基于语料库的我国职业性别无意识偏见共时历时研究*

朱述承[1]，刘鹏远[1*]，苏祺[2,3*]

（1.北京语言大学 信息科学学院，北京市 100083；2. 北京大学 外国语学院，北京市 100871；3. 北京大学 教育部计算语言学重点实验室，北京市 100871）

摘要：性别偏见是社会学研究的热点。近年来，机器学习算法从数据中学到偏见使之得到更广泛的关注，但目前尚无基于语料库的方法对文本数据中职业性别偏见的研究。该文基于标记理论，利用 BCC 和 DCC 语料库，从共时和历时两个层面考察了 63 个职业的性别无意识偏见现象。首先，以调查问卷的形式调研了不同性别和不同年龄段的人群对 63 个职业的性别倾向，发现和 BCC 语料库中多领域的职业性别偏见度呈显著的正相关。然后从共时的角度，利用 BCC 语料库中不同领域的语料，以及 DCC 语料库中 2018 年全国 31 个省级行政单位（不含港澳台地区）的报纸语料，发现从口语至书面语语体，大部分职业表现出对女性的性别偏见逐渐升高，且不同地区对职业的性别偏见存在着差异。最后，从历时的角度，利用 DCC 语料库 2005 至 2018 年的报纸语料进行统计分析，发现职业性别无意识偏见现象随着时间的推移，呈现总体弱化趋势。

关键词：语料库；职业；性别；无意识偏见；标记理论

中图分类号：TP391　　**文献标识码：**A

A Synchronic and Diachronic Study of Corpus-Based Gender Unconscious Bias in Occupations

ZHU Shucheng[1], LIU Pengyuan[1], SU Qi[2,3]

(1. College of Information Science, Beijing Language and Culture University, Beijing 100083, China; 2. School of Foreign Languages, Peking University, Beijing 100871, China; 3. Key Laboratory of Computational Linguistics (Peking University), MOE, Beijing 100871, China)

Abstract: Gender bias is a hot topic in sociology. In recent years, machine learning algorithms have learnt bias from data, made this topic gain more attention. However, there is no corpus-based approach to the study of occupational gender bias in text data. Based on the markedness theory, this paper examines the gender unconscious bias of 63 occupations in BCC and DCC corpora from the synchronic and diachronic levels. Firstly, the gender preference of 63 occupations among different genders and different age groups was investigated in the form of questionnaires. There is a significant positive correlation between the questionnaire and the occupation gender bias word frequency indicators in the BCC corpus. Then, from the perspective of synchronicity, using the corpus of different fields in the BCC corpus, and the newspaper texts of the 31 provincial administrative units (excluding Hong Kong, Macao and Taiwan) in the DCC corpus in 2018, most of the occupations are found from spoken to written language, showing a growing gender bias against women. There also are differences in occupational gender bias in different regions. Finally, from a diachronic perspective, using the DCC corpus from 2005 to 2018 newspaper texts for statistical analysis, it is found that the occupational gender unconscious bias phenomenon shows an overall weakening trend with the passage of time.

Key words: corpus；occupations；gender；unconscious bias；markedness theory

1 引言

“偏见”是人基于自身认知的一种态度，这种态度往往是针对一个群体的[1]。偏见是一

基金项目：教育部人文社会科学研究规划基金资助项目（18YJA740030）

* 通讯作者

种态度，是多种多样的，有职业偏见、种族偏见、性别偏见等等。社会学中，将性别偏见（Gender Bias）定义为对于一种性别较其他性别的倾向或歧视。性别偏见一直是社会研究的重点，随着世界平权运动的逐步兴起，越来越多的学者和团体开始关注“性别偏见”这一话题。目前，对于性别偏见的研究领域也十分广泛，从语言上的性别偏见，到教育中的性别偏见，再到职业中的性别偏见，国内外均有丰富的研究成果。普遍认为，职业性别偏见是一种客观存在的社会现象。社会学中，又将职业中的性别偏见定义为“职业性别隔离”（Occupational Gender Segregation），指男性群体和女性群体在不同职业中的分布不均衡，收入、职位等不公平的现象，这种现象在各个国家普遍存在，但程度有所不同。经济学中，则认为职业性别偏见会造成一定的经济效益损失。法学中，各国学者也呼吁立法，保障女性的就业权益。

语言学中，鲜有关注职业性别偏见的研究。大部分的研究是基于语言使用中的性别偏见现象进行的。男女在语言使用上具有显著差异[2-4]，进而可通过词汇、语法等不同层面判断作者性别[5]。不同语言间的性别偏见在词汇、语法、语用等不同层面也有不同程度的差异[6]。

语言作为社会交际的产物，不仅表达和反映了人们的思想，而且塑造了人们的思想世界观[7-8]。因此，透过语言剖析人们对职业的性别偏见便成为了一条行之有效的途径。目前发展最迅速的计算机领域，人们已发现人工智能一样存在着性别偏见[9]。人工智能的性别偏见实质上是训练数据中带有标注者的性别偏见。也因此说明，我们所使用的语料库中的语料数据也带有着我们个体乃至社会的性别偏见标签。因此，通过语料库研究职业中的性别偏见便成为了研究性别偏见的一种新形式，也成为今后研究的一种必然趋势。但目前还未有通过语言学，特别是语料库语言学对职业中性别偏见进行的相关定量研究。

本文将语言中的职业性别偏见定义为人们对从事某一特定职业的人群在性别上存在的无意识的、不明显的、潜在的不公平语言态度。例如，人们看到或听到“医生”一词，直接联想到的便是男性角色，这就是一种语言上的职业性别偏见。与之相对应的，是人们认知中的职业性别倾向，指的是人们主观上认为男性群体或女性群体更加适合某一特定职业的选择倾向，这种倾向可能会导致偏见。性别偏见可分为有意识偏见（Conscious Bias）和无意识偏见（Unconscious Bias）。无意识偏见不是偏见主体的主观意念，而是一种无意识的态度[†]。本文基于结构主义语言学中的标记理论（Markedness Theory）[10]，采用语料库方法，通过统计分析，研究了以下三个问题：

（1）不同性别、年龄的人群对特定职业的性别倾向是否存在差异？且语言能否反映出人们对于职业性别倾向的认知？

（2）在共时层面，不同语体、不同地区的语言中在职业性别偏见上是否存在差异？

（3）在历时层面，语言中职业性别偏见的程度经历了怎样的变化过程？

本文的组织结构如下：第一部分为相关工作，主要介绍职业性别偏见在不同学科中的相关研究。第二部分为研究设计，主要介绍了本文的数据来源、研究对象及研究方法。第三部分为调查与实验结果，主要分析了问卷调查的结果，并与语料库中的职业性别偏见度进行了相关分析。第四部分为共时历时分析，从共时和历时两个不同的角度，利用语料库的数据对我国的职业性别偏见现象进行描写刻画。第五部分为结语。

2 相关工作

2．1 社会学等领域的相关研究

社会学中，国内最早关于职业性别偏见的研究见于陆震（1994），该文指出“妇女在就业领域的培训、招聘、使用提升等各个环节受歧视，如今已是全社会公认的事实”[11]。之

[†] 若无特别指出，本文中的“偏见”均为“无意识偏见”。

后，国内的不少学者将焦点投向了女性就业偏见这一层面，普遍认为我国的就业市场存在着明显的性别偏见，女性的就业形势较男性更为严峻[12-13]。也有学者通过调查统计的方法，利用调查问卷和工资水平等指标验证了我国就业市场中客观存在着对女性就业的性别偏见现象[14-17]。

经济学中，有学者利用收入等相关测量指标并结合国家数据计算出不同地区、不同年份间的“职业性别隔离”指数，发现从历时角度上来看，我国的职业性别偏见日趋严重；而从地域上，东部地区的职业性别偏见较为严重[18-19]。

法学中，针对职业性别偏见现象，相关学者则强调通过建立健全法律法规完善保障我国女性就业的权益，进而防止职业性别偏见[20-24]。

国外的研究中，有学者对不同职业的性别刻板印象（Gender Stereotype）进行研究，人们认为工程师为具有男性气质（Masculine）的职业，而教师为具有女性气质（Feminine）的职业；在自我认知中，男性较女性认为男性更加适合军人、农民，以及研究型、技术型等职业[25-26]。

2．2 语言学领域中的相关研究

社会语言学领域关注语言与性别的关系，特别是言语性别分析。国内学者梳理了国内外针对汉语和英语的言语性别差异研究，普遍认为男女在言语表达上存在显著的差异 [2-4]。国外学者则利用计量语言学的指标对文本作者的性别进行鉴定，取得了较好的效果[5]。

在结构主义领域，多用标记理论进行性别偏见的研究。由于印欧语系大多属于屈折语，因此利用标记理论对印欧语系中的语言性别偏见现象研究较多。英语中带有女性后缀的词正逐年减少，表明了人们因观念的改变，影响了语言发展，逐渐减弱了语言上性别偏见现象[27]，且关于性别的形式标记、分布标记和语义标记可以体现出英语中的语言性别偏见[28]，英语中的职业称谓词和汉语中的社会称谓词也可以体现出性别偏见[29-30]。不同语言中语言性别偏见也存在着差异[6]。

语料库语言学与计量语言学领域，有学者研究利用语料库和定量研究的方法，发现不同人群的语言具有差异，其中包括男性和女性之间的语言差异，女性往往处于弱势地位[31-33]。

自然语言处理与人工智能领域，有学者发现机器学习中存在性别偏见现象，数据中隐含的偏见可能会导致机器学习过程中产生偏见[34]。语料库中的文本数据在收集过程中很可能就带入了收集者的偏见[35]。在识别图片中的复杂场景时，机器很容易根据带有性别偏见的数据信息而产生偏见，如在厨房中的人，机器更容易识别为女性[36]。

3 研究设计

3．1 研究数据

本文所用的共时不同文体的语料数据来自 BCC 语料库中文学、报刊、多领域、微博和科技 5 个模块的语料[37]。

所用的共时不同地区的语料数据来源于国家语言资源监测与研究中心平面媒体中心[‡]研制开发的 DCC 动态流通语料库，共选取了全国 31 个省级行政单位（不含港澳台地区）2018 年的 31 份报纸语料，规模共计 1.79GB，所选报纸名称及其所对应地区如表 1 所示。

所用的历时语料数据来源于 DCC 语料库中 2005 年至 2018 年共 14 年的《广州日报》、《南方周末》、《深圳特区报》、《华西都市报》、《今晚报》和《人民日报》6 份报纸的全部文本语料，其中 2016 年《今晚报》数据缺失，用当年《天津日报》的文本语料代替，2017 和 2018 年《广州日报》和《南方周末》数据缺失，分别用当年的《羊城晚报》和《南方都市报》的文本语料代替，全部报纸文本语料规模为 8.18GB。

[‡] http://cnlr.blcu.edu.cn

表 1 所选报纸名称及其对应地区

报刊名称	代表地区	规模	报刊名称	代表地区	规模
重庆日报	重庆	64.4MB	内蒙古日报	内蒙古	62.5MB
钱江晚报	浙江	67.2MB	沈阳晚报	辽宁	30.8MB
云南日报	云南	59.4MB	安徽日报	安徽	67.4MB
新疆日报	新疆	44.7MB	江西日报	江西	46.8MB
四川日报	四川	68.3MB	南京日报	江苏	65.5MB
西藏日报	西藏	56.5MB	吉林日报	吉林	61.6MB
滨海时报	天津	49.5MB	湖南日报	湖南	72.3MB
浦东时代	上海	30.2MB	湖北日报	湖北	76.0MB
陕西日报	陕西	76.2MB	哈尔滨日报	黑龙江	46.6MB
山西晚报	山西	71.9MB	郑州晚报	河南	61.5MB
齐鲁晚报	山东	23.2MB	燕赵都市报	河北	64.3MB
青海日报	青海	57.1MB	海南日报	海南	78.3MB
宁夏日报	宁夏	52.7MB	贵阳日报	贵州	29.3MB
兰州日报	甘肃	56.1MB	广西日报	广西	67.1MB
福建日报	福建	83.1MB	羊城晚报	广东	57.9MB
北京日报	北京	88.4MB			

3．2 研究对象

为了较全面的反映语言中各职业的性别偏见差异，利用网络资源并剔除在语料库中出现频次较少的职业，共选出 63 个职业作为考察对象。

表 2 所选 63 个职业名称及其职业类型

职业	职业类型	职业	职业类型	职业	职业类型
CEO	经管	军人	社会	翻译	艺术
按摩师	技能	科学家	研究	飞行员	研究
保安	社会	理发师	技能	服务员	事务
裁缝	技能	律师	经管	歌手	艺术
采购员	经管	漫画家	艺术	工程师	研究
出纳	事务	模特	艺术	工人	技能
厨师	技能	魔术师	艺术	管理员	事务
大使	经管	清洁工	技能	护士	社会
导游	经管	赛车手	技能	花匠	艺术
店员	经管	杀手	技能	化妆师	技能
法官	经管	商人	经管	画家	艺术
法医	研究	设计师	艺术	售货员	经管
会计	事务	兽医	研究	演员	艺术
机修工	技能	司机	技能	医生	研究
机长	研究	特警	社会	艺术家	艺术
记者	艺术	调酒师	技能	音乐家	艺术
建筑师	艺术	推销员	事务	邮递员	事务
教练	社会	外交官	经管	园艺师	研究
教师	社会	舞蹈家	艺术	运动员	社会
经纪人	艺术	舞者	艺术	主持人	艺术
警察	社会	消防员	社会	作家	艺术

根据霍兰德职业兴趣理论（Holland Vocational Interest Theory）将 63 个职业分为 6 个职业类别[38]。63 个职业名称及其在霍兰德职业兴趣理论中对应类型如表 2 所示。其中，艺术型（Artistic）职业偏好模糊自由的活动并在其中创造艺术作品，讨厌系统化的活动，想象力丰富，看重美的品质；社会型（Social）职业偏好对他人进行传授培训和教导，不喜欢与实物打交道，重视社会和伦理道德问题；经管型（Enterprising）职业喜欢领导角色和冒险活动，重视政治、经济上的成就；事务型（Conventional）职业偏好对数据资料进行明确有序的整理工作，看重商业和经纪上的成就；研究型（Investigative）职业偏好对各种现象进行观察、推理和分析，厌恶组织领导活动，看重科学研究；技能型（Realistic）职业偏好动手能力强和具体的任务，缺乏社交能力，擅长与物体打交道，看重具体事务的价值观。

3．3 研究方法

3.3.1 问卷调查

利用调查问卷的形式考察不同人群对 63 个职业的性别倾向认知。问卷的全部题项为必答项，需要参与者填写自己的性别、年龄段以及对 63 个职业的性别倾向。其量表采取李克特五分量表。参与者针对 63 种职业进行评分，调查参与者的认知中该职业更适合男性还是更适合女性。1 分为极适合女性，2 分为比较适合女性，3 分为无所谓，4 分为比较适合男性，5 分为极适合男性。若分数接近于 3，则表明该群体对该职业没有明显的性别倾向；若小于 3，则表明该群体倾向于女性选择该职业；若大于 3，则表明该群体倾向于男性选择该职业，且距离 3 越远，说明性别倾向性越明显。

3.3.2 基于语料库词频的职业性别偏见度测量

统计语料库中不同语料中的性别职业词词频。在这里，性别职业词指的是“男/女+63 种职业”，如“男医生”、“女机长”等等。根据标记理论，若一个职业中，出现“女+职业”的词频远高于“男+职业”的词频，则说明“女+职业”具有很强的标记特征，即为了强调女性特质（可能是突出女性取得该职业的不易或很少有女性能担任该职业），我们则认为该职业对女性有一定的偏见。具体到计算中，我们定义公式（1）为测量语料库词频的职业性别偏见度的方法。

$$R = 4 * \frac{f}{f+m} + 1 \qquad \text{公式（1）}$$

其中，f 为该职业女性职业（如：女医生）词频，m 为该职业男性职业（如：男医生）词频。根据标记理论，若女性职业词频在具有标记的职业词频（f+m）中所占比例很高，则说明该职业对女性的偏见较大。比值也消除了语料规模的影响。为了能与调查问卷的得分具有可比性，采用数学变换，得出职业性别偏见度，即 R。R 值越高，表明语言中该职业中对女性的偏见越大；R 值为 5 时，则男性职业词频为零，偏见度最大，与调查问卷的“极适合男性”相对应；R 值越低，表明语言中该职业中对女性的偏见越小；当女性职业词频与男性职业词频相等时，R 值为 3，说明该职业在语言上不存在性别偏见，与调查问卷的“无所谓”相对应；若 R 值小于 3，则表明语言中该职业可能对男性存在性别偏见。

4 调查与实验结果

4．1 调查问卷

于 2019 年 6 月 2 号至 6 月 5 号之间通过问卷星进行问卷调查，收回的有效调查问卷共 244 份。所有的参与者均为自愿参与，与其他利益因素无关，参与者的性别与年龄构成见图 1 和图 2。调查问卷的量表数据回收后利用 SPSS 针对调查问卷进行信度分析，调查问卷的总体及每个职业题项的克隆巴赫系数（Cronbach's Alpha）均高于 0.90，信度良好，可以进行下一步分析。接下来计算总体、不同性别群体和不同年龄群体对每一个职业的评分均值，

对不同群体结果进行比较，如表 3 所示。

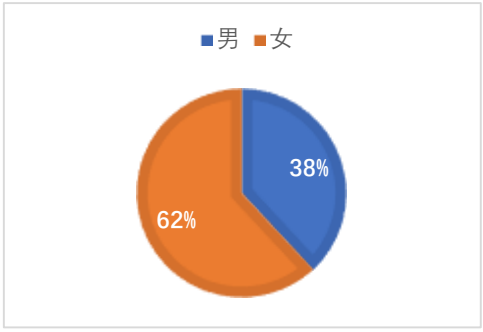

图 1 调查问卷参与者的性别构成

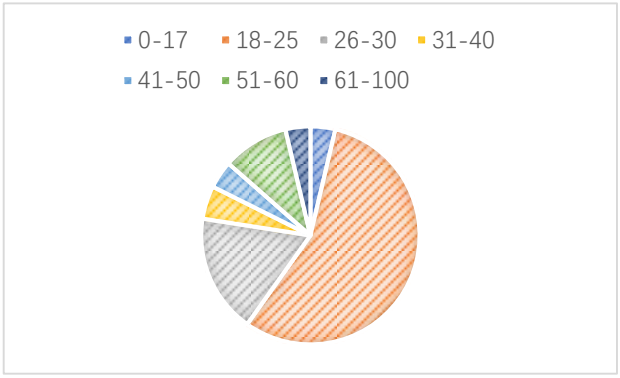

图 2 调查问卷参与者的年龄构成

表 3 总体及不同群体的职业性别倾向

	平均值 M	标准差 SD	p 值
男性群体	3.17	1.00	0.34
女性群体	3.14	0.82	
30 岁以上群体	3.20	0.96	0.16
30 岁以下群体	3.14	0.87	
总体	3.15	0.89	

如表 3 所示，总体来看，目前人们在认知上，没有表现出对职业有明显的性别倾向，且差异较小（M=3.15，SD=0.89）。到具体的职业上，人们认知中倾向于女性的职业（2.5 分及以下）有社会型的护士，事务型的出纳、会计，艺术型的化妆师、舞蹈家、模特。即人们对于事务型和艺术型这两类职业更倾向于女性。而人们认知中倾向于男性的职业（3.5 分及以上）有经管型的商人，技能型的司机、杀手、赛车手、机修工，社会型的警察、军人、消防员、保安、特警，研究型的飞行员、机长，事务型的邮递员。即人们对于社会型和技能型这两类职业更倾向于男性。

经统计学检验，男性群体和女性群体在职业性别倾向的认知上并没有显著性差异（p 值大于 0.05）。在具体职业方面，男性群体和女性群体认知差异最大的职业是护士和军人，均为社会型职业，但这种差异只是程度上的差异，且男性群体在认知上要更加极端一些，即男性群体和女性群体都认为女性更适合护士这一职业而男性更适合军人这一职业，但是男性群体的评分要比女性群体的评分更低或更高。男性群体和女性群体只有在艺术家、漫画家、画家和作家这四个艺术型的职业中有观念上的差异，即男性群体认为男性更适合这四类职业，而女性群体则认为女性更适合这四类职业。

将调查群体分为年龄在 30 岁以下的群体和年龄在 30 岁以上的群体这两个群体。经统计

学检验，这两个年龄段的群体在职业性别倾向的认知上也没有显著性差异（p 值大于 0.05）。在具体职业方面，30 岁以下群体和 30 岁以上群体在认知差异最大的职业是事务型的职业出纳和经管型的职业采购员，其中出纳是程度上的差异。30 岁以上群体在认知上要更加极端一些，即这两个群体都认为女性更适合出纳这一职业，但是 30 岁以上群体的评分要比 30 岁以下群体的评分更低。30 岁以上群体认为记者、经纪人和歌手这三个艺术型职业男性女性都适合，30 岁以下群体则认为这三个职业女性更加适合，但差异并不明显。30 岁以上群体认为男性更适合技能型的职业按摩师、艺术型的职业花匠和经管型的职业采购，而 30 岁以下群体则认为女性更适合。

4．2 职业性别偏见度

首先分别在 BCC 语料库中的报刊、文学、科技、多领域和微博模块中检索性别职业词，并统计词频，然后计算出不同职业在不同模块中的职业性别偏见度。为了尽可能地避免偶然性因素的影响，在统计过程中，删除了每一个模块中“女+职业”的频次小于等于 5 的职业。将每一模块中的职业按照指标从大到小排列，绘制折线图，如图 3 所示。

然后根据表 1 所选报纸，利用程序抽选并统计出 2018 年不同省级行政单位的性别职业词词频，将 63 种性别职业词词频按照女性和男性分别加和，并以此计算出每个省级行政单位 2018 年语言中的职业性别偏见度。根据值的大小绘制出全国（港澳台地区除外）热力图，如图 4 所示。

最后统计出 2005 至 2018 年 DCC 语料库中 6 份报纸语料中每年的男性性别职业总词频和女性性别职业总词频，并计算出每一年语言的职业性别偏见度，并绘制成图 5。接下来，将每年每种职业的性别职业词词频统计出来，删去女性性别职业词词频在 5 以下的职业（CEO、按摩师、裁缝、采购员、出纳、保安、厨师、大使、导游、法医、翻译、工程师、管理员、花匠、化妆师、工人、会计、机修工、机长、建筑师、经纪人、理发师、音乐家、邮递员、园艺师、消防员、舞蹈家、舞者、推销员、调酒师、特警、外交官、兽医、售货员、杀手、赛车手、清洁工、魔术师、漫画家），计算每种职业每年的职业性别偏见度。绘制出每种职业的职业性别偏见度的历时变化图，如图 6 至图 8 所示。

4．3 问卷调查与职业性别偏见度的相关性

将问卷调查的每一类职业的总体数据和 BCC 语料库中多领域模块计算出的每一类职业的职业性别偏见度利用 SPSS 进行相关性分析，得出的结果如表 4 所示。

表 4 职业性别偏见度与认知结果的相关性分析（注：*. 在 0.05 水平（双侧）上显著相关）

		职业性别偏见度	认知
职业性别偏见度	Pearson 相关性	1	.312*
	显著性（双侧）		.013
	案例数	63	63
认知	Pearson 相关性	.312*	1
	显著性（双侧）	.013	
	案例数	63	63

表 4 表明，BCC 语料库中的职业性别偏见度和认知在各个职业的性别偏见上呈显著正相关。因此，职业性别偏见度可以很好地反映人们对于职业性别偏见的认知，也因此证明我们选取的职业性别偏见度这一指标的可靠性与使用这一指标进行测量的可行性。

5 共时历时分析

5. 1 共时分析

5.1.1 不同语体中的职业性别偏见

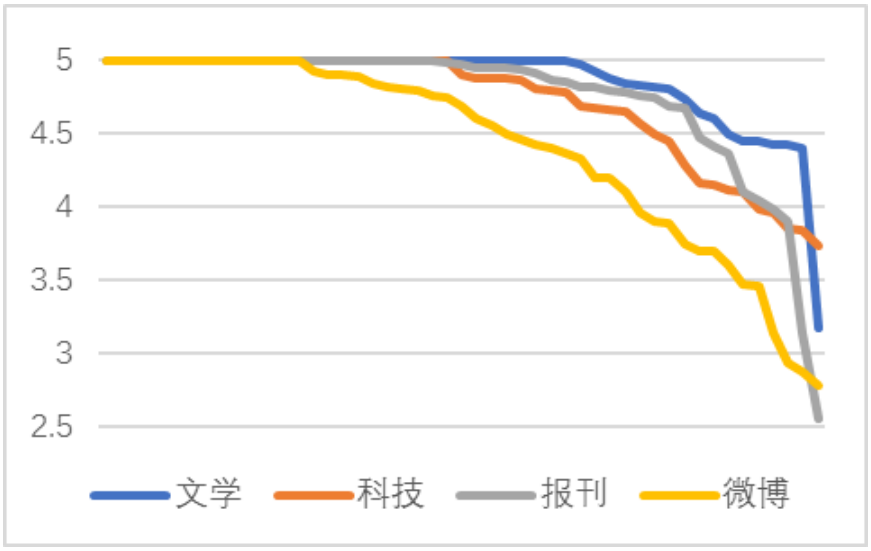

图 3 不同语体中职业性别偏见度分布

（每一模块中的职业按照职业性别偏见度从大到小排列，其中纵坐标为职业性别偏见度数值，值越高，则表明该职业对女性的偏见越高；横坐标为序列）

图 3 为 BCC 语料库中不同模块的职业性别偏见职业性别偏见度（公式（1））分布。由图 3 可知，微博模块的职业性别偏见度几乎一直位于报刊、科技和文学模块的下端，说明从报刊、科技和文学到微博，语言中对各职业的女性性别偏见逐渐降低。如果将微博视为口语语体，将报刊、科技和文学视为书面语语体，则首次发现从书面语语体到口语语体，语言中对各职业的女性性别偏见逐渐降低。

表 5 各语体中对女性偏见较小的职业

语体	职业	指标 R	类型	语体	职业	指标 R	类型
微博	杀手	3.89	技能	科技	运动员	3.99	社会
	理发师	2.88			教师	3.96	
	按摩师	2.79			舞者	3.86	艺术
	店员	3.61	经管		医生	3.84	研究
	运动员	3.96	社会		工人	3.74	技能
	护士	3.75		文学	工人	3.18	技能
	保安	3.14		报刊	演员	3.98	艺术
	医生	3.46	研究		模特	3.14	
	主持人	3.90	艺术		舞者	2.56	
	演员	3.70			运动员	3.90	社会
	舞者	3.70					
	魔术师	3.48					
	模特	2.94					

具体到每个模块的职业，根据职业性别偏见度（见公式（1）），将职业分为语言中对女性性别偏见较小的职业（R<4）和语言中对女性性别偏见较大的职业（R=5），其余则视为对女性性别偏见位于中间值的职业。如表 5、表 6。

由表 5 发现，艺术型的职业在语言中对女性的性别偏见较小，然后是社会型职业和技能型职业。语言中对女性的性别偏见较小的职业没有事务型的职业，说明该类型的职业普遍对女性具有性别偏见。具体职业中，工人、模特、舞者、演员、运动员、医生和主持人可以看作是语言中对女性性别偏见较小的职业。

表 6 各语体中对女性偏见较大的职业(R=5)

语体	职业	职类	语体	职业	职类
报刊	厨师 理发师 裁缝 清洁工	技能	微博	裁缝 清洁工	技能
	商人 大使 外交官 法官	经管		CEO 外交官 法官 商人	经管
	特警	社会		军人	社会
	推销员 邮递员 管理员 出纳 会计	事务		推销员 出纳	事务
	法医 兽医 机长	研究		法医	研究
	经纪人 音乐家 建筑师 设计师 艺术家 画家	艺术		音乐家 漫画家 翻译 建筑师	艺术
科技	清洁工 裁缝	技能	文学	清洁工 理发师 司机	技能
	外交官 大使 律师 商人	经管		律师 商人	经管
	特警	社会		教练	社会
	工程师	研究		管理员 出纳	事务
	舞蹈家 建筑师 艺术家 画家	艺术		工程师	研究
				舞蹈家 艺术家 主持人 画家	艺术

由表 6 发现，艺术型的职业在语言中对女性的性别偏见较强，然后是经管型职业和技能型职业。在各个语体类别中，均存在对女性性别偏见比较大的职业，且覆盖的职业类型全面。只有在科技语体中，对女性偏见较大的职业类型不包含事务型职业。具体职业中，裁缝、出纳、画家、建筑师、外交官、艺术家、商人和清洁工可以看作是语言中对女性性别偏见较大的职业。

有趣的现象是，艺术型和技能型的职业在语言中同时表现出对女性的性别偏见强和弱两

种特性。进一步观察可以发现，带有“家”或“师”后缀的艺术型职业普遍在语言中表现出对女性具有强的偏见性。

5.1.2 不同地区中的职业性别偏见

图 4 不同地区语言中职业性别偏见热力图

图 4 为我国不同地区（港澳台地区除外）的职业性别偏见度分布。由图 4 发现，2018 年各省级行政单位的报纸文体中的职业性别偏见在地理分布上存在着差异。图中颜色越深，表示语言中的职业性别偏见度越大，说明语言中各职业对女性的性别偏见越强。因此，对女性的职业性别偏见较小的区域集中在我国的中北部地区，西北、西南地区是我国对女性职业性别偏见较强的区域。

5. 2 历时分析

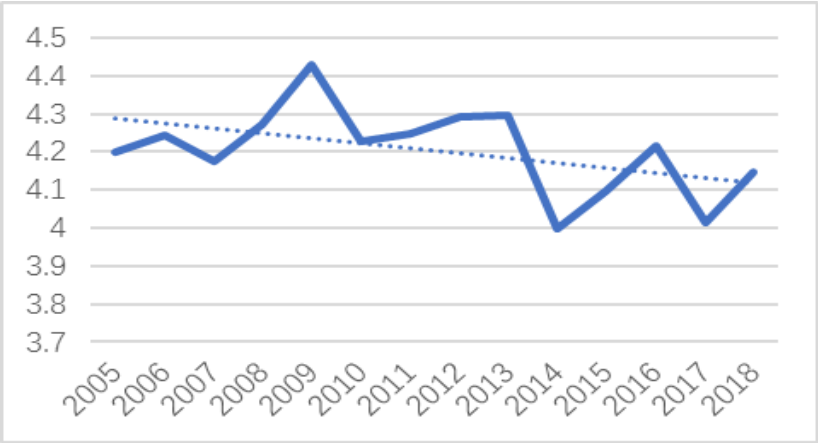

图 5 语言中职业性别偏见度历年变化图

图 5 为报纸语料中职业性别偏见度（见公式（1））的历时变化。由图 5 发现，14 年来，在语言中，各职业对女性整体上还存在着较大的性别偏见。不过从总体趋势来看，职业中对女性的性别偏见呈现下降趋势，虽然仍有反复，但自 2014 年起，指标较前 9 年维持着相对低的水平，说明近 5 年来，随着平权运动的不断兴起，越来越多的人们开始注意到语言中的职业性别偏见现象，并开始注意自己的用语，从而使语言的使用方面有所改变。特别是 2014 年联合国妇女署推出的“他为她”（HeForShe）行动，更将人们的关注目光引向了男女平权运动。一方面，平权运动的浪潮一波接一波的兴起，从国外一直影响到国内；另一方面，英

语等语言也逐渐走向了去除女性标记的道路，随着语言的接触，汉语使用者的认知也在悄然发生着改变，进而影响到汉语中，表现在语言中对女性的性别偏见程度呈现降低趋势。但在2009年的数据中，则出现了异常高的值，我们推测这与2009年国庆阅兵仪式有关，当年的阅兵仪式出现了大量女兵兵种，如女飞行员。因为从军者在大众认知中还处于男性性别倾向较重的职业，因此在媒体报道中使用的女性性别职业词激增，导致该年数据异常高，也印证了从军者这一职业有较强的女性性别偏见。从经济学的角度上，有学者经研究得出近年来，我国的职业性别隔离水平呈现日益严重趋势[12]。我们认为经济学上的职业性别隔离已经成为了一种歧视，是一种行为，而语言中的职业性别偏见还处于一种态度，从态度到行为的转变还需要一定的过程和时间。

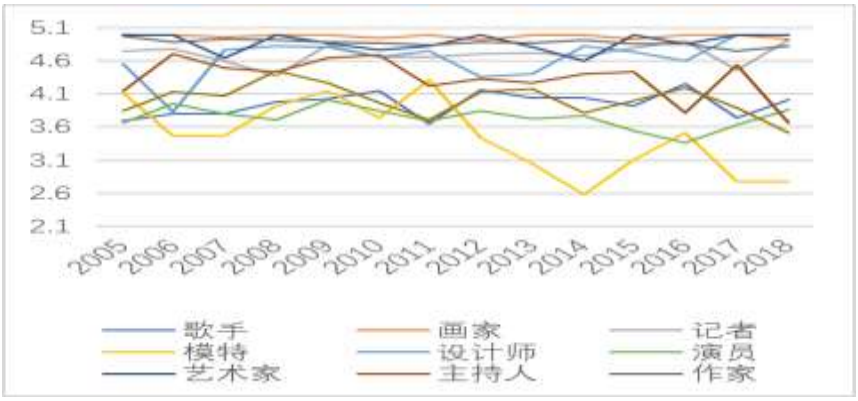

图6 艺术、技能型职业性别偏见度历时变化图

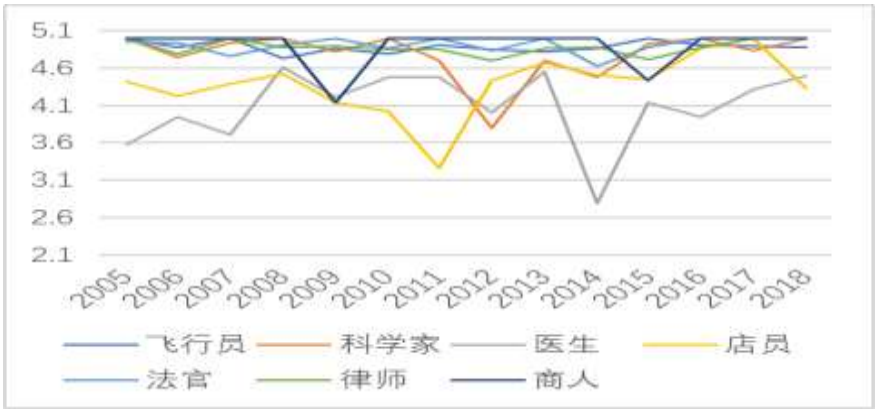

图7 研究、经管型职业性别偏见度历时变化图

图8 社会、事务型职业性别偏见度历时变化图

图6至图8为各具体职业的职业性别偏见度历时变化趋势，发现多数职业历年来的走势平稳。总体来看，护士、模特、运动员、歌手、医生在历年中可视为在语言中对女性性别偏

见较小的职业。尤其是护士，这一职业，在某些年份甚至表现出在语言上对男性存在性别偏见的现象。而语言中整体上的职业性别偏见呈现减小趋势可能与这些职业的性别偏见减小趋势有关。

当然，统计指标受诸多因素的影响，尤其是在图中一些变化差异较大的点，很可能受当时新闻或重大事件的影响，而并不能反映出语言中职业性别偏见的趋势。不过指标还是可以从总体上反映出语言和人们对社会现象的认知趋势，对我们了解语言和社会发展规律起到一定的帮助。

6 结语

语言作为一种社会中的交际工具，可以反映一个群体对社会现象的认知。通过基于语料库的方式和计量统计手段，我们可以看出当今人们对于职业中的性别倾向已经不那么明显，对于男性还是女性应该选择什么样的职业，无论男性群体还是女性群体，无论是年轻人还是中年人，亦或是老年人，似乎都没有表现出巨大的差异，人们普遍认为选择什么样的职业只关乎自己的兴趣与能力，与自己属于什么样的性别，好像并无太大关系。而利用语料库这一手段，我们更可以看出语言中关于职业性别偏见的蛛丝马迹。在共时层面，从口语语体到书面语语体，语言中各职业对女性的性别偏见越来越大；在不同地区的语言中，职业性别偏见也存在差异，我国中北部地区可以看作是语言中各职业对女性性别偏见较小的地区。从历时的角度来看，各职业从整体上表现出对女性的性别偏见逐年降低的趋势。与经济学上的结论相比较，我们呈现比较乐观的态度，即随着人们的认知不断改变，职业性别偏见这一现象将在未来逐渐减弱。

通过语言学，特别是语料库语言学的研究，不仅从语言现象上首次对职业性别偏见这一现象进行了刻画，而且用定量的方法也更能看出一种趋势，为未来的语言和社会发展做出一定预测。在未来的研究中，我们也希望发现更多的指标，对职业性别偏见这一现象作出更加合理客观的刻画和评估。当然，本研究还不能完全消除新闻事件等其他变量对指标的影响。但从总体趋势看，我们有理由相信，随着社会的不断发展和进步，人们对于职业的选择将不再受到性别这一自然生理属性的约束，人们也将更加自由地按照自己的兴趣和能力选择合适的职业。语言作为反映社会变化的镜子，我们也大胆预测，汉语中的职业性别偏见现象将会越来越少，而我们的社会也会变得越来越包容，越来越美好。

参考文献

[1] 陈国明.跨文化交际学[M].上海:华东师范大学出版社,2009:55-58.
[2] 戴炜栋.言语性别差异分析综述[J].外国语(上海外国语学院学报),1983(06):3-7.
[3] 李经伟.西方语言与性别研究述评[J].解放军外国语学院学报,2001,24(1):11-15.
[4] 史耕山,张尚莲.国内语言性别差异研究概述[J].外语教学,2006(03):24-27.
[5] Simaki V, Aravantinou C, Mporas I, et al. Sociolinguistic Features for Author Gender Identification: From Qualitative Evidence to Quantitative Analysis[J]. Journal of Quantitative Linguistics, 2016:1-20.
[6] Hellinger M, Motschenbacher H(Eds.). Gender Across Languages (Vol. 4). Amsterdam: John Benjamins Publishing Company, 2015:1-311.
[7] Sapir E. The Status of Linguistics as A Science[J]. Language, 1929, 5(4):207-214.
[8] Whorf B L, Carroll J B. Language, Thought, and Reality[M]. Massachusetts: The MIT Press, 1956: 23-33.
[9] James Z, Londa S. AI Can Be Sexist and Racist—It's Time to Make It Fair[J]. Nature, 2018, 12(559): 324-326.
[10] Richards J et al. Longman Dictionary of Applied Linguistics[M]. London: Longman, 1985:1-5.
[11] 陆震.妇女就业领域诸问题之我见[J].妇女研究论丛,1994(03):16-20.

[12] 云卉.中国女性就业性别歧视研究[D].哈尔滨工程大学硕士学位论文,2007.
[13] 金窗爱.中国当代女性就业问题研究[D].东北师范大学博士学位论文,2012.
[14] 毛海强.我国就业歧视研究[D].武汉科技大学硕士学位论文,2005.
[15] 季素萍.大学生就业中性别歧视：现象及影响因素[D].南京理工大学硕士学位论文,2006.
[16] 陈永伟,周羿.职业选择、性别歧视和工资差异—对我国城市劳动力市场的分析[J].劳动经济研究,2014,2(01):49-75.
[17] 曹娜.从性别刻板印象的视角解读女性研究生求职遭遇性别歧视[D].湖南师范大学硕士学位论文,2007.
[18] 张成刚,杨伟国.中国劳动力市场转型阶段职业性别隔离的新测度——基于 K-M 分解方法[J].人口与经济,2018(06):53-63.
[19] 任志敏.中国男女收入差距的分析[D].首都经济贸易大学硕士学位论文,2018.
[20] 朱懂理.试论我国消除就业与职业歧视立法[D].华东政法学院硕士学位论文,2004.
[21] 郭亦彦.《消除就业和职业歧视公约》实施研究[D].湖南大学硕士学位论文,2009.
[22] 刘冬.论我国女性就业歧视问题及法律对策[D].西南财经大学硕士学位论文,2010.
[23] 韩红颖.我国职场性别歧视的法律应对研究[D].江南大学硕士学位论文,2011.
[24] 游晓瑜.性别歧视的劳动法规制研究[D].上海师范大学硕士学位论文,2018.
[25] White M J, White G B. Implicit and Explicit Occupational Gender Stereotypes[J]. Sex Roles, 2006, 55(3-4):259-266.
[26] Ramaci T, Pellerone M, Ledda C, et al. Gender Stereotypes in Occupational Choice: A Cross-Secti -onal Study on A Group of Italian Adolescents [Erratum][J]. Psychology Research and Behavior Management, 2017, Volume 10:155-156.
[27] 苏晓玉.谈谈英语词汇中的女性后缀[J].解放军外国语学院学报,2000(03):24-25+54.
[28] 苗兴伟.从标记理论看英语中的性别歧视[J].四川外语学院学报,1995(03):51-55.
[29] 韦晓曙,陈佳敏.英语职业称谓性别歧视语及应对策略[J].科教导刊(上旬刊),2017(07):146-148.
[30] 张莉萍. 称谓语性别差异的社会语言学研究[D].中央民族大学博士学位论文,2007.
[31] Baker, P. Using Corpora to Analyze Gender. London: A&C Black, 2014: 19-176.
[32] 许家金,李潇辰.基于 BNC 语料库的男性女性家庭角色话语建构研究[J].解放军外国语学院学报,2014,37(01):10-17+30+159.
[33] 王显志.英汉语性别歧视现象的对比研究[D].中央民族大学博士学位论文,2010.
[34] John P, Penny P, Ernest J. et al. Big Data: Seizing Opportunities and Preserving Values[R]. Washington D.C.: Executive Office of the President, 2014.
[35] Durme B V. Extracting Implicit Knowledge from Text[D]. PhD essay of University of Rochester, 2009.
[36] Zhao J, Wang T, Yatskar M, et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: 2017 Association for Computational Linguistics, 2017: 2979-2989.
[37] 荀恩东,饶高琦,肖晓悦,臧娇娇.大数据背景下 BCC 语料库的研制[J].语料库语言学,2016,3(01):93-109+118.
[38] Holland J L. A Theory of Vocational Choice[J]. Theory & Practice of Vocational Guidance, 1968, 6(1):35-47.

作者简介：

朱述承（1994——），男，硕士研究生，主要研究方向为计算语言学；
刘鹏远（1974——），通讯作者，男，副研究员，主要研究方向为自然语言处理；
苏祺（1979——），通讯作者，女，副教授，主要研究方向为语料库语言学。

作者联系方式：

朱述承 北京市海淀区学院路 15 号北京语言大学信息科学学院 邮编 100083 电话 18701587631 电子邮箱 zhu_shucheng@126.com
刘鹏远 北京市海淀区学院路 15 号北京语言大学信息科学学院 邮编 100083 电话 13911510796 电子邮箱 liupengyuan@pku.edu.cn
苏祺 北京市海淀区颐和园路 5 号北京大学外国语学院 邮编 100871 电话 62752364 电子邮箱 sukia@pku.edu.cn