新文科视域下的计算社会语言学研究 *

袁毓林

（澳门大学人文学院中国语言文学系　澳门　999078）

［摘要］新文科建设的目标之一是：多种学科在研究内容和理论方法上进行交叉与融合，从而形成新的学科方向、研究范式，以及相应的复合型人才培养模式。为此，本文提出一种可操作的路径：利用已经发展起来的若干第一代交叉学科，融汇成第二代交叉学科。比如，对于“社交网络—语言的社会变异—计算建模”这种头绪繁多的研究领域，可以借助“计算语言学”与“社会语言学”等第一代交叉学科，形成第二代交叉学科“计算社会语言学”。这种操作路线可以细化与落实新文科建设，并提供必要的学科规训。文章还以当代社会的“物理—精神—信息”三元空间，以及其中说话者的社会身份、个体人格和社交互动对语言使用和变异选择的影响为例，阐明计算社会语言学的理论、方法和议题。

［关键词］新文科建设；第一 / 二代交叉学科；计算 / 社会语言学；三元空间

［中图分类号］H087［文献标识码］A［文章编号］1003–5397（2024）01–0005–12

DOI:10.16499/j.cnki.1003-5397.2024.01.005

Computational Sociolinguistics from the Perspective of New Liberal Arts

YUAN Yulin

Abstract: One of the goals of the new liberal arts is the intersection and integration of multiple disciplines in research content and theoretical methodologies, thus forming new disciplinary directions and research paradigms, as well as corresponding inter-discipline talents training models. To this end, this paper proposes an operational path: integrating several first-generation interdisciplinary fields that have already been developed into a second-generation interdisciplinary field. For example, in the research field of “social networking-social variation of language-computational modeling”, which is so complicated by multiple disciplines, we can use the first-generation interdisciplinary fields such as “computational linguistics” and “social linguistics” to form the second-generation interdisciplinary field of “computational social linguistics”. This path can refine and promote the development of the new liberal arts, and provide necessary disciplinary regulations. In addition, theories, methodologies and issues of computational sociolinguistics are also exemplified in this paper by the “physical–mental–information” space of contemporary society, where speakers’ social identities, individual personalities and social

［收稿日期］2023-08-30

［作者简介］袁毓林，澳门大学教授，主要研究理论语言学和汉语语言学。

* 本研究得到澳门大学讲座教授研究与发展基金（CPG2023-00004-FAH）和启动研究基金（SRG2022-00011-FAH）的资助，还承编辑部提出修改意见，谨此致以谢忱。

interactions impose influence on language use and language variation choices.

Keywords: development of new liberal arts; first/second-generation interdisciplinary; Computational/Sociolinguistics; ternary space

一　新文科建设视域中的社会计算

新文科建设的一个重要特征是，建设一批多学科交叉的人文社会科学的新兴研究领域或研究方向。这种愿景和理想是非常美好的，有助于培养能够应对未来挑战的跨学科复合型人才；但是，实现起来却是十分艰难的，从具体的学科群选择与组合、研究目标设计到操作路线与实施方案的制订，都需要进行不断的探索和尝试；并且，要冒着因尝试失败而沉没了机会成本的风险，甚至还可能要背上误人子弟的骂名。因此，不同的学科怎样寻找相关的伙伴学科，形成有议题（研究内容新颖而且重要）、可操作（有技术支撑）、有发展前途（带来理论突破或应用落地）的交叉学科，是新文科建设成败的关键要素和重中之重。

一般的印象，新文科不同于传统文科的一个标志是：有意识地在人文社会科学的研究内容或方法中，系统性地融入当代前沿的科学技术，以期形成新的学科方向和研究范式，以及相应的复合型人才培养模式。比如，随着大数据技术的跨越式发展，“计算”已经越来越成为人文社会科学领域的关键词。与“计算”相结合的多学科 / 交叉学科 / 跨学科研究领域也开始大量涌现，形成了数字人文（Digital Humanities）、社会计算（Social Computation）、计算社会科学（Computational Social Science）、计算传播学（Computational Communication）、计算社会学（Computational Sociology）等新兴学科或研究方向。[①]值得一提的是，2009 年，哈佛大学的 David Lazer 联合从事信息科学、社会学和物理学的 15 位学者，在 Science 杂志上联名发表文章（Lazer et al., 2009），首次提出了“计算社会学”（Computational Sociology）这一新兴学科。他们阐述了利用计算手段，从大数据中揭示社会学规律的学术思想和趋势，这标志着社会学研究进入数据计算时代。当代科学技术的显学是计算机科学技术和网络大数据技术，“计算社会学”显然可以算是“新文科”。近年来，计算社会学已成为人文社科领域最重要的研究范式。《科学》（Science）、《自然》（Nature）和《美国国家科学院院刊》（PNAS）等国际知名学术期刊上，涌现出大量计算社会学的研究成果；众多学术期刊出版专刊介绍计算社会学研究的进展；美国还成立了计算社会学学会；George Mason 大学也成立了计算社会学系，并成为世界上第一个正式授予计算社会学博士学位的机构。计算社会学无论对于揭示人类与社会规律，还是对于用户的个性化服务，均具有重要的意义。因此，基于社会媒体大数据的计算社会学研究，在学术界和产业界引起了广泛关注。这可以说是国际学术界新文科开拓与建设的一个成功范例，可以为我们的新文科建设提供经验。

我们认为，就利用社会媒体大数据的社会计算和计算社会学而言，并非只有原来从事计算机科学技术和社会学的学者才可以涉足。其实，语言学研究者也可以积极参与，并且还是可以大有作为的。因为，社会媒体的海量数据中，绝大部分是用自然语言写成的长短不一的文本，其中蕴藏了丰富的与用户有关的复杂的社会信息，是社会学、心理学（特别是社会心理学）和语言学（特别是社会语言学）等学科的重要研究对象。但是，这些学科所需的信息都隐藏在复杂的语言背后，需要利用自然语言处理和理解技术挖掘出来，才能被计算社会学研究进一步利用。而这种对语言数据的处理，正好是语言学、

自然语言处理和计算语言学等学科的强项。

近年来，随着机器学习和自然语言处理技术的发展，如何更好地分析社会媒体大数据中的自然语言（即文本信息），已经成为社会计算、计算社会学研究的热点，吸引了众多不同学科学者的研究兴趣，学科体系与范式已初具规模。并且已取得了不少令人鼓舞的成果，主要集中在以下四个方面：词汇的时空传播与演化；语言使用与个体差异；语言使用与社会地位；语言使用与群体分析。[②]诸如此类的研究，不仅具有社会学、心理学和语言学方面的学理价值，而且在舆论监测、社会管理和客户分析等方面，都具有重要的应用价值。

二　从“计算语言学”和“社会语言学”到“计算社会语言学”

众所周知，高等院校的学科体系和专业分工已经高度体系化。并且不同学科在学科传统、方法论和学术追求与价值观方面，也存在一定差别。这在历史上形成了所谓的“学科鸿沟”，或者如俗话所说的“隔行如隔山”。因此，要把不同的学科整合起来，殊非易事。对此，我们提出一种可操作的实践路径：尽可能有效地借鉴和利用已经发展起来的相关的第一代交叉学科，再次进行交叉与融合，从而形成第二代交叉学科。这样做好像是在比较坚固的旧楼上面加盖新楼，可以取得以旧出新、物尽其用、组合增效、事半功倍的效果。比如，面对“社交网络—语言的社会变异—计算建模”这种头绪繁多的研究领域，和相应的“大数据与计算科学—社会学—语言学”等多种学科和知识体系，可以借助已经成熟的“计算语言学”（Computational Linguistics，CL）与“社会语言学”（Sociolinguistics）等第一代交叉学科，形成“计算社会语言学”（Computational Sociolinguistics，CS）这种第二代交叉学科，从而使得相关的新文科建设不仅思路清晰、基础扎实，而且有迹可循、有章可依。

关于第一代交叉学科“计算语言学”的思想源头，可以追溯到研制电子计算机的当初。著名的“图灵测试”（Turing Test）就是以自然语言理解与翻译为思考背景的。相对于“自然语言处理”（natural language processing，NLP）和“自然语言理解”（natural language understanding，NLU）的研究方向与工程领域，学者们提炼出了“计算语言学”这种学科建制与学科体系，以利于学术探索和人才培养。

关于第一代交叉学科“社会语言学”的发展与兴盛，可以归功于 Labov（1966）和 Weinreich 等（1968）一系列关于语言与社会共变关系的研究。他们采用口头访问、书面问卷和民族志等方法，系统地调查和研究了说话者的性别、年龄、地理位置、社会阶层和权力关系等社会结构对个体与社群语言使用的影响，发现了说话人有关社会变量（social variables）与语言变异（linguistic variation）之间的对应关系，揭示了语言使用的阶层差异与历史演变的重要规律，推动了社会学、心理学和语言学的深入与细化。

现在，Nguyen 等（2016）认识到，随着大数据的发展，相关学科正经历着一场范式的转变。除了聚焦于传统的自然现象描写、理论发展以及计算科学，数据驱动的探索和发现已经成为许多学科方法论框架的有机组成部分，而计算语言学也在此发展之列。考虑到以往计算语言学主要是捕捉语言的信息维度和语言信息传递的结构，对语言的社会维度关注很少。最近二十年来，受社交媒体大数据的驱动，计算语言学对研究社会环境中的语言的兴趣越来越浓。社交媒体平台上的大数据为计算语言学的研究提供了新方向，也具有方法论意义。当然，此方向也面临着一些挑战，比如：（1）比起计算语言学传统上用的语料，社交媒体中的语言更加口语化、变异也更多。（2）社会变量和语

言之间的关系更为动态和脆弱，这也不同于计算语言学以往所关注的文意和结构之间的相对固定的关联。另一方面，传统的社会语言学用量化或质性方法来研究口语语料，调查和民族志方法是收集语料的主要手段，但是其语料规模往往较小。随着类似社交媒体平台语料的出现，大规模的数据为语言变异研究提供了更为宽阔的舞台。面对这些更为庞大也更为异质的语料，社会语言学需要新的方法论，而计算语言学则正符合这一期待。于是，他们大胆地构想一个计算语言学和社会语言学相结合的、可以被称之为“计算社会语言学”的新兴交叉领域，并且明确其目标是从计算的角度研究语言与社会的关系。

该文详细讨论了“计算社会语言学”的原理、范围及方法论，讨论了说话者如何使用语言来塑造对其身份的感知，并重点讨论了基于性别、年龄和地理位置的语言变异模型的计算方法。还从单个说话者转向成对、成组和社区，讨论语言在塑造个人关系、改变风格以及规范社区语言变化方面的作用。讨论了多语言和社交互动，其中概述了处理多语言交流的工具，如分析器（parsers）和语言识别系统（language identification systems），并从计算角度分析多语言交流模式的方法。最后，还指出了“计算社会语言学”这一方向所面临的挑战，也即这个新兴的多学科研究领域的研究议程：扩展调查范围，调整方法框架以提高兼容性，根据社会语言学研究的需要调整自然语言处理的工具。

我们认为，这种基于若干成熟的第一代交叉学科来构建第二代交叉学科的做法，不仅可以用于指导细化与落实新文科建设，并且可以为新文科建设提供一定的学科规训（古拉丁文 disciplina，英文 discipline/disciplinarit）[③]。下面，根据 Nguyen 等（2016）和刘知远（2021）等材料，再结合笔者的语言学工作经验和文献阅读体会，简要介绍和讨论一下计算社会语言学的有关理论假设、研究方法和主要课题。

三　计算社会语言学的理论、方法与课题

作为第二代交叉学科，计算社会语言学尝试整合社会语言学和计算语言学的有关方面，从大数据和计算的角度研究人们的语言（变异）和社会（参数）之间的关系，探讨对相关语言内容及其社会背景信息的数据收集、计算建模、结果分析以及理论含义的揭示等一系列方法，以便在新的技术和学科背景上，加深对语言运用中社会动态的理解，对在社会环境中使用语言这一主题产生新的见解。并且，通过社会语言学的语言研究，来改进自然语言处理的工具与方法，帮助建立更加丰富的语言计算模型，从而为媒体上的文本及其内容处理提供更多的学术支持。比如，基于对用户语言选择的分析，检测用户的性别、年龄、地理位置，甚至性格特点、兴趣爱好等的研究，可能会给自动用户分析工具（如用户建档）带来好处。反过来说，这种注重语言社会变异的研究，可以超越经典的自然语言处理工具背后的典型假设，即语言使用的同质性（homogeneity），从而让相关的语言计算工具更加贴近互联网语言运用的实际生态。

在社会语言学研究中引入计算建模方法，这是由网络时代语言运用的实际生态所要求的。因为，随着移动互联网的普及，数字信息世界这个虚拟空间已经成为人类生活世界（life-world）中一个不可或缺的组成部分。我们的社会突破了传统的“物理世界—精神世界”这种二元空间，已经全面进入了“物理世界—精神世界—信息世界”这种三元空间。[④]人们在无处不在的信息空间中频繁交往，不断地通过语言使用来建构和塑造自己的线上身份，维护与管理自己的线上社会关系网络，从而在这种以计算机为媒介的交际（computer-mediated communication，CMC）中，形成了大量与用户的社会变量相关的语

言变异，为社会语言学的研究提供了大规模的鲜活的素材。并且，信息世界通过全民互联和迅速更新的方式，对人们的观念、行为、时尚和情绪等舆情和趋势产生多方位的实时影响。比如，在社会预测方面，社会媒体中关于候选人的提及率就是很好的预测指标。例如，根据 Facebook 上的支持率就能够成功预测 2008 年美国总统大选结果（Williams & Gulati，2009）。可见，社会环境的空间结构变化了，在社会环境中运用语言的实际生态也变化了；网络环境中的语言运用已非传统手工研究方式所能应付，计算建模方法是很好的选择。因此，数字化时代，对于社会语言学来说，计算建模不仅是一种方法论，更是一种认识论，计算社会语言学这一研究范式应运而生。

比如，在语料收集方面，社会语言学的传统做法是观察旁听、口头访谈和问卷调查，等等。显然，这是一个耗时费力的过程，而所得的数据集往往很小。现在，随着网络媒体的兴起，微博、论坛、评论等社交平台上用户生成的内容极为丰富，并且这些自然、非正式的语言往往带有上下文信息（比如，用户、社交网络机构、生成时间，等等）。在一定的计算手段的帮助下，这些内容成为传统数据收集方法的一个有力补充。这种计算社会语言学范式下收集起来的网络语料，自然地规避了 Labov（1972）所谓的“观察者悖论”（observer’s paradox）：社区语言研究的目的必须是发现人们在没有被系统观察时是如何说话的，然而我们却又只能通过系统观察来获得这些数据。此外，计算语言学领域常用于获得各种大规模标注数据的“众包”（crowdsourcing）方式，也可以被计算社会语言学用以获取不同人群如何使用某种语言变体，以及不同人群如何看待不同的语言变体的大量数据。总之，计算的视窗一经打开，社会语言学的语料收集和处理方式就别开生面，如虎添翼。

在对语料的计算建模等研究方法方面，目前的计算语言学和自然语言处理，按照语言的结构层次和任务需求，已经形成了下列相对丰富和成熟的技术和系统：（1）词汇层，自动分词、词类标注、命名实体识别等；（2）句法层，自动句法分析、依存关系分析、层次结构和成分关系分析等；（3）语义层，词义消歧、语义角色标注、同义互释、文本蕴涵分析等；（4）篇章层，指代消解、共指消解、篇章结构、话题发现与跟踪等；（5）应用层，文本分类、信息抽取、智能问答、文档摘要、机器翻译等；[⑤]（6）算法模型层，除了传统的支持向量机（Support Vector Machine，SVM）、逻辑回归（Logistic Regression）、朴素贝叶斯（Naive Bayes）等算法，还有 n- 元语法（n-grams）、新兴的潜在变量建模方法（latent variables modeling approaches），以及最近十几年来发展起来的概率图模型（probabilistic graphical models）、神经网络方法中的深度学习（deep learning within a neural network approach）等。这些不同层面上的计算建模方法，可以在研究语言变异与社会变量的对应关系时选择性地使用。当然，生成能力强大的 ChatGPT 等大模型也可以使用。

在研究课题方面，计算社会语言学一方面继承社会语言学的两大主题——社会身份与语言变异的关系和社交环境与语言变异的关系，更加注重利用计算建模的方法来探讨和研究这些问题；另一方面，计算社会学已经开展的词汇的时空传播与演化、语言使用与个体差异、语言使用与社会地位、语言使用与群体分析等专题，[⑥]也可以融入和拓展上述两个方面。有关研究内容和计算方法，下面分三节进行介绍。

四　社会身份与语言变异的计算方法

下面举例说明如何建构与社会身份（social identity）相关的语言变异的计算方法。众所周知，社会语言学的一个重要假设是：说话者用语言来构建他们的社会身份，

语言（特别是其中的变异形式）是说话者用来塑造其身份的工具之一。当计算语言学认识到语言的使用可以揭示其使用者的社会身份后，许多研究就集中于从文本中自动推断作者的有关社会变量。这个任务可以看作一种自动的元数据检测，以期得到关于作者特征的有关信息。随着对社会趋势分析工具的需求的日益增长，人们对这类元数据检测算法的开发和改进也越来越感兴趣。在计算语言学研究中，与种族、社会阶层等群体变量相比，依据性别、年龄和地理位置等个体变量的语言变异受到了更多的关注。

在数据收集方面，早期的研究基于语料库中的正式文本，或者在当面对话或电话交谈等受控环境中收集。随着社交网络媒体的普及，人们从博客、推特、论坛等网络平台收集非正式文本。由于这类数据通常缺乏明确的关于用户的性别、年龄等身份信息，因而研究人员需要使用不同的策略，从用户提供的有限信息、注释或名字上来获得足够多的标签。

以性别建模为例，计算语言学领域的研究者研究过文本作者的自动分类。曾经用支持向量机、逻辑回归、朴素贝叶斯等算法，对作者进行基于生物学特征的二元分类。但是，社会语言学领域的研究表明，这种把性别作为说话者的一种固定属性的做法，忽略了说话者的主观能动性。从社会学角度看，性别是一种社会结构，性别行为是社会习俗的结果，而不是固有的生物学特征。如果联系会话伙伴、互动环境和社交网络，对语言使用中性别的特定模式进行计算研究，可以发现：尽管某些语言特征通常被男性或女性更多地使用（比如，在词类频率方面，男性更多地用介词、冠词，而女性更多地用代词，特别是第一人称代词；在风格方面，男性倾向于用长的词句和文本、更多地用詈辞，而女性更多地用情绪性词语、及“omg”“lol”之类典型的社交媒体词语），但是，个别说话者可能会偏离研究中所强调的惯常印象（例如：男性善于用“报告性”言谈来交换信息，女性喜欢用“亲善性”言谈来建立联系）。有研究发现，在同性别群体交谈时，他 / 她们更多地使用专属于其性别的语言变体。此外，性别因文化和语言的不同而形成不同的形态。这一切，有助于更好地证明：语言（运用）本质上是社会性的，语言的共时变异和历时变化与语言使用者的社会变量直接相关。

另外，怎样发现和分析年龄、地理位置与语言使用的关系？比如，什么年龄层次的人、处于什么场合更加容易偏离规范的标准语？怎样为年龄和位置的变化建模（离散的年龄段还是连续的生命周期，离散的行政区划还是连续的地理坐标）？怎样利用语言使用者在推特等社交媒体上留下的位置信息，或者他们在用户介绍中提供的位置信息？以及怎样对这些维度的调查结果进行解释？比如，年轻人更多地使用单数第一和第二人称代词，而老年人更多地使用复数第一人称代词及介词、定指词与冠词，这种倾向性与语言是否属于代词脱落型语言有没有关系？这些也是从计算角度研究社会结构如何影响语言使用的核心课题。

反过来看，如果研究清楚了性别、年龄和位置等变量决定的说话者的社会身份，如何影响了语言变体的选择，那么，这种成果肯定也可用于帮助改进基于身份信息的内容检测和文本分类等自然语言处理任务。比如，Dadvar 等（2012）训练针对特定性别的分类器来侦测网络霸凌的实例，结果发现，不同性别的侵扰者使用的语言是不同的。再比如，Hovy（2015）发现，训练针对特定性别或年龄的词嵌入向量，可以改善情感评价分析和话题分类等工作。这就走向语言学的社会研究和计算研究的双向对流、互惠互利和协同发展，也显示出计算社会语言学的应用潜力。

五　个体人格与语言变异的计算方法

事实上，对于个体的语言使用和变异来说，比社会身份更加隐蔽和关键的决定因素，可能是人格差异。人格心理学（personality psychology）和社会语言学的相关研究发现，人类个体的人格差异会反映在他们的语言使用特点上。因此，如何定量地建立起语言使用与个体人格差异之间的关联，是心理学、语言学和社会计算的重要课题。这个主题极具代表性的计算建模工作，是 20 世纪 90 年代 Pennebaker & King（1999）提出的“语言探询与词频统计”（Linguistic Inquiry and Word Count，LIWC）方法。这是一种基于词典的词语计数程序（dictionary-based word counting program），其基本思想是：以词汇作为定量分析语言使用的基本单位，首先通过人工收集、标注的方式，建立不同类别（如代词、数词、情感词等）的词语词典；然后在给定的个体或群体相对应的文本中进行词频统计，从而建立起个体差异（即不同人格）与词类比例（即语言使用特点）之间的关联关系。Pennebaker 教授的研究团队已经在这方面做了大量有影响的工作。他们发现：抑郁与自杀者往往会在其文本中发出可侦测的求救信号（Chung & Pennebaker，2007）；初次约会时，对象之间几分钟的对话就可以预测彼此的好感，而情侣间的对话也可以预测几个月后持续交往的概率（Ireland et al.，2011）；团队的凝聚力和合作倾向也可以通过其内部对话做出预测（Gonzales et al.，2010）；谎言的有关语言特性有助于分辨真假（Newman et al.，2003）；对语言使用进行分析，有助于结识新朋友（Pennebaker & King，1999）；语言使用还与年龄有千丝万缕的联系（Pennebaker & Stone，2003），等等。

目前，在网络社交媒体普及的背景下，更凸显出通过语言使用分析个体差异的必要性。一方面，很多在小规模数据集上建立起来的社会理论，需要在大规模真实数据集上进一步验证或再发现；另一方面，利用社会媒体用户产生的文本数据推测用户的人格或心理特点，可以在个性化推荐服务中发挥重要作用。正因如此，近年来，在社会计算领域中，研究人员提出了用户建档（也称为“用户画像”）的研究任务，旨在利用用户产生的内容来预测用户的各种属性，既包括用户的有关简单属性，如性别（Burger et al.，2011；Fink et al.，2012）、年龄（Goswami et al.，2009）和地理位置（Rao et al.，2010；Li et al.，2012）等；也包括用户的有关复杂属性，如兴趣（Yang et al.，2011）、政治倾向（Rao et al.，2010）、性格特点（Mairesse et al.，2007；Schwartz et al.，2013）和主观幸福感（Frank et al.，2013；Mitchell et al.，2013；Dodds et al.，2011），等等。这种研究成功地把语言使用特点与用户其他方面的特征（如用户的社会网络结构、在线行为模式等）综合起来进行有效的属性预测。特别是，在研究手段上超越了词频统计的层面，充分利用了机器学习和自然语言处理领域的新方法，如向量空间模型（Manning et al.，2008）、隐含主题模型（Steyvers & Griffiths，2007）、时间序列分析（Hamilton，1994）等，在定量分析的广度和精度上都向前推进了一大步。这种类型的研究，为计算社会语言学的建设和发展提供了新的领域和研究手段。

六　社交环境与语言变异的计算方法

语言运用往往是在成对、成组和成社群人员构成的社会互动环境中进行的。这给了不同的说话人顺应或塑造社会关系的机会，并响应特定的社交场合和相遇细节（如对话者或听众、话题和说话人的目标等）。这种与社交环境相关的语言变异研究，特别需要计算建模的方法。因为，首先，从数据源的角度看，各种线上社区、论坛、课堂等在线数据中，有大量详细的交互记录，已经推动并促成了计算语言学领域关于这一主题的大量

研究工作。其次，从上述语料中，我们可以通过一定的计算手段提取社会关系，以揭示社会关系的强弱、权力等级、礼貌策略、风格转换等对语言运用的影响。

语言运用不仅是一种信息交流的过程，也是一种表现自我和定位他人、以及反映说话人与会话伙伴的相对地位的社会行为。这种言语行为表现上的一致性，等于定义了会话角色。也就是说，从诸如此类的语言运用中，可以揭示相关说话人之间的若干社会关系。正是认识到了这一点，计算语言学领域已经展开了基于不同类型文本，来自动提取会话者的社会关系及其动态变化的研究，成功地从语言使用上发现了弱关系（比如熟人）和强关系（比如家人或密友）的区别。Bak 等（2012）用自动识别话题的方法，研究推特用户在强弱不同的关系中自我透露（self-disclosure）的差异。他们发现，推特用户面对强关系时会透露更多的个人信息，而面对弱关系时则会显示更多的正面情感评价。这种现象，也许可以用照顾初次相识这种社会规范来解释。其他一些研究，从更广泛的数据集中自动提取社会关系，通过语言使用上的不同表现，来判定线上互动时，发送消息的作者是向上言说（面向较高社会地位）还是向下言说（面向较低社会地位）。还有人用逻辑回归方法来对线上语料库中的权力关系进行自动分类，进而分析所提取出来的社会网络结构。比如，社会语言学调查了说话者如何使用语言来维持和改变权力关系，计算语言学探索了怎样从文本中自动识别权力关系。但是，对于不同言语社区之间的人们的社会互动，迄今的研究仍停留在比较简单的层面上。

关于不同权势的人们之间的语言互动，社会语言学理论曾经提出：地位低的发言者需要从语言上去适应地位高的听者，而地位高的人则不需要调整自己的语言方式去适应别人（Gonzales et al.，2010）。过去由于缺少相关大规模数据，因而有关理论一直缺少定量分析的支持。美国康奈尔大学的 Mizil 教授等人，选取线上和线下两个场景，验证了语言交流行为是如何体现权力关系的。两个场景分别是维基百科中编辑们的在线讨论和法院庭审现场的辩护对话。值得注意的是，这里所谓的语言使用情况，指的是虚词的使用，而不是实词的使用。他们调查了包括冠词、助动词、连词、高频副词、（非）人称代词、介词和量化词等 8 种标记，一共 451 个词项的使用情况。研究者观察了由甲引起的对话中，乙分别用了多少种标记来回应；并且考察了甲分别用了多少种标记，可能引起乙使用多少种标记来回应。值得注意的是，这种不同权势的对话者对虚词的不同的使用及其调整变化，甚至可能连对话者自己都没有注意到。然后，他们通过统计和定量分析及形式化刻画，验证了对话者之间权力的差异，会在两人如何回应对方的语言方式上有所体现（Mizil et al.，2012）。这种结论，也在推特平台上得到了验证。首先，他们利用介词等虚词的使用情况，考察了双方的语言风格是如何彼此适应的。然后，他们考察了交流双方之间影响的不对称性，以及这种不对称性与社会地位的关系，即，地位高的人不会去适应地位低的人，而地位低的人要付出更多去适应地位高的人。研究结果表明，虽然推特对交流增加了一些限制（非面对面，非实时，而且只能说 140 个词），但交流中仍然有比较明显的语言适应行为（Mizil et al.，2011）。

一般认为，社会交往中的礼貌行为，有助于维持社会和谐、避免社会冲突。Brown & Levinson（1987）发现，语言的礼貌行为受到下列三个社会因素的影响：社会距离（social distance）、相对的权力（relative power）、诉求的麻烦程度（ranking of the imposition, i.e., cost of the request，即请求的成本或代价⑦）。幸运的是，检测礼貌的自动分类器已经被开发出来，可用于大规模的礼貌策略研究。鉴于礼貌用语的使用与参与对话的人的社会地位之间具有密切的关系，Mizil 团队（2013a）分别对维基百科编辑和 Stack Exchange

论坛的讨论者进行了研究。他们把用户对他人提出请求时的对话摘录出来（其中，一句是真正的请求，而另一句是客套话），然后由标注者对其礼貌程度进行评价。研究结果表明，维基百科编辑在选举过程中试图获得更高地位时，会更加礼貌；而一旦选上以后，礼貌程度就会有所下降。这种情况，同样也出现在 Stack Exchange 上，明显地，人们的礼貌程度跟其地位呈反比关系。

最后，对社区语言使用情况的动态研究也是计算社会语言学的主题。因为，人们会根据谈话对象调整他们的语言使用。在社区内，一些规范会通过成员之间的互动而出现，例如，使用俚语和特定领域的行话，或者在推特上表示转发的约定。对于这一主题的早期调查，主要基于非公共社区的数据。最近的研究则使用了来自公共在线社区的数据，比如在线论坛和评论网站。这一方向的研究，显示了利用大量在线数据定量研究社区语言变化的潜力。当然，在这种分析中，应该仔细考虑数据中的偏差，特别是当数据的动态和内容没有被完全理解时。比如，据 Mizil 等（2013b）介绍，他们以两个大型啤酒评论社区作为研究对象，发现用户在社区中一般会经历两个阶段：在第一个阶段，他们刚进入社区，会积极学习适应社区的语言使用规则；而接下来，他们逐渐不再做出改变，任由规则变化；最后，逐渐退出社区主流群体。这项研究定量地探索了在社区与个人的相互作用下，语言使用规则变化的复杂性。可见，Mizil 等人的一系列研究，开创性地在社会媒体大数据上定量验证了社会语言学中的重要理论，并进一步利用该理论展开社会计算研究，为计算社会语言学研究树立了研究典范。

七　结语：基于计算建模的语言的社会变异研究

新文科建设的出发点是多种相关学科的交叉、融合与创新，通过在研究内容、研究方法和技术手段等方面的跨学科的交融、提炼与整合，形成新的学科方向与研究范式，以及相应的复合型人才培养模式。这种宏大的目标迫切需要明确的可依循的操作路径。有鉴于此，本文提出一种“在旧楼上加盖新楼层”的方法：尽可能利用已经发展起来的若干第一代交叉学科，融合贯通起来形成第二代交叉学科，以取得物尽其用、事半功倍的效果。就语言学而言，面对“社交网络—语言的社会变异—计算建模”这种头绪繁多的研究领域，和相应的“大数据与计算科学—社会学—语言学”等多种学科和知识体系，我们可以借助已经成熟的“计算语言学”与“社会语言学”等第一代交叉学科，形成“计算社会语言学”这种第二代交叉学科。因为，社会语言学关注在社会环境中使用的语言的社会维度，计算语言学关注在社交网络上使用的语言的信息维度，把二者结合起来，形成计算社会语言学这一新的研究领域和学科，可以整合这两个学科的优势和强项，更好地从计算的视角来研究语言（变异）和社会（变量）之间的关系，以便更加深刻地认识人类语言在社会环境中的运行机制，更加充分地为计算机处理人类的语言提供理论、方法和材料支持。我们希望这种操作路径可以帮助细化与落实新文科建设，并且为新文科建设提供系统的学科规训。

本文的创新点是从当代社会的“物理—精神—信息”三元空间这一特征切入，说明社会语言学的研究迫切需要计算建模这一方法。文章还以社会身份、人格特点和社交互动等对语言使用和变异选择的影响因素为例，说明了计算社会语言学研究的理论依据、语料采集与计算建模方法。

讨论至此，计算社会语言学的核心要务也就可以粗略地总结为：用计算建模的方法研究语言的社会变异，并为语言的计算建模而研究语言的社会变异。

［附　注］

① 详见《语言战略研究》微信公众号“观约谈”栏目 2018 年 12 月 19 日推出的《计算社会语言学》，https://mp.weixin.qq.com/s/saVHtd_QKvQTr9gs2_Pzdw。

② 详见汉语堂微信公众号 2021 年 12 月 11 日推出的刘知远文章《语言分析技术在社会计算中的应用》，https://mp.weixin.qq.com/s/NzvcP4eFlymwIsIuQqgBQA。

③ “学科规训”通俗地说是一门学科的规矩，重点是一门学科的标志性的规范化训练；具体指一种学科的特征、特点，学科人的信条、行事风格或价值取向，知识规划和训练的方法，学科知识与学科成员的准入门槛，等等。（黄维、胡弼成）2020。

④ 在信息技术上，“三元空间”指由信息空间、物理世界和人类社会所构成的复合性数据来源地（朱文武、王鑫）2021。本文则强调人类社会本身处于三元的“物理世界—精神世界—网络世界”之中。

⑤ 这些方法详见 Manning & Schütze（1999）和 Jurafsky 等（2000）。

⑥ 关于这四个专题研究的综述，详见汉语堂微信公众号 2021 年 12 月 11 日推出的刘知远文章《语言分析技术在社会计算中的应用》。

⑦ 这相当于我们日常所说的“人际交往中所欠的人情债的大小”。

［参考文献］

［1］黄　维，胡弼成 . 论学科规训对人才培养产生的隔壑作用［EB/OL］. 中国高校人文社会科学信息网：https://www.sinoss.net/uploadfile/2020/0403/20200403080404102.pdf，2020.

［2］朱文武，王　鑫 . 三元空间大数据网络关联表征［J］. 中国科学：信息科学，2021，（11）.

［3］Bak, JinYeong, Suin Kim, & Alice Oh. Self-disclosure and relationship strength in Twitter conversations［A］. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*［C］. 2012.

［4］Brown, Penelope & Stephen C. Levinson. *Politeness: Some Universals in Language Usage, volume 4 of Studies in Interactional Sociolinguistics*［M］. Cambridge：Cambridge University Press，1987.

［5］Burger, J. D., J. Henderson, G. Kim, et al. Discriminating Gender on Twitter［A］. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*［C］. 2011.

［6］Chung, C. & J. W. Pennebaker. The psychological functions of function words［A］. *Social Communication*［C］. New York：Psychology Press，2007.

［7］Dadvar, Maral, Franciska M. G. de Jong, Roeland Ordelman, & Dolf Trieschnigg. Improved cyberbullying detection using gender information［A］. *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*［C］. 2012.

［8］Danescu-Niculescu-Mizil, Cristian, M. & Gamon, S. Dumais. Mark my words! Linguistic style accommodation in social media［A］. *Proceedings of the 20th International Conference on World Wide Web*［C］. 2011.

［9］Danescu-Niculescu-Mizil, Cristian, L. Lee, B. & Pang, J. Kleinberg. Echoes of power: Language effects and power differences in social interaction［A］. *Proceedings of the 21th International Conference on World Wide Web*［C］. 2012.

［10］Danescu-Niculescu-Mizil, Cristian, M. Sudhof, D. Jurafsky, J. & Leskovec, C. Potts. A computational approach to politeness with application to social factors［A］. *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*［C］. 2013a.

［11］Danescu-Niculescu-Mizil, Cristian, R. West, D. Jurafsky, J. Leskovec, C. Potts. No country for old members: User lifecycle and linguistic change in online communities［A］. *Proceedings of the 22nd International Conference on World Wide Web*［C］. 2013b.

［12］Dodds, P. S., K. D. Harris, I. M. Kloumann, et al. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter［J］. *PloS One*, 2011,（12）.

［13］Fink, C, J . Kopecky, M. Morawski. Inferring Gender from The Content of Tweets: A Region Specific Example［A］. *Proceedings of the International AAAI Conference on Web and Social Media*［C］, 2012.

［14］Frank, M. R., L. Mitchell, P. S. Dodds, et al. Happiness and the patterns of life: a study of geolocated tweets［J］. *Scientific reports*, 2013,（1）.

［15］Gonzales, A. L., J. T. Hancock, J. W. Pennebaker. Language style matching as a predictor of social dynamics in small groups［J］. *Communication Research*, 2010,（1）.

［16］Goswami, S., S. Sarkar, M. Rustagi. Stylometric Analysis of Bloggers' Age and Gender［A］. *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*［C］. 2009.

［17］Hamilton, D. James. *Time Series Analysis*［M］. Princeton：Princeton University Press, 1994.

［18］Hovy, Dirk. Demographic factors improve classification performance［A］. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1：Long Papers)*［C］. 2015.

［19］Ireland, M. E., R. B. Slatcher, P. W. Eastwick, et al. Language style matching predicts relationship initiation and stability［J］. *Psychological Science*, 2011,（1）.

［20］Jurafsky, D., J. H. Martin, A. Kehler, et al. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*［M］. Upper Saddle River, NJ：Prentice Hall PTR, 2000.

［21］Labov, William. *The Social Stratification of English in New York City*［M］. Washington, DC：Center for Applied Linguistics, 1966.

［22］Labov, William. *Sociolinguistic Patterns*［M］. Philadelphia：University of Pennsylvania Press, 1972.

［23］Lazer, D., A. Pentland, L. Adamic, et al. Computational Social Science［J］. *Science*, 2009,（5915）.

［24］Li, R., S. Wang, H. Deng, et al. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations［A］. *Proceedings of the 18th ACM SIGKDD International Workshop on Urban Computing*, 2012.

［25］Mairesse, François, Marilyn A. Walker, Matthias R. Mehl, Roger K. Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text［J］. *Journal of Artificial Intelligence Research*, 2007,（30）.

［26］Manning, D. Christopher and Hinrich Schütze. *Foundations of statistical natural language processing*［M］. Cambridge, MA：MIT Press, 1999.

［27］Manning, D. Christopher, Prabhakar Raghavan and Hinrich Schütze. *Introduction to Information Retrieval*［M］. Cambridge：Cambridge University Press, 2008.

［28］Mitchell, L., M. R. Frank, K. D. Harris, et al. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place［J］. *PLoS One*, 2013,（5）.

［29］Newman, M. L., J. W. Pennebaker, D. S. Berry, et al. Lying words: Predicting deception from linguistic styles［J］. *Personality and Social Psychology Bulletin*, 2003,（5）.

［30］Nguyen, Dong, A. Seza Doğruöz, Carolyn P . Rosé, Franciska de Jong. Computational sociolinguistics:

A survey[J]. *Computational Linguistics*, 2016,(3).

[31] Pennebaker, J. W. & L. A. King. Linguistic styles: Language use as an individual difference[J]. *Journal of Personality and Social Psychology*, 1999,(6).

[32] Pennebaker J. W. & L. D. Stone. Words of wisdom: language use over the life span[J]. *Journal of Personality and Social Psychology*, 2003,(2).

[33] Rao, D., D. Yarowsky, A. Shreevats, et al. Classifying latent user attributes in Twitter[A]. *Proceedings of The 2nd International Workshop on Search and Mining User-Generated Contents*, 2010.

[34] Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, et al. Personality, gender, and age in the language of social media: The Open-vocabulary approach[J]. *PLoS One*, 2013,(9).

[35] Steyvers, M. & T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*[A]. *Latent Semantic Analysis: A Road to Meaning. Mahwah*[C]. NJ: Laurence Erlbaum Associates Publishers, 2007.

[36] Weinreich, Uriel, William Labov, Marvin I. Herzog. Empirical foundations for a theory of language change[A]. *Directions for Historical Linguistics: A Symposium*[C]. Austin: University of Texas Press, 1968.

[37] Williams, B. Christine & Girish Gulati. Social networks in political campaigns: Facebook and congressional elections of 2006 and 2008[J]. *New Media Society*, 2013,(1).

[38] Yang, S., B. Long, A. Smola, et al. Like like alike: Joint friendship and interest propagation in social networks[A]. *Proceedings of the 20th International Conference on World Wide Web*[C]. 2011.

(责任编辑 常文斐)

语言国情调查方案学术研讨会成功举办

2024 年 1 月 16 日，语言国情调查方案学术研讨会在教育部语言文字应用研究所成功举办。

本次会议汇集语用所承担的国家社会科学基金重大项目“‘两个一百年’背景下的语言国情调查与语言规划研究”和“少数民族地区国家通用语言推广普及策略研究”、国家语委重点科研项目“新时代语言国情调查方案研究”三个课题组的相关研究人员，以及语用所参与多项调查任务的科研人员，集中研讨了调查方案整体设计、推进落实试点调查计划、注重多类型数据的搜集和分析、加强各课题研究内容的统筹配合、提高研究成果的整体性等问题。

北京师范大学珠海校区黄行教授报告了制定语言国情调查方案的整体设想、任务、目标和重点举措。上海外国语大学赵蓉晖教授、北京语言大学王莉宁研究员、中国社会科学院贾媛研究员、教育部语言文字应用研究所刘丹丹分别报告了外语调查、汉语方言调查、民族语言调查及语言国情综合性调查方案的设计思路与推进计划。北京师范大学张秋玲教授、南京财经大学白先春教授、南京特殊教育师范学院孙计领副教授等提出了意见建议。

王敏副所长感谢各位课题组成员的辛勤付出，着重阐释了科学合理的语言国情调查方案对于提高语言政策规划的重要意义和价值，并对下一步工作推进提出了要求和建议。

（“‘两个一百年’背景下的语言国情调查与语言规划研究”课题组
“少数民族地区国家通用语言推广普及策略研究”课题组
“新时代语言国情调查方案研究”课题组）