的需要。

➤ 中心动态

李宇明教授出席第十一届中国社会语言学国际学术研讨会

2018年7月13日-15日，由中国社会语言学会主办、吉林大学公共外语教育学院暨高校外语教育研究中心承办的第十一届中国社会语言学国际学术研讨会在吉林大学召开。语言资源高精尖创新中心主任李宇明教授受邀出席会议并作主旨报告，吉林大学副校长吴振武教授出席开幕式并致辞。

本次大会主题是“全球化时代的中国社会语言学：机遇与挑战”，来自中国、中国香港、英国、美国、芬兰、荷兰、澳大利亚等近10个

国家和地区的 260 多名专家学者和研究生代表参加会议。大会共设 4 场主旨报告、4 场特邀报告、1 场主编论坛、1 场专家面对面和 14 场分组讨论。



在主旨报告环节，李宇明教授作了《中国的语言生活皮书及其有关问题》专题报告。他指出，中国的语言生活皮书已经形成“绿、白、蓝、黄”四大系列，是在中国语言规划

由“语言规范”“语言管理”发展到“语言生活治理”时期的重要表现。在皮书的发展过程中，近两年又相继出版了《北京语言生活状况报告》《广州语言生活状况报告》两部地方皮书，展示着“南北双雄”的城市语言生活景观。

李宇明教授指出，这些皮书共同的编纂追求是“有事有趣有思想”。所谓“有事”，就是要反映语言生活中的大事与热点，反映政府的语言政策与举措；所谓“有趣”，就是要尊重读者，



表述要尽量符合大众语言习惯，“说事”要重视用典型事例和真实数据；所谓“有思想”，就是不满足于把皮书只作为一个记事簿，而要在记事的同时，注意分析事件的形成原因与发展走向，揭示制约事件的因素与规律，努力淬炼出语言规划的新观念、新理论。

他强调，在“语言生活”基本理念指导下，在“有事有趣有思想”的编纂追求中，这些皮书各有侧重，相互配合，相得益彰。他指出，语言生活皮书坚持七大学术思想：一是关注语言生活，引导语言生活，构建和谐的语言生活；二是语言是资源，要珍惜它，爱护它，充分开发利用它，以期获取最大的语言红利；三是尊重各社区、各群体的语言权利，主张“地位平等、功能互补”的多语主义，使各种语言及其变体各得其所、各展其长、各安其位、相辅相成；四是努力提升个人的语言能力和国家处理海内外事务的语言能力；五是推进政府和学界的社会语言服务，关心国际、国家、社会各领域和家庭的语言规划，着力打造学界与社会的智力“旋转门”，探索用社会话语表述语言学研究成果；六是语言学发展的原动力，就是解决社会前进中遇到的语言问题，解决这些语言问题，需要国家、学界和社会的共同参与，需要多学科的合作共治，需要多种研究方法的综合共用，需要重视语言国情的调查与研究，特别是各种实态数据的收集与运用；七是信息化为语言生活提供了虚拟空间，为语言运用提供了语言技术和新媒体平台，为语言研究和语言规划提供了新手段，要全力促进语言信息化，过好虚实两个空间的语言生活，迎接人与机器人共事的语言智能时代的到来。

李教授在主旨报告中第一次对建国 69 年以来的中国语言规划作出系统梳理，对语言生活皮书作出全面总结，报告高屋建瓴，讲话深入浅出，充满睿智，引起与会人员的强烈共鸣。此外，李宇明教授还分别参加了主编论坛和专家面对面活动。



世界知识的中文表达

— “2018 海内外中国语言学者联谊会第九届学术论坛” 综述

2018 年 7 月 29 日，“2018 海内外中国语言学者联谊会——第九届学术论坛”在北京商务印书馆礼堂举行。本次论坛由中国社会科学院语言研究所、北京语言大学、中国语言学书院和商务印书馆联合主办，主



题为“世界知识的中文表达”。来自北美、欧洲、日本和中国（含香港、澳门地区）的五十余位语言学专家和中国语言学书院暑期研修班的部分学员参加此次论坛。

本次论坛主题的倡议者、北京语言大学语言资源高精尖创新中心主任李宇明教授致辞并做解题发言。他指出，基于对 2006-2015 年世界三大数据库的论文情况的分析，可以看到，随着语言全球化的进程，英语成为“世界通用语”，特别是成为科技（包括社会科学和人文科学）的主要“国际用语”。伴随着中国国力的提升，中文虽然也加快了国际传播的步伐，但是在国际科学技术领域却同英语之外的其他语言一样，被严重边缘化。这将对国人获取世界科技知识（亦即现代知识）形成严重障碍，



将对中华文化的发展强盛带来严重影响。拆除语言藩篱，解决世界知识的中文表达问题十分不易，须有耐心、有恒心、有信心。国家和知识界须有危机意识，除了提高全民的外语能力之外，要努力发展以中文为轴

心的多语种自动翻译技术，完善信息获取与推送的现代知识工程，让中文成为国人获取知识的路径依赖语言。更要通过政策支持建立“中文自信”的学术评价体系，提升中文文本的国际声望。这是更为根本的打破语言藩篱之策。

会上，北京大学哲学系陈波教授、上海外国语大学中东研究所朱威烈教授、伦敦大学伯贝克学院祝华教授、美国威斯康星大学张洪明教授分别作了《中文的学术地位与中国学术的国际影响力》《国际知识的中文供给》《中文研究在国际学术界的可信度与关注度》《“世界知识的中文表达”平议》主旨报告。在自由发言阶段，与会学者崔希亮、李行德、孙茂松、李长波、顾钢、林华、高雪松、史有为、周荐、周小兵、卢涛、洪波、徐杰、章宜华和陆俭明等先后发言，围绕着与主题相关的学术体系建设、全球视野、国际学术交往、翻译质量、术语表达规范、范式研究及话语权等展开了热烈讨论。中国社会科学院语言研究所所长刘丹青研究员做总结发言。

世界语言资源保护大会隆重开幕，中心积极参与

李宇明教授主讲《中国语言资源保护的理念与实践》



2018年9月19日上午，首届世界语言资源保护大会在长沙举行。湖南省委副书记、省长许达哲，教育部副部长、国家语委主任杜占元，联合国教科文

组织传播与信息助理总干事穆兹·查楚克，巴林公主努拉·宾特·哈利法，湖南省委常委、长沙市委书记胡衡华，副省长吴桂英等领导出席。

大会由联合国教科文组织、中国教育部、中国联合国教科文组织全国委员会、国家语言文字工作委员会、湖南省人民政府联合主办，湖南省语言文字工作委员会、湖南省教育厅、长沙市人民政府、北京语言大学、湖南大学共同承办，是联合国“2019国际本土语言年”的重要活动之一，也是中国语言资源保护工程重要的标志性成果之一。

大会以“语言多样性对于构建人类命运共同体的作用：语言资源保护、应用与推广”为主题，下设四个分论坛：语言文化多样性政策与措施、语言资源保护规范标准与人才建设、语言资源开发应用与推广、促进语言文化多样性的措施。40个参会国与会代表及联合国教科文组织嘉宾、部分国家驻华使领馆嘉宾、国内嘉宾等出席大会，并将共同见证大会成果《岳麓宣言》的诞生。

语言资源高精尖创新中心主任李宇明教授作为平行论坛中方总主持人应邀出席本次大会，并作了题为《中国语言资源保护的理念与实践》的报告。

李宇明教授指出，语言是一种社会资源，是近些年来才得到社会认可和重视的资源。中国在历史上也比较重视对语言资源的保护和开发利用，有许多经验值得总结和继承。但是，真正理性地认识到语言的资源意义，特别是把语言资源作为国家语言规划的重要理念，并在国家层面、在全国范围采取语言资源保护行动，还是十几年来做的事情。

他指出语言资源包括三类：1. 口头语言资源；2. 书面语言资源；3. 语言知识、语言技术、语言艺术等语言衍生资源。语言资源理念下的



“语言”，并不只是个“符号系统”，而更是个“知识系统”，语言保护不只是记录、保护语言符号，而是记录、保护人类的语言知识体系。就语言的文化资源视角来看，语言保护其实包含三个层次。其一是“语言保存”，

即通过书面方式和录音录像方式，将语言记录下来，建库保存下来。当前学者进行的多是语言保存层面的工作。其二是“语言卫护”，是通过各种措施来延长语言生命、维护语言活力。这类卫护工作，人类社会也在做不同的探索，积累了一些经验。其三是语言资源的开发利用，即对语言保存、语言卫护的成果进一步开发，获取语言保护的“红利”。

他强调社会正在向“智能时代”迈进，人工智能是智能时代的最主要的技术力量。无论从文化视角还是从人工智能的视角来看待语言资源，对语言资源的认识都需要加强，语言资源的保护需要人类社会的合作，包括不同地区、不同国家、不同国际组织的合作，不同社会部门和不同学科的合作。为了保证这种合作的开展，需要制定一系列国际标准，包括技术标准、工作标准。

刘晓海副研究员主讲《语保工程多媒体技术规范与语料采录加工方法》

首届世界语言资源保护大会上，来自 40 个国家和地区的代表在紧锣密鼓的主旨发言、平行论坛中，互相交流语言资源保护经验做法、多语社会的构建及语言政策等。在平行论坛环节中，北京语言大学语言资

源高精尖创新中心副主任刘晓海副研究员就《中国语言资源保护工程多媒体技术规范与语料采录加工方法》作了有关报告。

报告指出，中国语言资源保护工程（简称“语保工程”）是中国国家财政支持的重大语言文化工程。工程按照国家统一规划、专家和地方共同组织实施、鼓励社会参与的方式进行。语保工程第一阶段任务的重点是充分利用现代



化技术手段，收集记录汉语方言、少数民族语言和口头文化的实态语料，进行科学整理和加工，建成大规模、可持续增长的多媒体语言资源库，为后续的语言资源保护研究工作和深度开发应用工作，提供重要的基础性数据。

报告对中国语言资源保护工程的多媒体技术规范以及调查和整理阶段的多媒体语料采集、整理、加工方法进行了系统性介绍，并对如何利用日新月异的信息技术手段更高效地开展语保工程下一阶段工作进行了思考。

李宇明教授当选为中国辞书学会会长



2018年9月27日至30日，中国辞书学会第十二届年会暨学术研讨会在许慎故里河南漯河市成功召开。本次大会由中国辞书学会主办，漯河市人民政府承办，北京语言大学协办。语言资源高精尖创新中心

主任李宇明教授出席会议并当选为中国辞书学会新任会长。

原新闻出版广电总局副局长邬书林、教育部语言文字信息管理司副司长刘宏出席开幕式，并就中国辞书学会发展和新时代中国辞书事业发展做了讲话。原漯河市政协主席张社魁、中国辞书学会会长曹广顺致辞。中国辞书学会副会长、商务印书馆总经理于殿利主持开幕式。170余位来自全国各高校、科研机构和出版社的辞书学界和语言学界的专家学者参加会议。

本届年会一个重要议题是学会换届选举及章程修改。经过前期酝酿与协商，通过公开、透明、公正的选举程序，会议选举产生了以李宇明教授为会长的新一届常务理事会。新班子由来自辞书研究、教学、编纂、出版等机构和单位的代表组成。

新会长李宇明教授在闭幕式上从“辞书”“学”“学会”“中国”四个关键词切入，表示新一届理事会将在党的十九大精神统领下，团结全体会员，加强辞书学科建设，大力培养辞书人才，与时俱进，改革创新，为建设辞书强国、服务新时代文化建设事业做出新的贡献。

在主旨报告环节，李宇明教授作了《辞书与社会进步》报告。他从某大学数字图书馆的设计方案谈起，指出辞书是一个供人查检的知识体系，从而引出对辞书事业的一些思考。整个报告涉及辞书的本质，数字辞书在辞书发展史上的意义，以及未来的辞书学发展等内容。中国社会科学院学部委员江蓝生、商务印书馆总经理于殿利、南京大学教授魏向清分别作《〈现代汉语大词典〉的编纂理念与主要特色》、《从工具书到知识服务》、《知识服务驱动辞书创新研究》的主题报告。

此外，为弘扬中国辞书文化，打造辞书文化交流、研究和展示平台，依托漯河独特的许慎文化优势，漯河市

人民政府与中国辞书学会、商务印书馆签订协议，共建中国辞书文化博物馆。会上，举行了中国辞书展示基地揭牌仪式。李宇明教授作为学会会长参加了签字仪式，并与原漯河市政协主席张社魁共同为中国辞书展示基地揭牌。



本届年会学术气氛浓厚，以“新时代辞书发展与社会进步”为主题，组织了两场主题报告，并设置了“辞书维护与创新发

展”、“学习词典与用户需求”、“社长论坛：辞书发展与社会进步”等3个分论坛，对新时代辞书的发展进行了专项研讨。

年会上，最感人至深的环节就是致敬“辞书事业终身成就奖”专家。第四次中国辞书学会“辞书事业终身成就奖”共评选出11位辞书专家，他们是鲍克怡、晁继周、张万起、迟铎、汪耀楠、张柏然、周明鑑、杜瑞清、符淮青、李行健、孙义桢。这些老先生都年逾古稀，一生埋头苦干、兢兢业业，把毕生的精力和智慧都贡献给了辞书事业，取得了巨大的成就，获此殊荣可谓实至名归。“辞书事业终身成就奖”，是辞书人的最高荣誉。奖项设立的初衷，是要大力提倡为辞书事业甘于奉献的精神，树立光辉榜样，激励年轻的辞书工作者向他们学习，为辞书事业做出新的贡献。中国辞书学会曾于2006年、2008年、2011年三次开展奖项评选工作。

中心两位项目负责人获 2018 年政府特殊津贴



2018年享受政府特殊津贴 推荐人选名单公示	
北京大学	田新 冯 涛 潘晓军 陈学军 傅耀庭 刘永强 王树强 孙 华 陶新建 谢立宇 孙科峰 蓝安勇 郝建敏 郭 卫 余庆海 封卫东 郭耀庭 王向群 范海峰 陈 明 董 宇
清华大学	冯唯佳 孟 峰 杨殿阁 邵 峰 吕政英 张劲军 董树利 刘 杰 刘国辉 蔡 涛 张广海 梅 立 李 理 徐卫理 肖贵清 孙树强 曹本尧 屈 野 魏兆胜
中国人民大学	高福利 寇铁民 冯晓力 孟柏荣 刘耀军 林 嵩 伊志宏 尹晓飞 郭洪斌 王 斌
北京师范大学	冯晓敏 刘 勇 王岳存 莫基博 尹水水 鞠世峰 李 群 董文光 张耀庭 张金平
中国政法大学	张英俊 王 峰 李天恩 赵建凤 杨敏敏 郭静莹 江江强
北京师范大学	丁 毅 李 刚 李保林
北京语言大学	傅春宏 张劲松
北京科技大学	冯雪梅 曹和洪 徐华记 董德宝
北京理工大学	袁广峰 李春雷 蒋 治 卫 毅

北京语言大学语言资源高精尖创新中心项目负责人施春宏教授、张劲松教授入选教育部 2018 年享受政府特殊津贴人选。

施春宏教授：文学博士、教授，《语言教学与研究》主编。北京语言大学语言资源高精尖创新中心特聘研究员，语言资源高精尖创新中心与《语言教学与研究》合作平台项目负责人。

讲授的本科生课程主要有：语言学概论、现代汉语、语言应用理论与实践、现代汉语语法、对外汉语语法教学和词汇教学等；研究生课程主要有：语言学的观念和方法、汉语句法分析、语言学原理、现代句法理论和汉语句式分析、汉语语言学导论等。主要研究兴趣在汉语语法、词汇语义、语言应用和理论语言学方面，独立发表相关论文 50 余篇，合作发表论文数篇。出版专著《语言在交际中规范》（2005）、《汉语动结式的句法语义研究》（2008）、教材《作为第二语言的汉语概说》（2009），合著有《应用语言学理论纲要》（1999）、《语言哲学对话》（1999）等，并参与编撰《实用语文规范知识小词典》《学生多功能词典》《当代汉语小学生字典》等 5 部工具书。曾主持国家社科基金项目“句式构造和句式意义的互动关系研究”等多个句法、词汇课题和学校教改实验项目“语言学概论课程教改实验”。





张劲松教授：日本东京大学工学部电子情报工学专业博士、教授，北京语言大学语言资源高精尖创新中心特聘研究员，语言资源高精尖创新中心“面向智能语音教学的汉语中介语语音多模态语料库研究”项目负责人。

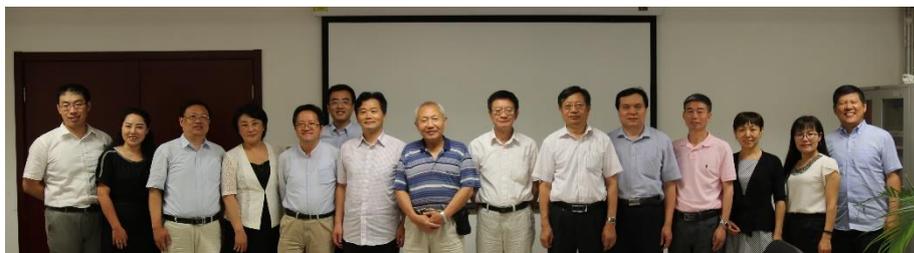
主要研究领域和方向：语音识别、合成技术、韵律建模、实验语音学、语音习得、基于语音识别、合成技术的计算机辅助语音教学技术。1990年以来主要从事语音识别系统和中文韵律信息处理的研究，参与了中国国家自然科学基金资助项目（“汉语连续语音识别和理解系统的研究”、“基于话者聚类的话者自适应语音识别技术的研究”、“基于韵律信息的汉语语音识别技术”）、日本总务省科研项目（“日中，日英语音翻译系统的研究和开发”）、日本文部省科研项目（“中文韵律及声调处理技术”）、日本情报通信研究机构科研项目（“日英、日中旅行会话翻译系统”）、教育部人文科学基地重大科研项目（“面向外国学生的计算机辅助汉语语音正音系统的研究”）等。目前的研究兴趣主要为基于计算机语音技术的高效二语语音习得技术。现为全国人机语音通信学术会议NCMMSC常设机构委员，NCMMSC学术交流工作组组长，ISCA（国际语言交流协会）、IPA（国际语音协会）、ASJ（日本声学协会）、中国语音、中文信息处理、中国声学等学会委员。担任国内外多家学术杂志审稿人工作，包括《IEEE Trans. On Speech and Audio Processing》、《Speech Communication》、《Computer Speech and Language》、《Journal of Phonetics》、《IEICE Trans. On Information and Systems》、《世界汉语教学》、《计算机工程应用》等。曾获ATR 2004 显著功绩表彰，2007IEICE 优秀论文奖。在国际学术会议及国内外

期刊上已发表 70 余篇研究论文，其中 SCI、EI、ISTP 可检索论文 20 余篇。

➤ 项目进展

“汉语中介语语料库建设创新工程”项目进展报告会顺利召开

2018年6月28日，语言资源高精尖创新中心（以下简称“中心”）“汉语中介语语料库建设创新工程”项目进展报告会在北京语言大学综合楼1006会议室召开。



本次会议邀请了北京大学袁毓林教授、杨德峰教授、

常宝宝副教授，北京师范大学朱瑞平教授、宋继华教授，北京语言大学宋柔教授、施春宏教授、曹文教授、刘长征教授、王治敏教授等专家出席。项目负责人张宝林教授及项目组部分核心成员和与会专家进行了深入交流。

张宝林教授首先从项目任务完成情况、存在的问题和解决方案、对未来工作的设想等三个方面进行了介绍，指出目前项目已经完成了语料收集、书面语语料录入、口语、视频语料转写、汉语中介语语料库标注规范的研制和语料标注平台的开发等工作。语料库检索系统开发技术总监郝振斌先生从技术架构、功能实现、进度计划三个方面对全球汉语中介语语料库综合平台的开发做了介绍。项目组成员饶高琦博士对项目所



支持的句法评测工作坊做了详细介绍，并提出句法评测工作坊旨在促进自动识别句法错误位置和类型的语言信息处理系统的发展。娄开阳副教授和王鸿滨副教授也对项目下一阶段实施提出了想法。张宝林教授指出该项目是一项极具挑战性的任务，集合了项目组所有成员的智慧，项目组将充分整合、共建、共享资源。

与会专家就项目设计理念、语料的收集标注和理论成果等方面提出了很多富有建设性的指导意见。宋柔教授肯定了语料库开放、市场化以及语料“浅层标注”的理念。袁毓林教授提出语料标注要充分考虑研究使用需求，并建议标注系统要和对外汉语教学系统做好对接。宋继华教授从“高精尖”的角度提出了项目要充分利用高新技术，探讨创新机制，要将“建资源、创模式”作为项目验收的重要依据。朱瑞平教授从语料库使用者的角度对语料建设提出建议，认为语料背景中的语别比国别更重要。杨德峰教授提出语料收集要包含个案追踪语料，语料标注要有专家干预。施春宏教授从多个角度对语料语体标注提出了看法，并建议要重视语体标注。曹文教授从“全球”概念出发，提出语料收集要具有针对性，要考虑所有国家全覆盖。刘长征教授针对语料收集、语料标注和项目成果等三个方面提出了可以进一步细化的建设指标。王治敏教授对语料库中母语者语料的可比性提出意见，并建议标注要实现人机交互。常宝宝教授从统计学的角度提出语料样本要有代表性，并建议要充分借鉴国外同类语料库的建设经验。

最后，中心常务副主任杨尔弘教授在总结发言中对项目组的辛勤工作和与会专家的不吝赐教深表感谢。她指出项目下一阶段需按照专家们的意见继续完善顶层规划，充分发挥团队力量，根据中心项目建设“顶天立地”的理念，扎实推进。

“新选中国名诗 1000 首：当代诗学名家经典选释系列”项目 进展交流会顺利召开



2018年7月15日，语言资源高精尖创新中心“新选中国名诗1000首：当代诗学名家经典选释系列”项目进展交流会在北京语言大学召开。中国社会科学院文学研究所陶文鹏教授，中华诗词学会副会

长、《中华诗词》杂志主编高昌先生，美国汉学界著名学者、伊利诺伊大学蔡宗齐教授，北京语言大学科研处处长聂丹教授，中心常务副主任杨尔弘教授出席会议，并与项目负责人韩经太教授及项目组核心成员北京大学钱志熙教授、葛晓音教授，中国传媒大学张晶教授，首都师范大学左东岭教授，华南师范大学蒋寅教授等进行了深入交流。

在研讨会上，项目负责人韩经太教授详细介绍了项目进展、阶段性成果以及成果海外传播计划等。他指出《新编“诗三百”》是该项目重要的阶段性成果，先有“诗三百”，终有“诗千首”，相信项目组各位专家的眼光、修养足以确保新选本能够经得起历史检验。项目组核心成员分别就本人负责部分的选目情况、注释体例、赏析视角等方面作了介绍。中心副主任刘晓海受项目负责人委托就项目宣传方案进行了说明。

与会专家就篇目选择、选释标准、翻译发行、宣传推广等方面提出了富有建设性的指导意见。陶文鹏教授和高昌先生肯定了项目的核心价值与意义，认为该项目选释诗目突出了“尽善尽美、脍炙人口”等特点，凸显了新时代“经典精选”的价值追求，同时再次强调选目须坚持“选名诗”的艺术标准，所选作品应具有涵养人性的审美文化功能。两位专

家还按照中国诗歌发展历史分期，对隋唐五代、两宋、明代和现当代等几个时期的选目样稿提出具体建议。蔡宗齐教授从项目成果海外传播和发行的角度建议可直接选用目前已有的权威名家翻译版本，同时可以考虑与国外年轻作者合作，改写选释诗目的“鉴赏”部分，以兼顾“体现中华文化精神的诗词经典选释”和“适宜于中外读者共同接受的诗词经典选释”，最终达到雅俗共赏、中外共享。

中心常务副主任杨尔弘教授对检查组专家及项目组专家的辛勤工作表示感谢，对项目取得的阶段性成果表示了肯定，并对项目提出两点希望：一是希望项目专家继续把握好成果的受众定位，拿出学术界和社会大众都认可的高质量成果；二是希望选释诗目要坚持“融思想历史和艺术审美为一体”的原则，凸显当代诗学名家的主体创造性和历代名诗的经典传承性，为学界树立新的审美文化标尺。

北京语言大学科研处处长聂丹教授认为该项目是中心极富特色的一个项目。语言与文化密不可分，诗词作为“语言艺术”的典范，是语言资源中最精华、最经典、最优雅、最富有文化特质的一部分。项目成果既面向学术界又面向社会大众，既面向国内诗词爱好者又面向海外中华传统文化爱好者，既达到了对内文化传承又实现了文化对外传播。聂丹教授指出该项目的终期成果会给中心的语言资源建设核心任务增加浓墨重彩的一笔。她祝贺项目组专家的辛勤耕耘初见成效，感谢与会的全体专家对北语科研的大力支持，并



诚挚邀请各位专家于秋冬时节相约北语开展系列讲座，让全校师生能够共飨名家的学识修养和项目的精华内涵。

“语言识别理论及语言数量统计的方法论研究”项目 进展交流会顺利召开

2018年7月20日，语言资源高精尖创新中心“语言识别理论及语言数量统计的方法论研究”项目进展交流会在北京语言大学召开。中心主任李宇明教授、中央民族大学丁石庆教授、北京语言大学司富珍教授出席会议，并与项目负责人黄行研究员及项目组核心成员南开大学冉启斌教授、中国农业大学许锋副教授、中央民族大学曲世锋老师进行了深入交流。

在研讨会上，黄行研究员详细介绍了项目进展情况、阶段性成果以及存在的问题。目前，项目三个子课题皆按计划不断推进：子课题一已根据《世界的语言》2017年第20版里全球7000多种语言的主要属性信息建成数据库，并正在进行中文翻译工作，此外还初步建成了世界语言“文字”数据库（据统计，世界已创建文字的语言有3870种，其中3017种使用的是拉丁字母）；子课题二的第一批理论成果已采取专栏形式集中发表于《语言战略研究》2018年第2期，为整个项目实施提供了比较全面的理论方法支撑；子课题三基于ASJP数据库，利用LDN距离计算方法、“拐点效应”（elbow effect）及分类树形图等科学方法，对世界语言与方言进行区分，子课题负责人冉启斌教授还计划在坚持学术标准的基础上，对相关变量和计算方法做进一步的探索，尝试在局部使用决策树模式进行语言分类研究，以构建一个兼容国际现有主流做法又体现中国特色的语言识别方法体系。

与会专家就语言识别的理论探索、方法研究、工作推进方案以及预期成果意义等方面提出了富有建设性的指导意见。丁石庆教授肯定了项目的核心价值，认为该项目立足学科前沿，攀登学术高峰，其成果能在国际舞台上发出中国声音，并希望项目组认真考虑语言和方言分类识别理论及方法的普适性与针对性，同时建议项目组做好项目进展规划和成果策略研究。司富珍教授对项目在研究语言划分时使用客观的跨学科方法表示认同，并认为语言与方言的识别和区分应主要着眼于语言本体，重视语言内部特征和规律，其他诸如政治、民族、文化等影响因子是次要因素，可作为参考。

李宇明教授最后对项目组的辛勤付出表示感谢，并指出对全球诸语言与方言进行分类与识别是世界性学术难题。该项目



以中国视角关注世界语言问题，不仅能让中国的学术声音传向世界，还能为解决国家周边战略安全问题提供重要的参考项。他希望项目组进一步扎实做好“四库”等基础性工作，用“已知”带“未知”，从理论和方法上解决全球不同类型、不同层面的语言及方言的划分和归类问题，并将项目成果用于解决我国语言资源保护工作中遇到的语言甄别、归类等问题。他建议项目组定期邀请国内外专家召开专题学术研讨会，针对世界语言身份识别和数量统计等问题不断碰撞思想，形成共识，切实为国家提供战略智库支持。

“面向冬奥会的机器翻译资源建设与技术研究”项目 进展检查会顺利召开

2018年9月17日，语言资源高精尖创新中心（以下简称“中心”）

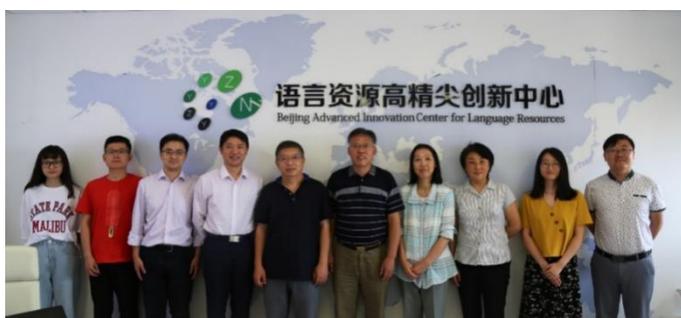
“面向冬奥会的机器翻译资源建设与技术研究”项目进展检查会在北京语言大学综合楼810会议室召开。会议邀请南京大学计算机科学与技术系陈家骏教授、北京交



通大学计算机与信息技术学院张玉洁教授、中国科学院自动化所赵军研究员等检查组专家出席会议，项目负责人清华大学计算机科学与技术系刘洋副教授与项目组部分核心成员对项目进展作了详细汇报，并与专家组进行了深入交流。会议由中心常务副主任杨尔弘教授主持。

刘洋副教授首先从项目立项背景、整体完成情况、存在的问题以及对下一步工作的设想等四个方面进行了介绍。目前，项目的汉英文本和语音平行语料库建设进展顺利，已收录近70万句对的领域文本语料和近10万句对的语音语料，汉英书面语和口语翻译系统在冬奥会数据集上取得了较好的翻译性能。另外，汉文-哈萨克文平行语料库的建设也即将启动。项目组成员王硕博士对项目开发的在线翻译系统、语音标注系统进行了现场演示。

与会专家就资源开发、应用前景和理论方法研究等方面提出了很多富有建设性的指导意见。专家组组长陈家骏教授对项目的建设及阶段性成果给予了肯定，认为该项目对提升北京冬奥会的多语言服务具有重大意义，并希望项目能够找准应用方向，做好系统工作。张玉洁教授结合历届奥运会的语言服务经验，从方法研究、技术实践的角度提出项目可以更加注重方法研究的体系化，并考虑引入人工评测机制。赵军研究员从应用的角度对项目提出意见，建议团队进一步明确应用场景与受众，



让项目落地后为北京冬奥会提供更有针对性的语言服务。

杨尔弘教授作了总结发言，她对项目组的辛勤工作表示感谢，并肯定了项目取得的阶段性成果。她表示，希望项目组充分吸收专家的建设性意见，进一步明确方向，把握重点，继续为北京冬奥会的语言服务和中心的“语言通”智能服务提供关键技术支撑。

“基于知识图谱的北京冬奥项目智能问答系统”项目 进展交流会顺利召开

2018年9月25日，语言资源高精尖创新中心（以下简称“中心”）



“基于知识图谱的北京冬奥项目智能问答系统”项目进展交流会在北京语言大学综合楼810会议室召开。会议邀请了中心学术委员会委员、中国科学院自动化所

宗成庆研究员、中国科学院自动化所赵军研究员、清华大学计算机科学与技术系李涓子教授、北京语言大学信息科学学院荀恩东教授出席会议，项目负责人中国科学院软件研究所孙乐研究员与项目组部分核心成员对项目进展、下一步工作要点作了详细汇报，并与专家组进行了深入交流。

交流会上，项目负责人孙乐研究员从项目整体规划、进展情况、取得的阶段性成果以及下一步工作计划等四个方面进行了介绍。项目立项至今以冬奥项目知识图谱的构建为核心，针对冬奥会信息服务存在冬奥知识碎片化问题、信息服务系统智能化和个性化不足、信息展示立体化和关联化不足的挑战，瞄准方向，扎实推进研究工作。目前主要的着力点在于数据资源收集，已收集了大量国内外与冬奥相关的半结构化与文本数据资源，领域实体、关系数据量突破7万，知识图谱包含了约83万个知识点。下一步项目组将通过构建QA（问答）对收集平台、积累训练语料等方式继续完善冬奥知识图谱。项目组成员韩先培副研究员及付成工程师对项目开发的冬奥智能问答原型系统与可视化知识图谱进行了现场互动演示。

互动演示环节结束后，与会专家就应用需求、领域资源建设和技术实践等方面提出了很多富有建设性的指导意见。专家组组长李涓子教授肯定了项目的核心价值与意义，领域知识获取与构建取得良好效果。宗成庆研究员就系统演示的实际效果，提出项目应该做好总体规则的制定，在系统的细节呈现方面多下功夫，同时建议重视问答对的资源积累。赵军研究



员从信息更新的角度建议项目要做好实体信息的动态更新与补充。荀恩东教授建议项目组应进一步聚焦系统的服务对象与应用场景，积极探索项目的实用性研究与学理性研究。

➤ 领域资讯

国际计算语言学协会亚太地区分会（AAACL）成立等资讯 3 则

一、冯志伟先生获自然语言处理与中文计算杰出贡献奖

中国计算机学会自然语言处理与中文计算大会于 2018 年 8 月 26 日-30 日在内蒙古呼和浩特举办。今年的杰出贡献奖得主为语言学家、计算机科学家冯志伟先生。杰出贡献奖于 2017 年设立，每年颁发给至多一位学者，首期得主为清华大学黄昌宁先生。同时，大会还为西湖大学副教授张岳，北京大学信息学院研究员孙栩颁发了青年新锐奖。



颁奖委员会为冯志伟先生撰写的颁奖词如下：

冯志伟教授是资深的自然语言处理和计算语言学专家，从事机器翻译、自然语言处理和计算语言学等基础理论研究和系统开发几十年，学术成就斐然，著作等身，已在多家著名出版社出版了二十余部专著，翻译国外重要论著数十篇（部），并用汉、英、法、德文等撰写了有代表性的重要学术论文百余篇，在国内外许多著名刊物上发表。

冯志伟教授年近八十，仍笔耕不辍，活跃在科研第一线，积极推动计算语言学的发展。他不仅在国际上担任很多知名组织、会议的委员、学术期刊的编委，也在国内计算机领域重要的学会中担任会长、理事、标准制定成员、评审委员、多所著名大学、研究所的兼职教授，以及核心期刊的编委等。

冯志伟教授是一位极为罕见的文理兼通、跨学科的研究型专家。他既懂得理科中的数学、物理、化学和计算机科学，又懂得语言学中的古代汉语、现代汉语、文字学、音韵学和普通语言学，深研过汉、英、法、德、俄、日等语言的语音、词汇和语法的自动处理，并把各方面的知识紧密地结合起来综合应用，在计算机上加以实现，完成各种研究任务。他在不同学科、不同语言研究中都取得重要的成就，分别成为这些学科的学术带头人。

二、百度王海峰博士出任 AACL 创始主席

第 56 届国际计算语言学协会 (Association of Computational Linguistics, ACL) 年会在澳大利亚墨尔本举行。开幕仪式上，ACL 主席 Marti Hearst 正式宣布成立国际计算语言学协会亚太地区分会 (Asia-Pacific Chapter of the ACL, AACL)。百度高级副总裁、AI 技术平台体系总负责人，ACL 前任主席王海峰博士出任 AACL 创始主席，清华大学副教授刘洋博士出任分会秘书。



ACL 成立于 1962 年，是世界上影响力最大、最具活力的国际学术组织之一。ACL 每年召开年会，供学者发布论文，分享最新成果。与会专家来自全球 60 多个国家和地区，是语言信息处理乃至人工智能领域最高级别的国际学术组织，代表了国际计算语言学的最高水平。为了更好地促进学术交流，1982 年和 1999 年，ACL 分别成立了欧洲分会 (EAACL) 和北美分会 (NAACL) 两个区域性分会。近年来，亚太地区在自然语言处理方面的研究进步显著。此次成立 ACL 亚太分会，将进一步促进亚太地区 NLP 相关技术和研究的发展。

据悉，王海峰是 ACL 五十多年历史上唯一出任过主席的华人，是截至目前最年轻的 ACL 会士。王海峰博士认为，人工智能的终极目标是让机器拥有像人一样通过学习掌握知识、理解语言，进而像人一样思考的能力。语言是知识和思维的载体，自然语言处理对于人工智能的意义不言而喻。

国际计算语言学协会亚太区分会将于 2020 年组织第一届 AAACL 会议，届时将与自然语言处理国际联合会议 (International Joint Conference of Natural Language Processing, IJCNLP) 联合召开。AAACL/IJCNLP 2020 日前已公开面向亚太地区征集会议承办单位。

三、第五届中文语法错误检测技术评测 CGED 举办

第五届中文语法错误检测技术评测 (Chinese Grammatical Error Diagnosis, CGED) 论文工作坊于 2018 年 7 月 19 日在第 56 届国际计算

语言学学会年会上举行。CGED 是自然语言处理领域关于中文，尤其是面向第二语言为汉语者进行语法错误自动检测、自动修改的权威技术评测，多年来与国际计算语言学学术会议（International Conference on Computer Linguistics, COLING）、国际自然语言处理联合会议（International Joint Conference of Natural Language Processing, IJCNLP）等顶级会议联办。技术评测旨在促进参赛选手利用自然语言处理技术和资源自动识别中文使用中的语法错误。

技术评测主席饶高琦博士在接受媒体采访时表示：“简单来说，我们的技术测评是促进学术界和工业界开发面向外国人学汉语的中文语法纠错系统。汉语母语者在语言关键期后，很少会犯语法错误，但对于第二语言学习者，这种错误常常发生。此外，儿童在小学、中学等阶段的学习中，也很容易犯错。”

该系列技术评测数据来自北京语言大学全球汉语中介语语料库项目（该项目在课堂上以及一些汉语类考试中积累真实的学生学习数据）。通过技术评测，将已加工为机读格式的数据形成训练数据集，供广大学者和企业开发出模型与算法，进行自动检错和改正，从而帮助第二语言学习者校对中文语法错误。

2015 年，哈尔滨工业大学团队夺得错误检测综合水平冠军；2016 年，阿里巴巴团队夺冠；今年，哈工大、科大讯飞联合实验室在错误检测和错误修正两方面均获得第一名，中国社会科学院语言研究所取得了第二名的成绩，阿里巴巴位列第三。据悉，目前该方向的系统尚不足以投入市场，其主要瓶颈在于数据稀缺。不过，该方向市场前景十分广阔。根据国家汉办的统计，近年来，每年平均新增注册汉语学习者达 210 万

人。汉语国际传播成为“中华文化走出去”和推进建设人类命运共同体等战略和构想的重要组成部分。在可预见的未来，本领域相应的技术、服务、产品需求十分巨大。

中文技术评测“亮相”国际顶级学术会议，也扩大了中文信息处理学术共同体在国际上的被关注度。据悉，2019年CGED技术评测将与基于经验主义方法的自然语言处理国际学术研讨会（Empirical Method Natural Language Processing, EMNLP）联合举办。

语言资源与智能领域 ESI 高被引论文 4 篇、热点论文 3 篇

基本科学指标数据库（Essential Science Indicators，简称 ESI），是目前普遍使用于衡量学科研究绩效、跟踪科学发展趋势的深度分析评价工具。通过分析 Web of Science（SCI/SSCI）收录的 11000 多种期刊的文献及其参考文献，可以定位学科领域中最具影响力的国家、机构、论文和出版物，以及研究前沿，并提供相应的排名数据。

某学科中的一篇论文，发表于十年内且被引频次位于该学科所有论文的前 1%，则称该篇论文为 ESI 高被引论文（Highly Cited Paper）；近 2 年内发表的论文，且在近 2 个月内被引频次排在相应学科领域全球前 1% 以内的论文则称为 ESI 热点论文（Hot Paper）。

由于 ESI 数据库为十年滚动数据，每 2 个月更新一次，不提供存档及永久使用权，更新后上一期数据在数据库中无法再查到，而高水平论文在引导创新、重视科研成果的影响力方面具有重要意义。因此，本栏

目旨在及时收集介绍语言资源与智能相关领域内具有代表性的高被引论文和热点论文，以飨读者。

一、语言资源与智能领域 ESI 高被引论文 5 篇

1. 《Word Sense Disambiguation: A Survey》（《词义消歧：一项调查》）

作者：ROBERTO NAVIGLI（意大利罗马大学）

发表时间：2009 年

发表刊物：《ACM COMPUTING SURVEYS》（《ACM 计算调查》）

摘要：Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner. WSD is considered as an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. We introduce the reader to the motivations for solving the ambiguity of words and provide a description of the task. We overview supervised, unsupervised, and knowledge-based approaches. The assessment of WSD systems is discussed in the context of the Senseval/Semeval campaigns, aiming at the objective evaluation of systems participating in several different disambiguation tasks. Finally, applications, open problems, and future directions are discussed.

词义消歧（WSD）是以计算方式识别语境中词语含义的能力。WSD 被认为是人工智能完备问题，也就是说，其解决方案至少与人工智能中最困难的问题一样难。本文介绍了解决单词歧义的动机，并提供

对任务的描述。文中分别概述了有监督、无监督和基于知识的方法。在 Senseval / Semeval 活动的背景下讨论了 WSD 系统的评估，旨在客观评估参与几个不同消歧任务的系统。最后，本文讨论了应用程序、开放问题和未来方向。

2. 《Orthographic Depth and Its Impact on Universal Predictors of Reading: A Cross-Language Investigation》（《正交深度及其对阅读普遍预测因素的影响：跨语言调查》）

作者：Johannes C. Ziegler, Daisy Bertrand, Dénes Tóth 等（法国 CNRS 和 AixMarseille 大学、匈牙利科学院心理学研究所等）

发表时间：2010 年

发表刊物：《PSYCHOLOGICAL SCIENCE》（《心理科学》）

摘要：Alphabetic orthographies differ in the transparency of their letter-sound mappings, with English orthography being less transparent than other alphabetic scripts. The outlier status of English has led scientists to question the generality of findings based on English-language studies. We investigated the role of phonological awareness, memory, vocabulary, rapid naming, and nonverbal intelligence in reading performance across five languages lying at differing positions along a transparency continuum (Finnish, Hungarian, Dutch, Portuguese, and French). Results from a sample of 1,265 children in Grade 2 showed that phonological awareness was the main factor associated with reading performance in each language. However, its impact was modulated by the transparency of the orthography, being stronger in less transparent

orthographies. The influence of rapid naming was rather weak and limited to reading and decoding speed. Most predictors of reading performance were relatively universal across these alphabetic languages, although their precise weight varied systematically as a function of script transparency.

字母拼字法的字母-声音映射的透明度不同，英语拼写法不如其它字母脚本透明。英语的异常状态导致科学家质疑基于英语语言研究的成果的一般性。文章调查了语音意识、记忆、词汇、快速命名和非语言智能在五种语言中的阅读表现的作用，这五种语言（芬兰语，匈牙利语，荷兰语，葡萄牙语和法语）分别位于透明连续体的不同位置。来自 1,265 名 2 年级儿童样本的研究结果表明，语音意识是与每种语言阅读表现相关的主要因素。然而，其影响受到正字法的透明度的调节，在不太透明的拼字法中更强。快速命名的影响非常薄弱，且仅限于读取和解码速度。大多数读物表现的预测因素在这些字母语言中相对普遍，尽管它们的精确权重随着脚本透明度而系统地变化。

3. 《Independent effects of bilingualism and socioeconomic status on language ability and executive functioning》（双语和社会经济状况对语言能力和执行功能的独立影响）

作者： Alejandra Calvo, Ellen Bialystok（加拿大约克大学）

发表时间： 2014 年

发表刊物： 《COGNITION》（《认知》）

摘要： One hundred and seventy-five children who were 6-years old were assigned to one of four groups that differed in socioeconomic status (SES; working class or middle class) and language background (monolingual or bilingual). The children completed tests of nonverbal intelligence, language tests assessing receptive vocabulary and attention based on picture naming, and two tests of executive functioning. All children performed equivalently on the basic intelligence tests, but performance on the language and executive functioning tasks was influenced by both SES and bilingualism. Middle-class children outperformed working-class children on all measures, and bilingual children obtained lower scores than monolingual children on language tests but higher scores than monolingual children on the executive functioning tasks. There were no interactions with either group factors or task factors. Thus, each of SES and bilingualism contribute significantly and independently to children's development irrespective of the child's level on the other factor.

175名6岁儿童被分为四组，每组均具有不同的社会经济地位（SES；工人阶级或中产阶级）和语言背景（单语或双语）。孩子们完成了非语言智力测试，用于评估基于图片命名的接受词汇和注意力的语言测试，以及两项执行功能测试。所有儿童在基本情报测试中表现相同，但语言和执行功能任务的表现受到社会经济地位和双语的影响。在所有措施中，中产阶级儿童的表现优于工薪阶层儿童，双语儿童在语言测试中得分低于单语儿童，但在执行任务方面得分高于单语儿童。没有与群体因素或任务因素的相互作用。因此，无论孩子在另一个因素的水平如何，社会经济地位和双语均对儿童的发展做出了显著和独立的贡献。

4. 《The rise of social bots》（《社交机器人的兴起》）

作者：Emilio Ferrara, Onur Varol, Clayton Davis 等（美国南加州大学、印第安纳大学）

发表时间：2016 年

发表刊物：《COMMUNICATIONS OF THE ACM》（《ACM 通讯》）

摘要：BOTS (SHORT FOR software robots) have been around since the early days of computers. One compelling example of bots is chatbots, algorithms designed to hold a conversation with a human, as envisioned by Alan Turing in the 1950s. The dream of designing a computer algorithm that passes the Turing test has driven artificial intelligence research for decades, as witnessed by initiatives like the Loebner Prize, awarding progress in natural language processing. Many things have changed since the early days of AI, when bots like Joseph Weizenbaum's ELIZA, mimicking a Rogerian psychotherapist, were developed as demonstrations or for delight. Today's social bots are sophisticated and sometimes menacing. Indeed, their presence can endanger online ecosystems as well as our society.

BOTS（软件机器人的缩写）自计算机早期就已存在。机器人的一个引人注目的例子是聊天机器人，用于与人类进行对话的算法，正如 Alan Turing 在 1950 年所设想的那样。设计通过图灵测试的计算机算法的梦想推动了数十年的人工智能研究，见证了像 Loebner 奖这样的举措，奖励自然语言处理的进步。自人工智能早期以来，许多事情都发生了变化，当时像约瑟夫·魏森鲍姆（Joseph Weizenbaum）的模仿罗杰里安

心理治疗师的伊丽莎白（ELIZA）这样的机器人被开发为示威游戏或者是喜悦游戏。今天的社交机器人很复杂，有时也很危险。实际上，他们的存在可能危及在线生态系统以及我们的社会。

二、语言资源与智能领域 ESI 热点论文 3 篇

1. 《A Probabilistic Programming Language》（《Stan：一种概率编程语言 Stan》）

作者：Carpenter, Bob; Gelman, Andrew; Hoffman, Matthew D.等（美国哥伦比亚大学、哈佛大学）

发表时间：2017 年 1 月

发表刊物：《Journal of Statistical Software》（《统计软件杂志》）

摘要：Stan is a probabilistic programming language for specifying statistical models. A Stan program imperatively defines a log probability function over parameters conditioned on specified data and constants. As of version 2.14.0, Stan provides full Bayesian inference for continuous-variable models through Markov chain Monte Carlo methods such as the No-U-Turn sampler, an adaptive form of Hamiltonian Monte Carlo sampling. Penalized maximum likelihood estimates are calculated using optimization methods such as the limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm. Stan is also a platform for computing log densities and their gradients and Hessians, which can be used in alternative algorithms such as variational Bayes, expectation propagation, and marginal inference using approximate integration.

To this end, Stan is set up so that the densities, gradients, and Hessians, along with intermediate quantities of the algorithm such as acceptance probabilities, are easily accessible. Stan can also be called from the command line using the `cmdstan` package, through R using the `rstan` package, and through Python using the `pystan` package. All three interfaces support sampling and optimization-based inference with diagnostics and posterior analysis. `rstan` and `pystan` also provide access to log probabilities, gradients, Hessians, parameter transforms, and specialized plotting.

Stan 是一种用于指定统计模型的概率编程语言。Stan 程序命令性地定义了以指定数据和常数为条件的参数的对数概率函数。从版本 2.14.0 开始，Stan 通过 Markov 链蒙特卡罗方法（如 No-U-Turn 采样器，哈密顿蒙特卡罗采样的自适应形式）为连续变量模型提供完整的贝叶斯推断。使用诸如有限存储器 Broyden-Fletcher-Goldfarb-Shanno 算法的优化方法来计算惩罚最大似然估计。Stan 也是一个计算对数密度及其梯度和 Hessians 的平台，可以用于替代算法，如变分贝叶斯，期望传播和使用近似积分的边际推理。为此，建立 Stan 以便容易获得密度、梯度和 Hessians，以及算法的中间量，例如接受概率。Stan 也可以使用 `cmdstan` 包从命令行调用，使用 `rstan` 包通过 R 调用，使用 `pystan` 包通过 Python 调用。所有这三个接口都支持基于采样和优化的推理以及诊断和后验分析。`rstan` 和 `pystan` 还提供对日志概率、梯度、Hessians、参数变换和专门绘图的访问。

2. 《Bi-Level Semantic Representation Analysis for Multimedia Event Detection》（《多媒体事件检测的双层语义表示分析》）

作者：Xiaojun Chang ; Zhigang Ma ; Yi Yang 等（澳大利亚悉尼科技大学、美国卡内基梅隆大学）

发表时间：2017 年 5 月

发表刊物：《IEEE Transactions on Cybernetics》（《IEEE 控制交换技术》）

摘要：Multimedia event detection has been one of the major endeavors in video event analysis. A variety of approaches have been proposed recently to tackle this problem. Among others, using semantic representation has been accredited for its promising performance and desirable ability for human-understandable reasoning. To generate semantic representation, we usually utilize several external image/video archives and apply the concept detectors trained on them to the event videos. Due to the intrinsic difference of these archives, the resulted representation is presumable to have different predicting capabilities for a certain event. Notwithstanding, not much work is available for assessing the efficacy of semantic representation from the source-level. On the other hand, it is plausible to perceive that some concepts are noisy for detecting a specific event. Motivated by these two shortcomings, we propose a bi-level semantic representation analyzing method. Regarding source-level, our method learns weights of semantic representation attained from different multimedia archives. Meanwhile, it restrains the negative influence of noisy or irrelevant concepts in the overall concept-level. In addition, we particularly

focus on efficient multimedia event detection with few positive examples, which is highly appreciated in the real-world scenario. We perform extensive experiments on the challenging TRECVID MED 2013 and 2014 datasets with encouraging results that validate the efficacy of our proposed approach.

多媒体事件检测一直是视频事件分析的主要工作之一。最近提出了各种方法来解决这个问题。其中，使用语义表示对于人类可理解的推理所具有的有前景的性能和理想能力已得到认可。我们通常使用几个外部图像/视频存档，并将训练它们的概念检测器应用于事件视频。由于这些档案的内在差异，所得到的表示可能对某个事件具有不同的预测能力。尽管如此，没有太多工作可用于评估来自源级别的语义表示的功效。另一方面，认为某些概念对于检测特定事件有噪声是合理的。受这两个缺点的影响，我们提出了一种双层语义表示分析方法。关于源级，我们的方法是从不同的多媒体档案获得的语义表示的权重进行学习。同时，它抑制了嘈杂或无关概念在整体概念层面的负面影响。此外，我们特别关注有效的多媒体事件检测，几乎没有积极的例子，这在现实场景中受到高度赞赏。我们对具有挑战性的 TRECVID MED 2013 和 2014 数据集进行了大量实验，结果令人鼓舞，验证了我们提出的方法的有效性。

3. 《A linguistic intuitionistic multi-criteria decision-making method based on the Frank Heronian mean operator and its application in evaluating coal mine safety》（《基于 Frank Heronian 均值算子的语言直觉多准则决策方法及其在煤矿安全评价中的应用》）

作者: Hong-gang Peng, Jian-qiang Wang, Peng-fei Cheng (中国中南大学、湖南科技大学)

发表时间: 2018 年 6 月

发表刊物: 国际机器学习与控制论杂志 International Journal of Machine Learning and Cybernetics

摘要: Coal mine safety has been a pressing issue for many years, and it is a constant and non-negligible problem that must be addressed during any coal mining process. This paper focuses on developing an innovative multi-criteria decision-making (MCDM) method to address coal mine safety evaluation problems. Because lots of uncertain and fuzzy information exists in the process of evaluating coal mine safety, linguistic intuitionistic fuzzy numbers (LIFNs) are introduced to depict the evaluation information necessary to the process. Furthermore, the handling of qualitative information requires the effective support of quantitative tools, and the linguistic scale function (LSF) is therefore employed to deal with linguistic intuitionistic information. First, the distance, a valid ranking method, and Frank operations are proposed for LIFNs. Subsequently, the linguistic intuitionistic fuzzy Frank improved weighted Heronian mean (LIFFIWHM) operator is developed. Then, a linguistic intuitionistic MCDM method for coal mine safety evaluation is constructed based on the developed operator. Finally, an illustrative example is provided to demonstrate the proposed method, and its feasibility and validity are further verified by a sensitivity analysis and comparison with other existing methods.

煤矿安全多年来一直是一个紧迫的问题，在任何采煤过程中都必须解决这个问题，这是一个不可忽视的问题。本文着重于开发一种创新的多标准决策 (MCDM) 方法来解决煤矿安全评估问题。由于在煤矿安全评价过程中存在大量不确定和模糊信息，因此引入语言直觉模糊数 (LIFN) 来描述过程所需的评价信息。此外，定性信息的处理需要定量工具的有效支持，因此语言尺度函数 (LSF) 用于处理语言直觉信息。首先，针对 LIFN 提出距离，有效排名方法和 Frank 操作。随后，开发了语言直觉模糊 Frank 改进加权 Heronian 均值 (LIFFIWHM) 算子。然后，基于开发的算子，构建了一种用于煤矿安全评价的语言直觉 MCDM 方法。最后，提供了一个说明性的例子来说明所提出的方法，并通过灵敏度分析和与其他现有方法的比较进一步验证了其可行性和有效性。