

大数据时代语言研究的方法和趋向

刘海涛^{1, 2} 林燕妮¹

(1. 浙江大学 外语学院, 浙江 杭州 310058;

2. 广东外语外贸大学 外国语言学及应用语言学研究, 广东 广州 510420)

摘要: 文本围绕大数据时代的语言研究这一主题展开探讨。首先从信息时代背景下语言学家的角色谈起, 阐释当今时代语言研究的变化, 强调语言材料的真实性对发现语言规律至关重要, 介绍大数据为语言研究带来的新契机, 并论述“语言是由人驱动的复杂适应系统”的观点。其后, 从科学哲学的角度阐明采用科学方法研究语言的必要性, 并讨论数据密集型语言研究范式及问题。之后以团队的研究成果为例, 介绍基于数据的语言研究具体如何开展。最后阐述基于数据的方法在语言学的学科建设与发展中所起的作用。

关键词: 语言学; 大数据; 数据密集型研究方法; 科学研究范式

中图分类号: H1-0

文献标识码: A

文章编号: 1005-9245(2018)01-0072-12

引言

自20世纪下半叶起, 人类社会从工业时代逐步迈入信息时代。随着信息化浪潮席卷全球, 信息爆炸问题日渐凸显。人类历史上从未遇到过这么多的信息, 人类几乎生活在一个被信息所包围的世界里。对处理海量信息和知识的迫切需求, 促使人们思考如何使用计算机帮助人类完成一些繁杂的工作或解决一些问题, 例如抽取信息、自动翻译等, 使人们可以集中精力做更重要的事情。因此, 计算语言学和自然语言处理领域应运而生, 并呈现出了蓬勃发展的态势。

然而, 正是在这一颇具发展潜力的语言应用领域中, 却时而能听到对于语言学家的质疑声。例如, 美国工程院院士、自然语言处理专家弗雷德里克·杰利内克(Frederick Jelinek)据称曾说过这样

一句话:“每当我解雇一位语言学家, 系统的性能就会改善一些。”^{[1][2]}(有关这句话的来龙去脉, 可参考 https://en.wikipedia.org/wiki/Frederick_Jelinek#cite_note-6。)当中也许有些玩笑的成分, 我们却无法忽视一个事实: 目前在主流的计算语言学和自然语言处理领域中, 几乎很难见到语言学家的身影。理应来讲, 这些应用性领域主要的处理对象是语言, 而作为研究语言的基础学科, 语言学本应能够为语言实践与应用提供一些帮助和指导, 这个时代本应是语言学家大展宏图的时代, 但现实却对语言学家如此残酷, 究竟有何原因? 当代语言学家的作用该如何体现? 这成为触发我们反思语言研究及其与当今信息时代之间关系的起因。伴随着信息化进程不断推进, 近年来, 以规模性(Volume)、多样性(Variety)、高速性(Velocity)和价值性(Value)的“4V”特征^[3]著称的“大数据”(Big Data), 开始改变人类的社会生活和思维方式, 并形成了新的研究范式^[4],

收稿日期: 2017-04-10

基金项目: 本文系国家社科基金重大项目“现代汉语计量语言学研究”(11&ZD188)、中央高校基本科研业务费专项资金资助(浙江大学大数据+语言规律与认知创新团队)的阶段性成果。

作者简介: 刘海涛, 浙江大学求是特聘教授, 广东外语外贸大学“云山领军学者”, 国际世界语研究院院士; 林燕妮, 浙江大学外语学院博士研究生。

在自然科学和人文社会科学领域均有不少新发现。由此可见，信息时代对语言研究提出了挑战，同时也带来了新的机遇。

本文以大数据时代的语言研究为主题，尝试就如下几方面问题进行探讨：信息时代背景下，语言研究产生了怎样的变化？基于数据的方法能否为语言研究带来新的思路？作为建立在数据基础上的语言学分支——计量语言学持有怎样的语言观？其研究范式如何体现出科学性？采用数据密集型方法的语言研究具体如何开展？当今“双一流”建设背景下，该方法又能对语言学的学科建设与发展起到什么作用？

文章余下内容的组织结构如下：第一部分阐述信息时代语言研究的变化；第二部分讨论数据密集型语言研究方法及其问题；第三部分介绍几项基于数据的语言研究；第四部分给出关于学科研究与发展的一些思考；最后部分为余论。

一、信息时代语言研究的变化

本部分阐述信息时代语言研究的变化。首先结合一位世界著名语言学家从“花园”走向“灌木丛”的学术经历，强调当代的语言研究必须注重语言材料的真实性，并突破以往研究方法的局限；其后指出在大数据时代，语言研究将获得新的发展契机；最后介绍基于数据的语言分支——计量语言学，阐释其定义及语言观。

（一）语言研究的转变：从“花园”到“灌木丛”

2016年8月，词汇功能语法（Lexical Functional Grammar）^[5]的提出者琼·布里斯南（Joan Bresnan）获得了计算语言学学会（Association for Computational Linguistics）授予的终身成就奖。布里斯南的获奖感言后来发表在《计算语言学》（Computational Linguistics）2016年第4期上，题为《语言学：花园与灌木丛》^[6]。文章中，布里斯南回忆了自己从语言学的“花园”走向“灌木丛”的经历。她认为，目前大多数传统意义上的语言学理论，与现实社会所需要的语言学理论存在着本质的区别。包括生成语法在内的传统语言学属于“花园里的语言学”，主要分析语言学家依靠精挑细选或内省得出的语言现象，并通过句法树、短语等符号来进行定性概括。而“灌木丛中的语言学”或“野地里的语言学”研究的是人们日常交流所使用的真实语言，通常借助条件概率、信息量等来进行定量分析。当面对的不

再是花园里那些整整齐齐、完美精致的花儿，而是大片杂草纷乱的野生灌木丛时，花园里用的那一套工具与方法就极有可能失效。

布里斯南是乔姆斯基的博士生，她在文中还回忆了自己20世纪60年代在麻省理工学院跟随乔姆斯基读博士的情况。那个时期，整个世界都为乔姆斯基的想法所吸引。语言被视为符号模式所组成的集合，通过采用符号逻辑公式，分析人类语言结构，探索人类的语言与心智——这当然是非常激动人心的。当时被这个想法所鼓舞的人很多。其中有一位工科博士，比她在麻省理工入学早几年，甚至一度打算从他攻读的信息论专业转到语言学。但由于他导师不同意，他只好把信息论的博士读完^[7]。这个人正是后来说要“解雇语言学家”的杰利内克。这不禁令人疑惑：语言学发展的几十年间，是什么使得像杰利内克这样一位热衷于理论（形式）语言学的热血青年，变成一个“解雇语言学家”的冷面老板？最大的问题可能出在主流语言学的研究材料和方法上。如上所述，自然语言处理需要面对真实的、多样化的语言，如同在大千世界里自然生长的灌木丛。如果像栽培花园里的花朵一样，只用几个精选好的句子，可能难以发现真实语言的规律。

无论是传统语言学还是现代语言学，研究的对象都是人类语言。不管语言学家是否准备好了，信息时代都已来临。信息的主要载体之一是语言，信息时代的语言研究可能要同时考虑人和计算机的需要，这是一种信息时代的语言观。自然语言处理所面对的是真实的语言材料，真实语言最显著的特点是概率性，即，语言的合法性介于可能与不可能之间，具有梯度性，而不是非此即彼的简单二分。科学研究一般均涉及抽象建模的过程。模型的特征对应的是研究对象可观察的属性。理论并不能直接解释现实世界本身，而是要通过抽象之后的模型以及它所对应的现实来进行解释。因此，理论的预测能力取决于模型和现实之间的对应关系。如果在建模的过程中忽略了研究对象最本质的特性，没有反映其真实面貌，那么通过这一模型发现的成果最后就很难被别人使用。这可能是绝大多数语言学家被计算语言学所抛弃的重要原因之一。当然，我们不能仅以此例来评价语言学存在的意义与价值。但布里斯南从“花园”走向“灌木丛”的经历，说明信息时代的语言研究可能正面临着重要转变。

毋庸置疑，20世纪50年代起，乔姆斯基所倡导的语言形式化方法与理论为我们带来了一场语言

学革命。然而，这几十年语言研究的理论与实践均表明，语言研究可能还需要一些新的转变。具体而言，第一，在研究对象上，应更多地关注真实的语言材料，关注人与语言系统的关系；第二，在研究方法上，需要根据真实语言材料的特点，采用先进的技术手段与研究方法，以此来弥补内省法或定性手段的不足；第三，在模型选择上，更需要关注模型的跨语言有效性，而不囿于某种特定的语言，因为语言学研究的是人类的语言，语言学家所发现的规律更多的应该是人类语言的普遍规律。否则，语言研究者可能会离这个时代越来越远。

（二）大数据时代为语言研究带来新机遇

信息时代在给当今的语言研究带来挑战的同时，也为实现上述转变提供了新的契机。前文提到的转变，实质上更多的是方法的转变，即从内省方法到数据驱动方法的转变。数据驱动意味着语言研究也可以具有或应该适应信息时代的另一个特征，也就是我们今天常听到的“大数据”。虽然“大数据”这个提法不太严谨，因为“大数据”除了规模大之外，还具有种类多、处理速度快、价值密度低等特点^[3]。但无论是“大数据”还是最近提的“厚数据”，说的都是我们正处在一个数据唾手可得的时代。对于语言学家而言，我们应该更看重“数据”这个时代特征，更关心数据驱动的语言研究路向，而不只是数据的多少。换言之，我们更应该关心的是能用数据来解决哪些语言学问题，或者能发现那些过去我们注意不到或无法研究的语言规律。从这个意义上说，数据为我们提供的是一种研究范式、一种观察研究对象的方法和工具。

首先，基于数据的方法为我们提供了感知研究对象的量化维度，令我们对研究对象有一个更清晰、更精确、更细微的认识。宛如从不同的距离和视角观察同一个事物，从宏观到微观，随着观测距离的推近与拉远，所看到的世界以及给人们带来的体验会很不一样。有了更多的真实语言材料，有助于更深入而真实地反映语言的概貌。基于数据的方法能反映语言的一些本质特征，其中一个特征是语言的概率性^[8]。例如，在以内省法为研究手段的语言学中，打星号（“*”）标记的句子，按母语者的语感是不符合语法或不能接受的。然而在日常生活中，这些打了星号的句子实际上有相当一部分人在使用。大量研究表明，人们理解或产出的语言，按照规定性语法，并不是“能接受”与“不能接受”的绝对二分，而是介于两者之间。假如有大量语言

数据的支撑，那么在很难描述某种说法的合理性时，也就便于更细致地区分语法上可接受的程度。数据手段有助于更好地反映语言的真实状态和本质特点，正如伯纳德·科姆里（Bernard Comrie）在《语言共性和语言类型》前言中的最末一句话：“语言学研究语言，而语言是民众实际所讲的语言。”^[9]

此外，数据能更好地帮助我们研究人类的语言规律和认知规律之间的关系。我们知道，语言是一个符号系统。而以往的很多研究把人与语言分离开来，只做纯粹的形式符号分析。但实际上，语言是由人驱动的符号系统，或更精确地讲，是一种人驱复杂适应系统。语言的结构模式和演化规律均受到生理、心理、认知等内部因素，以及自然、社会等外部因素的综合影响。其中，内部因素的普遍性决定了语言的共性，外部因素的差异造就了语言的多样性。一方面，认知普遍性在一定程度上决定了语言的普遍性。例如，递归被认为是人类语言最本质的特点^[10]，但实际上递归并非是无穷的，三层以上的递归现象在实际使用中很少出现^{[11][12]}。人不能完全等同于机器，人是受到认知因素约束的。另一方面，人生活在一定的自然环境和社会环境中，这些自然、社会、文化等因素可能会对语言有所影响，从而形成了世界上多种多样的语言。因此，从大量来自于真实语言运用的数据出发，有助于我们更好地发现或解释人类语言的普遍性和多样性。

（三）离不开数据支持的语言学分支——计量语言学

杰利内克后来曾发文指出，语言学家研究语言现象与物理学家研究物理现象十分相似。因此，如果工程师能从物理学家的真知灼见中获益，自然语言处理的研究者也应该从语言学家处汲取营养^[2]。换言之，物理学家发现的是物理世界的规律，而语言学家应该研究的是语言的结构以及演化规律。既然如此，为什么主流的语言学研究成果难以应用于自然语言处理实践？除了以上提及的研究资源与方法的问题外，研究的精确性与科学性也是一个值得注意的问题。如果从采用科学的方法发现语言系统规律的角度看，计量语言学是一个值得倡导的语言学分支。

计量语言学采用定量的方法，对各种语言现象、语言结构、结构属性以及他们之间的相互关系进行定量分析和动态描写，以揭示各种语言现象的关系、地位、规律和总体面貌，探索语言系统的自适应机制和语言演化的动因，力图提高语言研究的精确化

和科学化^[13]。

那么，计量语言学与本体语言学有什么联系与区别？首先，作为语言学的一个分支，计量语言学和语言学的其他分支一样，均以探索语言规律为目标。而计量语言学在语言观、语言材料、研究方法和抽象程度方面，与本体语言学是有差别的。很多情况下，本体语言学是由与某种语言现象有关的具体问题驱动的，主要通过具体例子或用法，借助语感来进行分析，试图通过内省法，并或多或少借助形式化的手段探求语言结构的规律，以研究大脑的语言处理机制。计量语言学从系统的角度，把语言看作一个复杂适应系统，使用真实的语言材料，以定量方法为主、采用数学的手段来探求语言结构和演化规律。总之，它具有精确、真实、动态的特点。值得注意的是，计量语言学与大部分的本体语言学相比，在研究对象的抽象程度方面是有差异的。计量语言学希望通过建立语言系统的模型，在更抽象的层面上探讨语言系统及其运作规律。虽然采用的是真实文本，但是很少涉及其中具体的字词句。从规律的发现与呈现方式看，计量语言学追求的语言规律更接近于物理学家发现的有关物理世界的规律。

当然，从具体的语言结构出发也是很有趣的，两种视角很难说孰优孰劣，两者都以探讨语言规律为目标，只是方法不同。人类语言是一个非常复杂的动态系统，为了探求系统的运作机理和演化规律，我们可能需要同时采用不同的方法、结合各自的优势，来对语言系统进行全方位的探索，从而对人类语言系统有一个更全面、更完整的认识。

（四）计量语言学的语言观：语言是复杂适应系统

计量语言学将语言视为一种复杂适应系统，这种语言观与以往对语言的看法是有所不同的。语言作为一种符号系统的观点，很早以前就由以索绪尔为代表的语言学家提出来。遗憾的是，长期以来，语言甚至被认为是一种可脱离于人而存在的符号系统。1995年，霍兰在《隐秩序》^[14]一书中提出了复杂适应系统理论。该理论的核心思想是：个体的适应性导致了系统的复杂性。在这一思潮的影响下，遗传算法、神经网络、演化博弈论等复杂网络方法逐渐被引入对社会系统的研究之中^[15]。近年来，一些语言学家基于语言事实，提出语言也是一种复杂适应系统^{[16][17][18]}。

按照系统科学的定义，“系统”是指组分及其

之间的关系所构成的整体。哲学认为运动是绝对的。现实系统总是不可避免地要承受来自环境或系统自身的各种扰动^[19]。因此，正常情况下的系统是动态的，为了一个共同的功能目标而运作。如果语言是一种系统，那么它应该具备系统的共性。作为一种动态系统，其运作的主要目标是为了完成作为人类交流工具这一主要功能。当然，语言还有其他的功能，如文化的容器、身份的象征等。为了实现交际最优化，语言系统的各个组分，受省力原则的支配，需要在词汇、句法和语义等层面上协同起来，以共同完成这个目标。然而，过去很多研究却把语言看成了静态系统。系统是动态还是静态，研究起来存在着本质差异。

“复杂”主要指系统的整体行为不等于组分行为之和，即具有涌现性。对于语言系统，以一个由五个词构成的句子为例，把这五个词简单加起来，有时候并不一定能得到整个句子的意思，即存在部分之和不等于整体的情况，这也是现实世界中复杂系统的主要特点之一^[20]。除此之外，复杂系统还具有不确定性、非决定性、随机性等特征^[21]。从某种意义上讲，复杂总是与不确定性或概率相关。

“适应”针对的是有目标限定的动态系统。语言系统具有适应性。所谓“适应”是指在一定的外界环境下，系统通过自组织过程适应环境而出现新的结构、状态或功能^[19]。适应系统具有一套自我调节机制，以维持自身的平衡，语言也是如此。以语言的词汇系统为例，我们从词汇系统中抽象出词的一些属性，包括词的频度、长度、多义度、与其他词的结合能力等，这些属性之间密切相关。在一个平衡的词汇系统中，一个使用频度高的词，长度通常比较小——注意这不是绝对的，而是统计规律，之前也提到过，语言是自然状态下呈现出千姿百态的“灌木丛”。如果一个之前不太常用的词，使用频率突然增加了，那么其词汇协同子系统会作出反应，即这个词会自动地、暂时性地缩短长度，以满足交际的需求，这是系统适应能力的具体表现。

作为一种复杂适应系统，语言与人是共同演化的。前文提到，语言是由人驱动的系统。语言系统处在不断变化发展中，人也处在不断变化发展中。语言系统的发展变化由人这个使用者带动，来自人内部（生理、心理等）和外部（自然、社会等）两方面的因素影响了语言的普遍性和多样性，所以我们不能撇开人的因素来孤立地看待语言现象。

如果语言是一个系统，那么按照研究一般系统

的方法来研究语言，是一个很自然的思路。从系统论的角度研究语言，需要通过对各种语言现象进行细致的观察，对语言系统的组分、结构、过程、行为、功能和环境等方面展开研究，这些同样需要来自真实语言材料或语言行为试验的数据作为支撑。

二、数据密集型语言研究方法和问题

从以上有关计量语言学的定义和语言观的讨论里，不难体会到计量语言学追求精确的特点，这与语言作为一门科学的题中之义相契合。在这一部分，我们将从科学哲学角度阐释计量语言学的研究范式，并就语言研究中与数据相关的几个问题展开讨论。

（一）采用科学研究方法的必要性

从规律发现的角度看，采用科学的研究范式研究语言是十分有必要的。哲学的分支之一——科学哲学对于什么是科学、什么是理论、什么是科学研究范式等问题有专门的阐释。当今科学界认为，科学研究必须采用科学的方法进行。“语言学是一门科学”的理念虽然被大多数语言学家所接受，但长期以来，语言学在科学界却难以获得普遍认可，其中原因与语言研究者在多大程度上认同并遵循科学研究范式不无关系。如果我们认为语言学是科学，却又不遵循科学的方法开展研究，在逻辑上恐怕难以成立。

当然，即便是按照传统方法完全不使用任何数据也并无不妥，因为每一个严肃地做自己的研究的人都应该受到尊重。只是现在我们手上掌握了大量数据以及操作数据的新方法，不去使用总感觉有些可惜。更重要的是，数据或许真能帮助我们获得一些新发现。举个生活中的例子，在摄影的时候，使用长焦、标准、广角、鱼眼等不同的镜头拍同样的景物，拍出的照片给人的感觉会不一样。当分别用显微镜和望远镜去看同一个事物，所见所感也是很不一样的。没有体验过的人可能很难想象由此带来的灵感和启发。那么，是不是当我们掌握了更多的数据，对语言的感受和认识可能会不一样？过去我们没有类似显微镜和望远镜这样的工具，现在触手可及，又何乐而不为呢？

（二）计量语言学的研究范式

前文指出，计量语言学采用的是数据密集型研究范式，具有精确、真实、动态的特点。其中，“精确”是指采用数理手段对语言进行定量描写；“真

实”是指使用日常交际所使用的真实的语言材料；“动态”是指把语言视为一个变化着的复杂适应系统。因此，计量语言学采用的是接近于自然科学的方法。

用定量的方法来研究语言历史悠久，但长期以来没有形成一个系统的学科。20世纪60年代，德国学者加百利·阿尔特曼（Gabriel Altmann）开始系统地研究语言学和科学哲学的关系。他在分析了大量实例后，完全按照科学哲学的方法，制定了一套比较详细的方案，构拟出现代计量语言学的理论架构。在研究范式上，阿尔特曼总结了计量语言学的研究范式，给出了五个基本研究步骤：1. 提出与实证相关的并可以进行检验的假设；2. 用统计的语言来表达这些假设；3. 寻求合适的统计方法对假设进行统计检验；4. 根据统计检验的结果，决定能否拒绝假设；5. 解释假设。计量语言学的这一研究范式，就是当今我们所理解的符合科学哲学意义的研究范式。美国学者大卫·爱丁顿（David Eddington）曾经写过一篇文章，就叫作《语言学与科学方法》^[22]。文中写到，如果要对真实的语言作出有效的解释，必须采用科学的方法。而且从一定意义上来说，语言学的进步取决于研究者在多大程度上采用了这种科学家所公认的、标准的科学研究方法：观察现象、提出假设、收集数据、验证假设、得出结论——也就是今天我们所说的实证研究方法。

在这个时代开展基于数据的语言研究，首先要考虑有哪些问题是需要数据的，或是否有需要数据去解决的问题。通常会遇到两种情况：一种情况是假设驱动，即按照科学研究范式，先提出假设，然后收集数据、验证假设并得出结论；另一种情况是数据驱动，即尽管暂时还没有假设，但先掌握了大量的数据，然后分析这些数据所展现出来的模式，发现并解释其中的规律。验证假设也是需要数据的。尽管内省法是日前主流语言学家的选择，但如果我们也可以科学家公认的方法来验证假设，弥补内省法的不足，得到的结论也许会更令人信服。

关于科学研究范式，李国杰院士在为《可视化未来》撰写的序言中曾这样写道：“数据密集型科学研究已经上升到与科学实验、理论分析、计算模拟并列的科学研究‘第四范式’……大数据对科学的变革意义，与伽利略首次将望远镜指向太空对天文学的意义一样重大。”^[23]迄今为止，科学家们采用数据密集型范式开展研究，在诸多领域已经有了很多有趣的发现^[24]。

（三）数据密集型研究方法的几个问题

在介绍数据密集型研究范式之后，以下围绕与该方法相关的几个问题展开讨论。

1. 大数据时代量化研究方法的特点

用量化的方法研究语言的历史并不短暂。以往的语言定量研究也是以发现语言规律为目标的。受技术等条件约束，依靠传统的卡片式等收集方法所获得的语例比较有限。但是在今天，只要打开计算机联网，语言材料随手可得。从数据规模上看，全世界几十亿人，几乎每个人每天都在说话，真要全部收集起来，数据量必然非常大。海量数据及操作技术为我们这个时代的语言学家提供了更有利的条件，这有助于反映不同场景下的语言样貌，加深我们对语言的了解和认识。数据大当然有大的好处，但并不是越大越好。当语料库达到一定的规模后，其发现规律的功能不一定会随着规模变大而同步增长。对于文科的学者来说，要处理好大量的数据，可能也存在一些技术上的困难。此外，从建模的角度看，以往定量研究中的统计模型是验证驱动的，强调先有假设，再通过数据验证假设的合理性；而大数据模型是数据驱动的，强调建模过程以及模型的可更新性^[4]。这是一个比较大的区别，但对于语言研究而言并没有本质的差别。因为数据终究不能完全代替人，我们需要思考如何在数据的基础上作出更科学的解释，思考如何用数据回答关于语言结构规律和发展规律的问题。

2. 两种数据观：数据是否会说话？

大数据时代存在两种常见的数据观：一种观点认为数据会说话，不依赖于人，也很少受到人的影响；而另一种观点认为数据自己没法说话，是我们在为它说话并赋予它意义。

首先，我们认为数据当然不会说话，是人在用数据说话。比如“1”“2”在不同场景下代表的意义不一样，这只有人才能够理解这一点。所谓“数据会说话”是指人使用了数据，话可以说得更合理有据。定量方法或基于数据的方法，能够帮助我们更科学地验证过去的一些假设，或者更好地发现在小数据或没有数据的时代难以发现的一些模式。但是如果一个人对所研究的领域一无所知，那么数据再多也毫无用处。所有这些过程都需要人的主动参与，尤其是高级的研究活动，如发现、分析、归纳、解释和预测等，人的主动参与作用无法被机器所取代。因此，大数据最大的价值并不在于数据本身，而在于如何将数据与知识、社会、文化、行为以及

人联系在一起，并通过数理统计方法，更科学地发现数据背后隐藏的有关人类认知、行为的模式以及人与社会、自然交互的规律。

其次，至于数据的中立性，前文也已经阐述过，人对现实世界的观察和抽象是有选择性的，这也是建模的一般问题。以语料标注为例，标注语料的过程中或离不开人的直觉分析，或受到现有语言理论的影响，这是在所难免的。分析一个句子的句法，需要通过大脑的认知机制和语言系统，识别出主语、宾语或状语等等，然后标注出来。标注过程就反映了标注者对于这个句子乃至这种语言句法的认识，标注的过程实际上是人向机器传授语言知识的过程。如果有足够多的这种标注过的句子，机器就可以从中抽象出这种语言的句法知识。当然这里蕴含着一个问题：既然是人，对同一个句子的分析可能不一样，那么标注体系也就不一样。从句法模型看，句法主要包括：研究词间关系的依存句法、研究句子结构中部分与整体关系的短语结构句法以及把这两者结合起来的句法框架。不论是哪一种模型，都涉及人类语言中句法的抽象和建模过程，与其他科学领域一样，我们需要对现实世界抽象到一个高度，构建模型之后，再去研究这个模型。当然，从现实到模型的抽象不可能面面俱到，要涉及因素的取舍问题，这是所有科学研究都无法避免的，但只要模型能反映研究对象的主要特点即可。建模之后开始标注语料，标注过程中会有一些语言现象存在争议，因为每个人的语言直觉是不一样的。标注过程中，当然可以争论哪种标注方法更为合理。实际操作中，对于同一个有争议的现象，只要统一标注方式，对于规律的发现不会有什么大的影响。此外，也应该意识到，标注过程中能引起争议的部分毕竟很少，在整个系统中占的比例通常也是很小的。

也许有人会追问：如果搁置有争议的那一小部分数据，是否会对研究整体产生影响？一般不会。语言是一个动态复杂系统，在正常情况下处于平衡状态。平衡状态意味着我们可以用这种语言来完成基本的交际功能。反之，如果语言中所有的组分及结构都有争议，那么这个语言是不稳定的，我们无法用它进行交流，所以，有争议的只是其中极小的一部分，不影响全局，这是动态系统的特点：它是不断变化发展的，而语言系统的核心具有稳定性，是它能够作为交流工具而存在的基础，这使得我们能科学的方法来研究整个系统的核心。以词性标注为例，10000个词中，有10个词很难界定词性，

那么可以先临时搁置这 10 个词,因为规律最大的可能是在这 9990 个词里面。总之,要把语言看成是一个系统,而不是孤立地纠结一两个词,这可能是和传统的分析方法不太一样的地方。除此之外,我们也应该时刻提醒自己,语言是一个复杂适应系统,这意味着绝大多数语言规律可能都是统计规律。

3. 与大数据相关的几个误区

关于大数据,有一本畅销书曾被广泛推介,名为《大数据时代》^[25]。该书可能出于宣传目的,把核心内容压缩成了三句比较简短的口号:“要全体,不要抽样;要效率,不要绝对精确;要相关,不要因果。”这三句口号曾一度引发争议。在此需要指出的是,口号中说的“不要”并不是意味着完全抛弃,只是在强调重点发生了转移,我们的思维和处理方式需要转变。

关于第一句“要全体,不要抽样”,过去的技术手段难以处理规模过大的数据,需要借助随机抽样,以最少的数据来获取最多的信息。当今天的机器软硬件等技术条件日臻成熟,当机器可以支持处理关于全体的大数据,就不必抽样了。当然如果仍想用抽样同样也是可以的,要根据研究问题来决定。

至于第二句“要效率,不要绝对精确”,统计关注的是趋势,追求的不是绝对的精确。用计算机高效快速地处理完数据后得到数据的模式和趋势即可。大数据的核心是预测,例如气象大数据经过计算机处理得到模式和趋势之后,可以用来预报大约 5 个小时之后某个地区会降雨,提醒人们出门记得带伞就行了,无需把降雨时间精确到 5 个小时后的几点几分几秒。大数据模型擅长做预测,但不具有演绎性,这与追求必然性的物理定律不同,但并不意味着它不科学,只是二者各自有其适用的范围,目前还不能过多地苛求其精确性^[4]。

第三句“要相关,不要因果”引发的争议较大。我们知道,以理性主义为代表的学术研究追求的是因果关系。有人也许会问,如果不研究因果,我们还搞什么研究呢?如果不研究因果,要数据还有什么用?大数据寻求的是模式,然后在此基础上进行预测,如预测购买行为、天气状况、流行病传播等等,能解决实际的问题就行。但是不是意味着就彻底抛弃了因果?事实并非如此。作为一个学者,当然要探索因果关系。如果两个要素之间关系非常简单,容易发现因果,那么当然要研究因果。很多时候,涉及人与社会的情况错综复杂,利用大数据有助于我们发现相关,但进一步厘清因果则非常困难。

比如,我们投入了大量的精力与物力才对“吸烟有害健康”有了一个初步的因果认识。大量的行为实验难以重复的事例也说明,涉及人与社会的因果关系是很难一时半会儿厘清的,因为这样的系统大多是非线性系统,而“因果”更多的是线性系统的特点^[20]。笔者认为,因果关系是相关关系的一种,相关的偶然性蕴含着因果的必然性;如果相关已经能满足需要,就不一定再追求单一的因果关系了。大数据有助于发现因果关系,至少可以在相关的基础上接近因果。

由因果关系又引申出来一个小问题:目前基于数据的语言研究,发现的大多是一些可复现的模式;那这些模式与我们寻求因果的语言研究有什么联系?我们知道,寻求因果的研究多是由好奇心驱动的。用大数据做研究的人,同样具有学术好奇心。只要是研究,不论是大数据、小数据甚至无数据,都是有点好奇心的人才去做的。数据密集型研究范式,正如李国杰院士所说,是一种工具。人们用望远镜去观测星空,探求过去用肉眼难以感受到的宇宙深处的斑斓奇幻,现在感受到了,会不会更加好奇?工具能让我们发现一些从前看不到的模式,而这些模式可能进一步激发我们的好奇心,去思索为什么会形成这样的模式。而好奇心是所有学术研究的动力,它也许能更好地促进我们探求这种语言现象背后的原因——这就转到了因果关系的探索上。

三、几项基于数据的语言研究

关于信息时代的语言研究,前文已从研究方法上作出阐释。本部分以我们此前的几项研究成果为例,具体介绍如何开展基于数据的语言研究。

(一) 依存距离最小化研究

首先介绍依存距离最小化研究的案例。依存语法是建立在词间关系基础上的语法理论^{[26][27]}。我们知道,一个句子中的词是呈线性排列的,两个有句法关系的词在句子中可能紧挨着,也可能间隔其他的词。根据依存语法,两个有依存关系的词在句中的线性距离称为依存距离。依存距离有远有近,一般通过间隔词数来计算。通过依存距离,我们分析了过去心理语言学家做过的一些句子,发现心理实验中被认为难的句子,一般依存距离比较大。这说明依存距离可能与心理、认知因素有关,如工作记忆。如此,文本计量指标就可以和人的认知机制联系在一起,或者说有可能用经过依存句法分析的文

本来研究人的认知。假设依存距离与工作记忆有关,那么所有语言的依存距离应该十分接近,因为前文提到过,语言具有认知普遍性,受到认知规律的约束。十余年前,笔者开始基于20种语言的真实语料展开了进一步的研究^[28]。这是世界上首次采用大规模、跨语言的真实语言数据来进行的依存距离最小化研究。结果非常清楚地展现了至少有十几种语言的平均依存距离几乎是一样的;而人类语言的依存距离比所构造的非人类的随机语言的依存距离小。这验证了我们的假设:依存距离最小化有可能是人类语言的普遍规律。依存距离最小化展现了一种我们过去所看不到的模式,这种模式的特征展现了人类语言的普遍特征,体现了(大)数据的作用。

值得指出的是,依存距离最小化如果是人类语言的普遍特征,很容易让人感觉平淡无奇,因为一些不太了解依存距离最小化原理的学者,可能认为这是对乔姆斯基普遍语法的验证,但实际上两者是不同的。乔姆斯基认为,普遍语法是人与生俱来的一种大脑机制,它决定了人类语言的普遍性。但我们的研究认为,依存距离最小化实际上是由于工作记忆容量的约束而导致的,人在线性化造句的时候,依存距离应尽可能小。工作记忆当然不是专司语言的,而是人类普遍认知系统的一部分。换言之,依存距离最小化的特征是由人的普遍认知机制约束的,这并没有证明、也无法证明人脑中存在一个生物学意义上的专门负责语言或普遍语法的机制。也就是说,依存距离最小化并没有验证普遍语法存在与否。

在过去的十几年间,笔者团队从不同的角度继续完善对依存距离最小化的理解,比如“为什么汉语的依存距离比较大,我们却感觉不到它难”等等类似这样的研究。从这个意义上说,采用大规模、多语言的真实语料,可以帮助我们发现一些平常注意不到的语言普遍特征。

(二) 基于依存方向的语序类型研究

第二个例子是基于依存方向的语序类型研究。依存语法分析有三个要素:支配词、从属词和依存关系。一个句子中,支配词或位于从属词之前,或位于它之后,即存在支配词前置和后置两种不同的依存方向。采用依存方向比例这一指标考察了二十种语言的依存方向分布。基于大规模真实语料的数据,我们发现依存方向可以作为判定语序类型的指标;语序类型是一个连续统,任何一种语言均可以在这个连续统中找到自己的位置,并根据依存距离

的远近来进行聚类分析^[29]。例如,过去说某一种语言是“SOV语言”或“SVO语言”,但实际上每种语言可能都有SOV的成分,只不过可能某些语言中SOV的比例更大一些。这加深了我们对语言类型学的认识,也是(大)数据给我们带来的新发现。

(三) 基于依存距离的语言产出机制研究

第三个案例是从系统的观点研究语言的产出机制。既然语言是一个复杂自适应系统,当中则会涉及到调节的问题。一个句子的依存距离要尽可能最小,人们交流起来才可能更省力。对于一个只有3个词的短句,依存距离不会太大;而对于一个有30个词的长句,依存距离则有可能很大。在遇到长句时,语言的自适应机制被触发,从而使得这个句子的依存距离尽可能地小。我们知道,自适应系统在调节自身的过程中必然要围绕一个目标。如果我们从系统的角度研究语言,也需要有一个明确的设定值。如果依存距离最小化是句子线性化的目标或设定值,那么当人们产出一个长句时将会怎么做?我们通过计算机模拟的手段、采用真实语言标注的语料库对比的方法发现,在处理长句的过程中,很可能产生一种动态的语言单位,即组块。组块可以大大地减少长句的平均依存距离,在引入组块之后可以达到依存距离最小化^[30]。这是从系统的角度对语言产出机制的探索。

以上均是由笔者通过数据去验证或发现的,这些探索加深了我们对语言规律及语言处理机制的认识和理解。由此可见,数据密集型语言研究不但是可以进行的,而且能帮助我们发现过去难以发现的语言模式与规律、解决过去解决不好的问题。

四、关于学科建设与发展的一些思考

本部分阐述与语言学学科发展相关的一些思考,关注基于数据的方法在其中所起的作用。首先提出语言学的学科教学需尽量体现时代特点,并结合社会的需求;其次介绍当前高校“双一流”建设的背景下,基于数据的研究方法对于推动中国语言研究的科学化和国际化进程的作用;最后就跨学科语言研究加以讨论。

(一) 与时代需求相适应的课程设置与教学内容

关于语言学家的作用,杰利内克认为自然语言处理界其实一直期盼得到语言学家的帮助,但他们所需要的是将语言学与数据驱动的统计法相结合,使得机器能更好、更有效地反映人的语言

知识^[2]。类似的说法还有：“当雇佣一个受过良好训练的语言学家的时候，树库就会更好。”^[31]这与杰利内克的那句话遥相呼应。在今天，大部分建立在统计机器学习的自然语言处理以及建立在神经网络基础上的深度学习，需要大量的语言材料来进行训练。如果我们为语言材料赋予了句法或语义信息，机器就能够更好地学到句法或语义知识，从而能更好地处理人类语言。这些标注过句法或语义等信息的语料库被称为树库。树库是机器学习的知识来源。值得一提的是，世界上最早的大规模句法树库正是在杰利内克的支持下建立起来的^[7]。他最初想通过树库归纳出语法，从而为语音自动识别服务^[2]。如此说来，我们可能会感觉到语言学家是通过标树库来做贡献的。但遗憾的是，并不是谁都能从事这样的标注工作。一个“受过良好训练”的语言学家，至少应该知道目前自然语言处理界所采用的主流分析方法是什么。例如，就句法而言，在自然语言处理领域，大量实践经验已经证明了短语结构句法等分析方法存在一定的局限，目前主要采用的是依存句法理论。近年出现的基于普遍依存关系的依存句法标注体系，力图面向全世界的人类语言，最新版本已包括了50种语言的70个树库^①。但是学校里的语言学课程却很少涉及这些内容，或者说语言学专业的学生很少有机会了解自然语言处理领域的现状。因此，从这个意义上说很难学以致用。尽管社会需要“受过良好训练”的语言学家，但我们的专业教学内容却在一定程度上远远落后于时代，难以满足社会的实际需求。如果语言学家仅满足于创造各种概念，然后又围绕这些难以反映真实语言样貌的概念争论不休，那便真的可能近似于在讨论一个针尖上，究竟有几个天使能在上面跳舞了^[32]。当然，我们不是说在花园里养的花没有价值，即使是塑料花或绢花，也能为人们的生活增添色彩，但我们不能只生活在花园中，人类也许更需要面对真实的世界，因为这姹紫嫣红的大千世界不会由于我们的忽视就不复存在。因此，语言学家只有与时俱进，面对真实自然的语言材料，采用更科学的研究方法，所发现的语言规律以及得出的理论才有可能更好地服务社会。而语言学专业可能需要相应地开设一些课程，使得我们未来的语言学家有意也有能力从事具有鲜明时代特征的工作。

（二）数据密集型方法与语言研究的“两化”目标

从2010年至今，笔者一直在各种场合明确提出语言研究的目标——中国语言学的国际化与语言研究的科学化（简称为“两化”）。我们为什么需要这么做呢？

一方面是中国语言学的国际化。根据定义，语言学研究的是语言系统的规律，它应该具有普遍意义。记得多年以前，笔者对语言感兴趣是从学外语开始的。后来一次很偶然的机会，读到一句话：“学语言是给个人增加新知识，研究语言学是给全人类增加新知识。”^[33]这句话令人触动颇深。语言学研究应该具有普遍价值。我们知道，中国拥有的语言学研究队伍可能是世界上最为庞大的。但实事求是地讲，改革开放以来或者更早一些，我们中国大陆本土的语言学家对于世界语言学的贡献是比较有限的。这不是说我们自己的研究没有价值，而是世界很少知道我们的研究。当然其中牵涉各种各样的问题。无论出于什么样的原因，世界确实对于我们的研究所知甚少。这显然和中国整体的经济和科学发展水平不相适应。国家和社会的迫切需求使得中国语言学研究必须走国际化之路，尤其在现在的“双一流”建设背景下，更要求我们的学科不应该关门来自己干，而应把优秀的成果拿出去与世界分享。成为世界一流的前提是让世界知道。当我们现在提倡要建设世界一流，而世界却从未听闻，又怎能谈得上一流？正如一个人总是说自己是某个体育项目的世界冠军，但却从未在世界各种体育比赛中露面，恐怕是不符合常理的。只有让世界知道，站在世界比赛的起跑线上与别人同场竞技，才有可能谈得上争取世界一流，才有可能证明中国的语言学家也可以研究一些有趣的问题，也能对世界语言学的发展作出贡献。

另一方面是语言研究的科学化——这不只是中国语言学家的事情，可能也是全世界语言学家追求的目标。在高水平的科学期刊上发表学术论文，是获得科学共同体认可的有效途径。但目前看来，这样的文章很难发表。如果一个学科在科学家认可的期刊上几乎很少有文章发表，那么它如何成为科学，甚至于是“领先科学”？之所以难，一个很重要的原因在于科学的研究需要采用科学的方法。反思语言学的发展现状，以科学方法开展研究是语言研究

①下载地址：<http://universaldependencies.org/>。

走向科学化的必经之路。

那么，数据密集型研究范式与“两化”有什么关系？细究起来，除了语言障碍之外，很多时候还有其他原因，其中也包括研究问题和研究方法。在研究问题的选择上，如何从汉语中的特殊问题引向更具有普遍意义的语言学问题，是值得我们深思的。在研究方法方面，数据密集型的研究范式，可能比纯粹思辨的、内省的方法更容易获得当今学界的认可，而无论是验证假设还是发现模式，都是需要数据的。我们应该思考如何在发挥传统优势的基础上，结合学界通用的方法，把中国好的研究推向世界，让世界知道中国人也可以做出好的研究成果。因此，数据密集型研究范式无疑是能促进“两化”具体实现的。

（三）大数据时代的语言学跨学科研究

近年来，跨学科研究成为学术界的热词之一。我们知道，最早的时候学科划分是不存在的，历史上文理兼通的人数不胜数。后来由于技术的发展，探索的手段层出不穷，趋于复杂多样，而一个人也不可能同时掌握那么多知识和技能，因此学术分工更加精细，形成了学科划分。经过几十年的精细化研究历程，我们发现精细化的方法近似于采用一种盲人摸象的方法，从整体来说对大象的认识还是需要合起来。因此在探讨同一个研究对象的时候，倾向于采用不同的方法和工具。比如在研究语言的时候，借鉴生物学、物理学或数学的方法，这时便出现了所谓的跨学科或者交叉学科。

今天不少人存在一个误区，以为任意几个不同专业的人合在一起做事就是跨学科。这样做的效果往往并不太理想，主要在于没有厘清并落实研究问题。对于现阶段的语言学跨学科研究，从理论上讲，应该是借用别的学科的方法来研究语言学问题。比方我们对一个语言学问题比较好奇，当本学科现有的方法难以研究这个问题时，是不是可以借用其他学科的方法？

这里举一个语言学跨学科研究的例子。儿童语言习得研究发现，儿童大概在两岁时，母语的句法会出现一次飞跃。如果把语言看成一个复杂适应系统，那么儿童的母语句法会出现涌现现象。尽管掌握的词汇不如成人，但在两岁的时候说出的句子的句法模式可能已接近成人。过去的心理语言学、儿童语言习得的实例观察均发现了这一现象，但很难清晰地展现出来。前几年，西班牙的一些学者用

复杂网络展示了两岁左右时儿童母语句法的涌现现象^[34]，十分直观形象。可见，“跨学科”并不是漫无边际地“跨”，“跨”的本质是一种“拿来主义”，即从别的学科尽量借鉴一些方法来解决本学科的研究问题。

笔者团队近几年在语言学跨学科研究方面也取得了一点成果，例如借用复杂网络的方法，对斯拉夫语言进行了类型学研究。当今语言类型学的主流是语序类型学，而在分析形态变化比较丰富、语序相对自由的斯拉夫语族时，过去的语序类型学方法不太适用。我们从统计物理学中借鉴了复杂网络的方法，基于十二种斯拉夫语言的真实文本，采用复杂网络的指标对这些语言进行了分类研究^[35]。大家可以参看我们发表在《生命物理学评论》(Physics of Life Reviews)上的两篇采用复杂网络研究人类语言规律以及如何采用依存距离来发现人类语言线性化模式的文章^{[36][37]}，体会跨学科语言学研究的旨趣。

前面提到的两个案例中，“跨”并不是跨到物理学中——当然从物理学的角度来讲，也拓展了复杂网络方法的应用领域，提供了蕴含普遍性的真实网络实例，丰富了复杂网络理论。而对于语言来讲，采用复杂网络帮助我们解决了过去不太容易解决的语言学问题。当然，随着两个学科彼此借用越来越频繁、关系越来越密切，极有可能形成一个交叉的学科，甚至可能形成新的研究范式。交融的程度加深，使得这个新学科不同于原来的任一学科，例如可能有一天分不清究竟是物理语言学，还是语言物理学。

那么大数据是否也有助于跨学科发展？从实际操作层面来看，语言学的跨学科研究需要对所“跨”的领域有一定的了解。如果我们把语言学定义为“研究语言结构模式和演化规律”的学科，这当然是很狭义和传统的定义，因为语言学中还涵盖着很多内容，不过归根结底还是要处理语言数据的。在处理语言数据时，要用统计学、数学和计算机科学的知识，例如借助生物学用来研究网络的软件，来研究从语言数据构造出来的网络，也属于语言学的跨学科研究。此外，在语言作为复杂适应系统的视域下，从真实文本材料中得到的规律，有可能指导当今颇具潜力和发展前景的计算语言学和自然语言处理，那么我们实际上还是在和语言数据打交道。因此，基于数据的方法显然也能促进语言学的跨学科研究与发展。

五、余 论

语言学是一门科学,但不只是我们自己嘴里说而已的科学,而应该得到科学共同体的承认。这些年的努力使我们体会到:语言学研究可以实现科学化,但前提是采用科学的方法。显然,科学的方法,需要我们付出更多的努力去学习与掌握。从长远来看,这样的付出是值得的,不论是对语言学的学科建设还是个人的学术发展都不无裨益,而且非常必要。要有付出,敢于啃硬骨头,才能有所突破。一个谁都可以轻易入门、指点江山的学科,可能很难与科学挂上钩。复旦大学的葛兆光教授曾写过一篇文章《人文学科拿什么来自我拯救》,面对人文学科日渐衰落的境况,他在文中指出:“打铁还需身板儿硬”。文章最末,他还写道:“如果大学人文知识就是这些业余可以模仿习得的东西,那么何必还要这些拥有博士、教授头衔的人在这里坐馆?”^[38]人文学科都需如此,何况号称科学的语言学?

(致谢:感谢为本文提供问题的所有同学与老师,你们对语言研究的兴趣是促成本文的主要动力。感谢陈衡、陈芯莹、黄伟、蒋景阳、梁君英、陆前、徐春山、于水源对本文初稿提出的建议,你们的参与使得这篇访谈更像是一篇论文了。)

参考文献:

- [1]Hirschberg J. “Every time I fire a linguist, my performance goes up”, and other myths of the statistical natural language processing revolution (Invited speech)[R]. 15th National Conference on Artificial Intelligence (AAAI-98), Madison, Wisconsin, 1998.
- [2]Jelinek, F. Some of My Best Friends Are Linguists[J]. Language Resources and Evaluation, 2005, 39(1), 25-34.
- [3]陈工孟, 须成忠. 大数据导论: 关键技术与行业应用最佳实践[M]. 北京: 清华大学出版社, 2015.
- [4]李国杰. 对大数据的再认识[J]. 大数据, 2015, (1): 8-16.
- [5]Bresnan J, Asudeh A, Toivonen I, et al. Lexical-Functional Syntax[M]. 2nd Edition. John Wiley & Sons, 2015.
- [6]Bresnan J. Linguistics: The garden and the bush[J]. Computational Linguistics, 2017, 42(4): 599-617.
- [7]Jelinek F. The dawn of statistical ASR and MT[J]. Computational Linguistics, 2009, 35(4): 483-494.
- [8]Bod R, Hay J, Jannedy S. Probabilistic Linguistics[M]. Cambridge, Mass: The MIT Press, 2003.
- [9]伯纳德·科姆里. 语言共性和语言类型[M]. 沈家煊, 译. 北京: 华夏出版社, 1989.
- [10]Hauser M, Chomsky N, Fitch T. The Faculty of Language: “What Is It, Who Has It, and How Did It Evolve?”[J]. Science, 2002, 298(5598): 1569-1579.
- [11]Sampson G. Depth in English grammar[J]. Journal of Linguistics, 1997, 33(1): 131-151.
- [12]Karlsson F. “Syntactic recursion and iteration”, Hulst H V D, editor, Recursion and Human Language[M]. New York and Berlin: Mouton de Gruyter, 2010: 43-67.
- [13]刘海涛. 计量语言学导论[M]. 北京: 商务印书馆, 2017.
- [14]Holland J. H. Hidden Order: How Adaptation Builds Complexity[M]. NY: Basic Books, 1995.
- [15]米勒, 佩奇. 复杂适应系统: 社会生活计算模型导论[M]. 隆云滔, 译. 上海: 上海人民出版社, 2012: 309.
- [16]王士元. 语言是一个复杂适应系统[J]. 清华大学学报, 2006, 6(21): 5-13.
- [17]Kretzschmar W. Language and Complex Systems[M]. Cambridge: Cambridge University Press, 2015.
- [18]Ellis N C, Larsen-Freeman D. Language as a complex adaptive system[M]. New Jersey: Wiley-Blackwell, 2009.
- [19]许国志. 系统科学[M]. 上海: 上海科技教育出版社, 2000.
- [20]Solé R V, Goodwin B. Signs Of Life How Complexity Pervades Biology: How Complexity Pervades Biology[M]. New York: Basic books, 2008.
- [21]埃德加·莫兰. 复杂性思想导论[M]. 陈一壮, 译. 上海: 华东师范大学出版社, 2008.
- [22]Eddington D. Linguistics and the scientific method[J]. Southwest Journal of Linguistics, 2008, 27(2): 1-17.
- [23]艾登, 米歇尔. 可视化未来: 数据透视下的人文大趋势[M]. 王彤彤, 沈华伟, 程学旗, 译. 杭州: 浙江人民出版社, 2015.
- [24]Hey T, Tansley S, Tolle K M. The Fourth Paradigm: Data-intensive Scientific Discovery[M]. US: Microsoft research Redmond, WA, 2009.
- [25]舍恩伯格, 库克耶. 大数据时代: 生活、工作与思维的大变革[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.
- [26]刘海涛. 依存语法和机器翻译[J]. 语言文字应用, 1997, (3): 87-93.
- [27]刘海涛. 依存语法的理论与实践[M]. 北京: 科学出版社, 2009.
- [28]Liu H. Dependency distance as a metric of language comprehension difficulty[J]. Journal of Cognitive Science, 2008, 9(2): 159-191.
- [29]Liu H. Dependency direction as a means of word-order typology: A method based on dependency treebanks[J]. Lingua, 2010, 120(6): 1567-1578.
- [30]Lu Q, Xu C, Liu H. Can chunking reduce syntactic complexity of natural languages?[J]. Complexity, 2016.
- [31]Eberhard-Karls-Universität Tübingen. Linguistic Treebanks and Data-Intensive Parsing (ESSLLI 2005): Treebanks: An Overview[EB/OL]. <http://www.sfs.uni->

- tuebingen.de/~kuebler/esslli05/treebank-intro.pdf.
- [32]Percy W, Samway P. Signposts in a Strange Land[M]. New York: Farrar, Straus, and Giroux, 1991: xv, 428 p.
- [33]徐烈炯.生成语法理论[M].上海:上海外语教育出版社, 1988.
- [34]Corominas-Murtra B, Valverde S, Sole R V. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition[J]. Advances in Complex Systems, 2009, 12(3): 371-392.
- [35]刘海涛, 丛进.基于平行词同现网络的语言聚类[J].科学通报, 2013, 58(5):432-437.
- [36]Liu H, Xu C, Liang J. Dependency distance: A new perspective on syntactic patterns in natural languages[J]. Physics of Life Reviews, 2017. Retrieved from [http://doi.org/10.1016/j.plrev.2017\(03\):002](http://doi.org/10.1016/j.plrev.2017(03):002).
- [37]Cong J, Liu H. Approaching human language with complex networks[J]. Physics of life reviews, 2014, 11(4): 598-618.
- [38]葛兆光.人文学科拿什么来自我拯救[J].上海采风, 2012, (9):96.

Methodology and Trends of Linguistic Research in the Era of Big Data

LIU Hai-tao^{1, 2} LIN Yan-ni¹

(1.School of International Studies, Zhejiang University, Hangzhou 310058; 2. (National Key)Research Center for Linguistics & Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou 510420)

Abstract: This paper presents methodology and trends of linguistic research in the era of big data. We begin with a discussion on the role of linguists in the information society, and illustrate the chances and challenges linguists are currently facing. After highlighting the significance of using authentic data in linguistic research, we argue that language is a complex adaptive system driven by humans. Then, from the perspective of philosophy of science, we introduce the research paradigm of quantitative linguistics with several cases. Finally, we address that China's linguistic research will benefit from the data-intensive approach in terms of scientification and internationalization.

Key words: Linguistics; Big Data; The Data-intensive Approach; Scientific Research Paradigm

[责任编辑：马瑞雪]

[责任校对：李 蕾]