大数据时代语言学理论研究的路径与意义

刘海涛 郑国锋*
（浙江大学，杭州，310058；华东理工大学，上海，200237）

摘要：郑国锋、惠特曼（2020）指出，由于计算语言学的崛起，当代语言学理论研究面临着发展危机，作为旨在发现与总结语言结构及其演化规律的语言学理论研究，在大数据时代需要在研究对象和研究方法上做出调整。刘海涛认为，考虑到来自计算语言学等领域的挑战，语言学理论研究有必要在研究范式上转向数据驱动，在研究对象上面向真实语料，这不仅为理论语言学的发展提供了机遇，也为中国语言学在世界范围内取得领先地位提供了最佳时机。

关键词：理论语言学；计算语言学；数据驱动
［中图分类号］H0 ［文献标识码］A ［文章编号］1674-8921-(2021)02-0005-14
［doi 编码］10.3969/j.issn.1674-8921.2021.02.001

1. 前言

郑国锋、惠特曼（2020）指出当代语言学理论研究面临的不是内部竞争，而是外部竞争，尤其是来自计算语言学的竞争。本次访谈从语言研究的形式与语义两方面，聚焦大数据时代语言学理论研究危机的根源、语言学理论研究的方法与对象、语言学理论研究呈现的“两张皮”现象、语言与言语的区分、依存距离研究、言语平均数、语言人工智能研究等主题。当代的语言学理论研究有必要借助现代化的研究工具，锚定实际语言应用，探求真实语言规律，走语言理论研究科学化与国际化之路，在“百年未有之大变局”中做出自己的贡献。

2. 访谈内容

郑国锋：按照康奈尔大学语言学系惠特曼（Whitman）教授的观点，语言学研究目前存在危机，但其根源不是语言学研究内部各个流派之间的竞争，而是源自计算语言学的进步和各种算法的出现。您是否同意目前语言学研究存在危机，以及计算语言学的进步是造成这一危机的根源？

刘海涛：如果把语言学看作是一个纯粹的、基础的、理论的、能自圆其说的

作者简介：刘海涛，浙江大学求是特聘教授、博士生导师。主要研究方向为数字人文、计量语言学、依存语法和语言规划。电子邮箱：htliu@163.com
* 郑国锋（通信作者），博士，华东理工大学外国语学院副教授。主要研究方向为对比语言学、语言类型学、运动事件的语义-句法接口研究。电子邮箱：zgfcl@ecust.edu.cn
引用信息：刘海涛、郑国锋. 2021. 大数据时代语言学理论研究的路径与意义[J]. 当代外语研究(2)：5-18，31.

学科，我觉得不存在危机。但实际上，语言学不只存在于理论层面，除了探究语言系统运作的规律性或理论性问题之外，人类还会遇到很多与语言应用相关的其他问题。但在过去，这些问题都是由应用语言学家解决的，而理论语言学家不关心、也不屑于解决这些问题。应用语言学家，例如语言教学、语言规划等领域的专家则致力于解决自己领域的问题，其间或许会运用一些语言学的理论，但如果没有这样的理论或理论不适用，他们就自己想办法来解决遇到的问题。

近些年，由于互联网的普及，操不同语言者之间的交流较过去更快、更密切，这使得古老的语言障碍问题更为突出，这是其一。其二，互联网的出现导致大量的、不同语言的信息涌现出来，网络信息爆炸，更确切地说是多语信息爆炸。此时，人们自然想通过语言学理论与计算机技术的融合解决信息爆炸的问题。这就涉及人工智能，即用计算机处理语言。

自 20 世纪 40 年代计算机出现以后，人们就设法用计算机解决语言问题，比如机器翻译，计算语言学也由此开始。起初，计算语言学家使用语言学理论提供的一些模型，主要是一些形式模型，但收效甚微。这有几个原因：首先，计算机算力本身的发展需要一个过程；其次，语言是一个开放系统，10 条、20 条，甚至 1000 条规则，也只能处理封闭环境下有限的句子，难以面对无限的人类语言。

90 年代以后，计算语言学经历了一次革命，即不再采用语言学家所谓的语言规则，因为他们认为这些规则只能解释一些封闭环境下的句子，不适用于语言的无限性。人们发现，计算机能够设法从人类语言中学习语言知识。当然，如果在机器学习的过程中，能够获得语言学家的帮助，就能够更快、更好地学到这种语言的句法知识。但是，机器要求的不是所谓的离散规则，如 S 可以重写为 NP＋VP，因为这样的规则在大规模真实的语言使用过程中被证明效果不是太好。基于此，近几年的人工智能领域改进了机器学习的算法，以及机器学习到的知识的表达方法，然后把学习的结果表示成各种形式的人工神经网络，这便是深度学习。在技术不断突破以后，机器学习的效果比以前改善了很多，比如现在我们要翻译常用的或一般用途上的文字，机器翻译的水平要好于一般的外语学习者。

郑国锋：可以看出，在计算语言学大步向前的时候，语言学理论研究几乎隐形，这个现象值得深思。

刘海涛：基础研究可以忽视研究的用途，但其发现的规律需要验证。而计算语言学或者自然语言处理就是验证这些规律的最好领域。他们发现理论语言学研究出来的规律不好用，而计算语言学者自己摸索出来的东西反而更好用。我们处于最需要语言学理论的时代，但是语言学家却集体失声了。一批像惠特曼教授一样有良知的人就开始反思，认为我们存在危机，而这个危机可能是来自计算语言学的进步和挑战。在这个意义上，危机是存在的。

郑国锋：惠特曼教授还认为以代数计算为特征的形式语言学将会持续繁荣，主要是因为该流派可以更好地对接计算技术的进展，但您在“智能时代的语

言研究:问题、资源、方法”主题报告中指出,有代数是不够的(刘海涛 2020)。请问为什么?

刘海涛:采用代数的形式研究语言是形式语法的做法。理论上,它比我们采用的其他方式,如在形式语法诞生之前的传统语言学的其他分支,更容易与计算机对接。但是,就像我们前面谈到的,从 20 世纪 50 年代到 90 年代的 40 年间,语言学家基本上是采用形式的方法来研究语言,包括理论语言学和计算语言学,但是结果证明在面向真实的语言应用场景的时候,这种方法是不够的。最有名的例子是 1970 年左右基于系统功能语法搭建的维诺格拉德的积木世界,应用场景虽小,但效果非常好,当时震惊了全世界。但人不是生活在游戏的空间里,更不是生活在人造的花园里。当人们走出花园,走出人造的小世界时,会发现只有那些是不够的,而且修修补补也不能解决真实世界中更大、更广泛的问题。这也是为什么从 20 世纪 90 年代以后,计算语言学家们另起炉灶的主要原因,因为那种方法很难应对我们真实世界的语言使用场景。

洪堡特(Humboldt)(1999:116-117)认为语言是有限手段的无限运用,这给人一种错觉:似乎只要搞明白了“有限手段”,我们就能掌握“无限运用”。语言是很复杂的现象,在经过多年实践之后,人们发现,就人类语言而言,理论上的“有限”,实际上是一个“无限”,理论上的“无限”,实际上是“有限”。我们过去总认为机器不能像人一样从已有的文本(语言使用)中学到语言的一切,是因为我们面对的是“无限”,而你的知识是“有限”的,但是人类就是这么学的。语言本身不是非黑即白的二元系统,而更像是一个多阶的灰色概率系统。在采用数理手段研究语言的时候,不能忘记语言最根本的属性——动态性和概率特征。

实际上,代数系统、形式方法在描写计算机程序设计语言时取得的成功,要远大于分析人类语言时取得的成功。计算机程序设计语言是人造的、形式化程度极高的形式系统,与人类语言区别明显。从这个意义上讲,我们说形式方法是有用的,不过只有它是不够的。我们现在最大的危机和挑战在于,基于语言使用的计算语言学,在许多应用方面比过去只用形式、规则的方法取得了更好的成果。我们所面对的危机是要解释:我们过去认为行不通的做法却被他们证明是可行的;我们过去认为可行的方法却行不通。如果我们是有良知的语言学家,就要从理论的角度回答这个问题。

郑国锋:语言学研究的对象可以是语言,也可以是语言使用者。人类是语言使用者,萨丕尔-伍尔夫假说认为使用者与语言互相影响,甚至会被语言左右。语言使用者是生理、心理、社会、自然的统一体,这些也是语言表述的主要内容。请问目前的语言学理论研究对这个系统的探讨充分吗?这些研究的主要不足是什么?

刘海涛:如李宇明先生(2018)所言,语言学是一个学科群,我们可以从不同的角度研究语言,而语言和人密不可分。如你所言,语言学研究的对象可以是人的语言,也可以是这个人。可以研究人产生语言的过程,也可以研究人产生

出来的言语，即语言。《现代汉语词典》对语言学的定义是“研究语言的本质、结构和发展规律的学科”(2016:1601)。如果再准确一点，语言学是研究语言结构的模式和演化规律的科学。如果语言学是科学，其目的是探究语言的结构的组成规律、演化规律，那么语言学家作为最懂语言的人，特别是理论语言学家，其注意力就要放在语言上，而不是放在人身上。这从逻辑上讲是对的，但是语言又和人密不可分，是人的认知机制的产物。学者们提出过各种各样的说法，最早从索绪尔开始，语言成了一个符号系统，研究者从符号学的角度研究语言，但是这种研究把人置身事外。另外，动态性、演变性、概率性是语言的根本属性，也没有体现出来。这样研究出来的成果，只能反映语言系统的一部分特点，因此也不太好解决相关的问题。

郑国锋：语言既然是一个系统，就需要采用系统的方法来研究。王士元先生(2006)认为语言是一个复杂适应系统，您近年来更进一步，提出语言是一个“人驱的复杂适应系统”(Liu 2014, 2018;刘海涛、林燕妮 2018)。请问这样的考虑是什么？

刘海涛：二战以后，出现了大量系统的、专门的科学理论，人们又借用这些系统的理论和方法研究语言。王士元先生认为语言是一个复杂适应系统，但这个观点没有突出人的作用。所以从前几年开始，我们倡导语言是一个“人驱的复杂适应系统”。人驱系统的意义在于，无论结构模式也好，演化规律也好，驱动语言这个系统存在的、发展的原始动力是人，而人本身有两个特点：首先，人具有生物学意义上的普遍性，它决定了各种语言间会有一定的共性；其次，人又具有社会属性，这些生理之外的时空的、社会的因素也会影响语言，并成为人类语言多样性的一个源头。因此，当语言被表达为一个人驱复杂适应系统以后，有助于我们更好地解释和理解语言的规律及其认知动因和社会动因。

郑国锋：我们经常提到研究中的“两张皮”现象。当前语言学理论研究要么强调“语言”，要么突出“人”。如果我们想让自己研究的成果或规律(无论是结构的规律，还是演化的规律)能够被那些需要语言规律的研究者使用，或如果要迎接计算语言学的挑战，我们需要做些什么调整？

刘海涛：首先，语言学的研究对象应该与计算语言学所处理的对象一样，只有这样发现的规律才能够被计算语言学所使用。例如，我们研究怎么把塑料花或木头花做得更漂亮，即使研究的整个流程看起来很完美，这些研究成果在真正的园丁手里却没有什么用，因为他面对的是经历风吹、日晒、雨淋的真花。因此，如今我们急需的或要反思的是理论语言学要面向大量真实的语言材料。而且我们不要因为不好处理，或不好解释，不太符合我们心目中的语法，就忽略这些材料。机器帮我们处理语言工作的时候，它们面对的也是这些材料，因此语言学就不能不去面对它。

其二，在研究方法上，我们现在很多语言学理论研究是靠内省法和语感，而且还要靠母语者的语感，这便催生了另一个问题，例如中国的外语学者不能研

究外语，只能研究汉语，因为他们不是外语的母语者。通俗地讲，世界上有任何一个号称科学的学科要研究蚂蚁，那么研究者就要变成蚂蚁吗？不需要。所以只有通过语言的使用来发现语言的规律，才能找到真正的、符合科学意义的规律。我们可能不懂这种语言，但只要有科学的方法，也有可能找到这种语言的结构规律和发展规律。而这可能是真正的语言系统运作的规律。当然，为什么会有这样的规律？我们需要给出解释。只有意识到这种转变，我们才能有资格迎接计算语言学的挑战。搞计算语言学的人做机器翻译并不需要必须懂这两种语言，这在过去是不可思议的一件事，但是现在这些研究者做到了，而我们语言学家还停留在只能由蚂蚁研究蚂蚁的阶段。

郑国锋：语言形式是语言学研究的重要领域。如您所言，人的生物学意义的普遍性加上社会意义的多样性可能是语言形式普遍性与多样性的源头。在大数据的背景下，这对于语言形式研究有什么启发？

刘海涛：基于这样的认识，我们需要用科学的方法来研究普遍性的源头。例如，由于认知生理意义上的普遍性导致语言的普遍性，我们需要通过认知科学的方法研究那些与语言有关的认知普遍性。认知普遍性不一定都和语言有关，只有和语言密切相关的特征，例如工作记忆的容量，才会导致语言普遍性的产生。而这种普遍性的产生不是说研究者自己想通就可以了，也不是研究者用一两个例句就解决问题了。要用真实的语言材料，而且不能只用主谓宾齐全的简单陈述句做研究。我们要用人类各种真实场景下用于传递信息的语料去研究，从这些语料中挖掘语言的特点，而不是挖空心思地去找自圆其说的普遍性。

郑国锋：传统意义上，语言形式研究是语法研究，主要包括词法和句法。在英语作为事实上的通用语的今天，大量研究以英语形式为对比项，出现了许多牵强的结论。这与研究者过于重视普遍性或多样性有关。请问这样的形式研究偏误如何解决？

刘海涛：普遍性是从多样性中挖掘到的普遍性，或者是从普遍性出发来解释多样性。无论是从多样性抽象出普遍性，还是从普遍性分散为多样性，只研究一种语言是不够的。世界上 7000 多种语言中，有和英语结构类似的，但是大部分未必和英语结构一样。因此，只参照英语研究人类的普遍性与多样性是不够的。

导致多样性的因素可能会有很多，大部分可能是来自语言外的因素。一种不同的语言可能和英语的发展和演化不一样，那么导致它多样性的因素也是不一样的。田野调查很重要，为我们提供了基本材料。但是，当我们有了基本材料以后，我们要使用系统科学的方法来挖掘多样性，或者透过多样性寻找语言的普遍性。无论如何，多样性和普遍性是人类语言的一个事实。如果语言学是寻找语言结构的模式和演化规律的话，那么就要回到现实的语言使用中，只有这样才能发现真正的规律，以及多样性背后由于认知或其他生物学意义上的普遍性而导致的语言普遍性。否则，那些研究虽不能说是在做无用功，但至少可

能不是我们日常使用的语言规律，只是研究者自圆其说的规律。

郑国锋：随着时代的变迁，人们对语言的使用变得日益不同，传统的语言形式研究理论，比如词法理论，不再能够解释当前的语言现象，主要原因是什么？作为理论语言学家，应该如何应对？

刘海涛：首先，语言（形式）是变化的，因为人所处的社会环境是变化的。人处在不同的时空里，而时空是不断变化的，语言（形式）因此也会变化。我们可以用静态的方法研究语言（形式），但是这遗漏了语言作为一个动态系统的最重要的特征——动态性。理论上讲，一个人一生中说的话在不同时期是变化的；不同的人在不同的时期，或不同的人在同一个时期说的话也都是变化的，所以变化是语言（形式）的根本属性。

语言学理论研究对语言（形式）的各种变化都要持一种开放的态度，但是过去人们很难掌握这种变化以及在这种变化中做研究，研究者只能找一个语言（形式）的横切面来研究。作为一名语言学家，如果想让我们的研究成果有用，一定要在这种变化中寻找那些稳定的部分，因为正是这些稳定的部分形成了语言作为人类交流工具的基础，变化的是语言（形式）外围的一些特征。这要求我们在研究语言规律的时候，首先要把握核心的、基础的、相对不容易变化的特征，然后识别哪些特征是容易变化的。我们要切记不能把那些变化的、相对不稳定的因素当成一个语言的核心要素来研究，否则研究出来的规律可能很难解释我们遇到的语言问题，也很难被其他领域的需要规律的研究者使用。

郑国锋：100 多年前，索绪尔将语言研究二分为“语言”和“言语”，语言学研究应该怎么应对这对互伴互生的概念似乎成了永恒的话题。大数据时代，语言形式与语言技术高度发达，我们还要作此区分吗？

刘海涛：100 多年前，索绪尔的《普通语言学教程》（简称《教程》）出版以后，普通语言学的主要任务就是发现语言的普遍规律。索绪尔在《教程》里区分了语言和言语。后来的很多语言学流派也都进行了区分，只是术语不同。在那个时期，这种区分是可以理解的。语言具有复杂性，限于当时的技术手段，学者们很难从看起来杂乱无章的海量语言材料里发现规律。索绪尔在《教程》里是这样区分语言和言语的：言语就是我们一般的言语活动，和社会密切相关。他同时强调，在大量的言语交互中，人们慢慢地会在大脑里形成一种言语活动的平均数，而个人是言语的主人（索绪尔 1999：30-34）。如果我们能够全部掌握储存在每个人脑子里的词语形象，也许我们就会接触到构成语言的社会纽带。在索绪尔时期，言语是分散的，每个人大脑里的言语会通过言语的平均数构成语言，但每个人脑子里的语言都是不完备的，只有在集体中它才能够完全存在。我们假设有 100 个人在讲同一种语言，在这 100 个人互相交流的过程中，个人的言语通过频繁的交流会形成一个平均数，即使达不到 100%，也可能会形成 70% 到 80%的程度。由于这个平均数的存在，这 100 个人才可以用这种语言进行交流。索绪尔虽然区分了语言和言语，但是他也明确了语言和言语之间的关系，

即语言实际上是言语的平均数，语言不是个体的，是集体的。

100 多年前，他没办法计算平均数，但今天我们可以把这 100 人 20 年内说的话收集在一起，则有可能求得言语的平均数。所以现在当我们重读经典的时候，要理解先驱者过去为什么要做像语言和言语这样的区分。他们想到了这件事，但没有办法这么做，于是选择了另外一条路。100 年后的我们，比 100 年前的先驱更有条件来做好这件事。我们在现代技术的辅助下，通过大量的语言使用求得平均数，而这个平均数可能比我们基于个人语言实践得出来的平均数能更好地把握语言的规律。这就是我们为什么认为今天的人工智能（计算语言学是它的一部分）比以往更好，因为他们实现了前辈语言学家的想法：从数以万计的语言使用者产生的真实的语言材料中，获得了语言的规律，然后使用这种规律去分析解决语言问题。但是，我们现在还缺乏这种数据驱动的语言理论研究，这可能也是为什么我们难以解释计算语言学成功的主要原因。

语言和言语已经“分离”了 100 多年，现在是我们要终结这个“悲剧”的时候了。前辈语言学家不得已而为之，让它们分离，尽管实际上它们可能是难以分开的。今天也许是时候让它们合二为一了。因为只有合在一起，我们才能够从具体到抽象，真正抽象出人类语言的规律。也只有用这样基于大规模的、大量语言使用者的数据方法，我们才有可能实现语言研究的转变。从我们关注的像塑料花、木头花、绢花这样的花园里，走到真实的人类语言的灌木丛中，去发现人类语言使用的真正规律。

郑国锋：如果形式是“表”，语义就是“里”。自从建筑学家苏利文（Sullivan 1896）提出“形式遵从于功能”后，功能或者意义成为表达的主导。语言学研究也一向将语义作为研究的中心，比如语义的产生、语义的磨蚀等。您认为目前的语义研究主要有哪些路径？

刘海涛：形式语法出现后很长一段时间里，很多研究重视语言的形式方面。依照索绪尔的观点，语言是一种符号系统，是形式和内容的统一体。语言既是交流工具，也是思维工具，形式研究的主要意义和价值是通过形式掌握它的内容。因此，任何面向实用的，或者面向语言作为一种交流或者思维工具的研究，就避不开内容，也就是语义，这才是语言无论是作为交流工具还是思维工具最重要的功能。

目前语义研究的主要方法是语义分解法，也就是为了表达、理解清楚语义，很多时候我们要创造一种元语言。常见的形式语义学是用逻辑语言表达语义，这里的问题是逻辑表达本身，即这些符号，也构成了一种人造语言系统。已经有不少学者认为这套系统的表达能力不如自然语言，不能涵盖我们日常语言表达的所有内容。因此采用形式语义学的方法，就难以表达自然语言表达的所有内容。第二个是采用分解的方法表达语义，例如使用义素分析法、语义场分析法等区分词义，实际上又创造出一套人造的体系，这个体系认为意义可以分解，采用的是逻辑方法。它们源于哲学，特别是形式逻辑，认为自然语言中有很多

模糊不清的内容，表达的意义也是模糊不清的。因此，学者们尝试寻找或创造一种完美的语言，然后可以更清楚地表达我们的思想。在17、18世纪的欧洲大陆和英国，很多重要的哲学家如笛卡尔、莱布尼茨等都积极参与这个叫作“寻找完美语言”的运动(Eco 1995)。他们想找一种普遍语言，像数学般精确，比人类语言更精确地表达我们的思想，但是失败了。

郑国锋：这些聪明的哲学家和语言学家，为什么会失败呢？

刘海涛：我们可能需要仔细分析人类自然语言表达的语义，因为它可能本质上就是模糊的。从数学的角度看，这个世界是不完美的，但是人类社会的知识仍然在不断增长，人类社会仍然在不断向前发展。我们对于这个世界，对于我们自己的认识也越来越深入。同样，从精确的数学意义上讲，自然语言、日常语言都是比较模糊的，是有缺陷的，但不妨碍我们进行思想交流。我们没有办法用代数的方法来解决一个原本用代数解决不了的问题，可能需要用微积分才能解决。我们可能也需要反思现有的语义研究理论和方法，是不是语义的这种不可分割、不能量化的特点，使得我们需要采用不能分割、不能量化的方法来研究。

郑国锋：既有的语义研究有一个现象值得注意：即很多时候把精力花在了对语义的猜测上。以汉语研究为例，认知语言学家认为汉语使用者经常采用总体扫描的模式生产语义，既抽象，又难以操作。请问这样的语义研究问题在哪里？应该如何避免？

刘海涛：关于语义研究的方法，我们以上已经谈及一些，主要的问题是多数建立在自圆其说的基础上和对个别词语的分析上。这种分析的弊端在于如果你采用一种方法分析了一个词，又采取另外一种方法分析另外一个词，刚才的分析又要调整。这样即使分析了100个词，可能也发现不了如何更好地掌握某种词义的规律。一个人可能通过研究100个词写了一些文章，但这样的规律别人能使用吗？计算语言学家可以在计算机上实现这样的规律吗？

索绪尔在《教程》里还有这样一句著名的话：语言既然是一个系统，它的各项要素都有连带关系，而且其中每项要素的价值都只是因为有其他各项要素同时存在的结果。这意味着我们研究一个词的意思，还要观察它与同时期的其他词的关系。在研究过程中，在一个共时的横切面的研究里，语言学家很多时候会把很多词语也拉进来。比如，在猜测、分解或区分词义时，语言学家认为使用的10种或更多的知识已经足够。但当我们把这10种知识(如词的搭配特征)告诉或输入到一个完全不懂这个词的人或计算机内，就会发现这个人或计算机对这个词的理解不如预期，这是因为人类实际上没有能力靠内省来完全剥离自己在做一件事时用到的知识。

可以看出，用内省法做出来的研究，只是找到了处理问题所用到的部分知识，但是可能遗漏了关键的或更重要的东西，这导致发现的规律不能被别人使用，或者别人使用的效果不好。以义素分析法为例，它实际上是通过逻辑的思路，使用可量化、可分解的语义方法处理语言材料。研究者认为这样可以弄清

语义，但事实上只是可以弄清一部分，而不是全部，当这样的发现放在别的地方时，就不好用了。

维特根斯坦(2005:150)在《哲学研究》中认为词义是词汇在语言中的用法，符号自身是死的，是在使用中有了生命，注入了生命的气息。要掌握一个词的意思，就要尽量熟悉这个词的用法，观察它的使用。A 看了一个词的 10 种用法，B 看了 100 种用法，C 看了 1 万种用法，那么见过这个词 1 万种用法的那个人，就比其他人更好地掌握这个词的用法。这意味着一个人对词使用的场景知道得越多，对词义的把握就越准确，这是哲学家关于词义研究的转向。但是很多语言学家似乎还没有意识到这个问题，还在采用分解的方法去切割不能切割的东西。而从事自然语言处理的研究者、计算语言学家，之所以今天他们的自然语言处理系统和计算语言学应用能更好、更有效地处理语言，就是因为他们应用了从大规模的真实语料中学习词的用法。可以说今天这些自然语言处理、计算语言学应用的成功之处就在于他们比前人更好地实现了维特根斯坦的用法论，他们的做法使得计算机比人都更懂、更能精确地把握词的用法。从他们的成功中，我们做语义理论研究的人，能得到什么启示呢？

郑国锋：提及词义关系，我们谈谈依存距离。它指的是构成依存关系的支配词和从属词之间的线性距离，这是判断词义关系的重要标准。您还提出“依存距离最小化”是人类语言的一种普遍倾向(Liu *et al*. 2017)。请问依存距离词义研究与传统词义研究有何不同？能够对我们的语言使用做出更合理的解释吗？

刘海涛：首先，我们如果要更好地掌握一个词的意思，就要把握它和其他词的关系。如上所述，索绪尔说一个语言单位的价值存在于它和其他单位之间的关系。维特根斯坦和索绪尔的说法一脉相承。从句法、句子形式的角度看，需要通过更多它与其他词语之间的形式关系来掌握词语的句法特征。一个词表达的意思是有限的，如果一个人要给别人传递一条复杂的信息，自然就要把更多的词组织在一起，就会出现这些词语的关系问题。例如，我们要传达一条信息的时候，大脑会检索到 8 个或者 10 个词，这实际上是在大脑里构成了一个语义网络结构。由于人类的生理限制，我们要把二维的语义网络结构转换成线性结构，转换出来就是一个句子。句子中的实词从二维结构转变为线性结构时会有先后顺序，这是有规律的，这就是语法或者句法。这个过程中每种语言的手段不一，比如有的语言靠固定的语序，也就是一个实词在这里是一种句法功能，换个位置就是另外一个句法功能，前后不能颠倒。

如果我们要表达的思想比较复杂，就需要很多实词，那么仅靠语序可能会混乱，让接收者难以解码，这时就需要有虚词。虚词表达功能意义，就是连接，在一个二维结构转变为一维线性结构时，实词之间的关系可以因此更清晰。当然也有语言不通过语序变化告诉接收者词在句子中的句法功能，而是通过词本身的变化，例如词尾的形态变化。总之，当信息从说话者的大脑里传达出来或

者写出来时，由于二维结构要转变成线性结构，要么用语序加虚词，要么用词汇本身的形态变化来告诉接收者这个词的句法功能。我们的问题是：说话者在组合这些词语时，是否遵循某种普遍规律。比如，1 个句子中含有 10 个实词，由于语言不同，为了更清楚地表达说话者的信息，10 个词有可能变成 12 个，增加 2 个虚词。无论是 10 个词还是 12 个词，不管使用的是虚词加语序的手段实现线性化，还是采用词形变化实现线性化，我们认为背后有一个普遍规律。因为所有人在将二维语义表达线性化的过程中，都要受到自身认知的约束。假如两个词在我们的大脑中是密切相关的，在组合它们的时候，受自身约束，我们不可能在两个词之间插入很多词。否则，受工作记忆容量的约束，我们就记不住这些关系。所以在组合这些词时，大部分情况下我们会把它们安排的比较接近。而且，看到或听到这个句子的人，也同样受工作记忆的约束和其他认知约束。这种想法早已有之，1909 年德国学者巴哈格尔发表过一篇文章，从德语、拉丁语等几种语言版本中，找到了实际的语例，但因为时代的限制，他主要是从一些真实的文本语料中寻找某些语言的句子进行观察。他发现大概有这样的规律：语义上比较接近的词语，在句子线性序列里离得也比较近。这符合我们刚才的假设，但是他挑选的是个别的例子，也不是连续的语料，而且涉及的语言也不多。但是他至少在 100 多年前就发现并用多种语言的语例验证了这样的假设。

郑国锋：的确如此，以往受制于技术手段和硬件设备，很多想法无法得到实现。在大数据时代，我们甚至可以计算出语言平均数，这对上一个问题中提及的论题的解决有什么促进吗？

刘海涛：由于大量语料库和计算技术的出现，我们可以通过依存句法标注语料库，采用更多语言、更大规模的真实语料来验证这样的假设。我们的研究结果发表在 2008 年，用了 20 种语言的真实语料，还做了两种随机的、不符合语法的语言，即人造语言，以便比较。我们发现，无论这 20 种人类语言的类型如何，确实有依存距离最小化的倾向。我们称之为“倾向”，没有说是规律，因为语言的特点就是概率性，即大多数的句子都有最小化或者优化的倾向，但是我们不能叫最优化。最优化是所有有关的词语都挨着，多数情况下不可能，比如一个句子有 30 个词，这些词语不可能都挨着。但正是这种不可能，才能够看出这种最小化的倾向，因为虽然词汇做不到都挨着，但会想办法尽可能地挨着。依存距离最小化的发现证明认知结构、认知机制在一定程度上决定了语法的结构，而这个语法结构是一种普遍现象。这个研究用真实的数据打通语言的普遍性和语言的认知，可能是人类语言的一个真实的规律。这篇文章发表 12 年来，得到了包括计算语言学界、自然科学界在内的学者们的认可，还被最好的语言学杂志 *Language* 中的文章引用(Futrell *et al*. 2020)。这说明语言研究的范式到了要转变的时候。

郑国锋：大数据背景下，社会中流动的语言数据越来越多，这为语言学理论

研究提供的既是“沃土”,也是“挑战”。假如我们依旧采用过往的研究范式,语言学理论研究将会面临越来越多的挑战。据我对您官网的统计,您的研究呈现出一个鲜明的特征:从 2019 年 3 月份以来您已经完成了 33 场学术报告或座谈,这其中 14 场与大数据或人工智能有关。这是否意味着数据驱动的语言研究或语言人工智能研究的时代已经到来?

刘海涛:根据前面的讨论,我们可以说数据驱动的语言研究或者语言人工智能研究的时代已经到来。重要的是,就像惠特曼教授与你讨论时指出的,有良知的语言学家已经意识到,我们正面临来自计算语言学家的挑战。尽管计算语言学家可能不清楚他们为什么这么做就成功了,但作为这个世界上从理论上讲最懂语言的语言学家,我们有义务、有责任回答计算语言学家的疑惑:为什么我这样可以,按照你们说的办法却不行。

你提到“沃土”,“沃土”是人家已经做成了,人家庄稼的长势比用你传授的耕作方法要长得更好。同样一块土地,用你的方法,他发现庄稼长得不好;而用他自己的方法,产量和质量都更好。在这种情况下,语言学家作为本来最懂土壤、最懂种子的“庄稼汉”,不能只说自己的方法好,别人的方法有问题。我们不能只挑几颗奇怪的种子,例如用“我一把把把把住”这样的句子来挑刺。因为这是一般情况下正常人不会说的话。当大部分你不熟悉的种子已经长势良好的时候,你不应该找一些奇怪的种子去否认这些进步。我们现在要反思,过去我们采用完全形式、内省的方法,是不是遗漏了语言系统最本质的某些特点,因而发现的规律有可能没有反映,或者没有完全反映语言系统的运行规律,而这导致需要使用规律的学者用到这些规律时做不成事情。对于一个动态自适应的概率系统来说,用定性的、一成不变的标准来检测这些产品,当然会出问题。

郑国锋:为什么要强调数据驱动的语言研究和人工智能的关系?语言学理论研究应该如何适应这样的全新场景?

刘海涛:一本大数据的书里提到(Sarangi & Sharma 2020),与人工智能有关的学科包括计算机、数学、医学、心理学、工程,语言学也名列其中。一个研究对象重要,研究这个对象的领域并不一定也重要,领域重要的意义在于它的成果能够被别人使用,但现在理论语言学的成果,别人使用不了。人工智能在 50 年代开始出现,但作为一个学科,只是这几年才让大家觉得机器产生的智能行为能够与人类抗衡。比如,人类在围棋或其他棋类游戏中已经无法战胜机器,机器从人类那里已经学习不到下围棋的新技能了。

正在进行的人工智能革命不是基于规则的,也不是将通过内省法提取出来的知识再输到计算机里面,让计算机具有这种知识,然后再用这种知识解决问题。这是人工智能领域过去的做法,也是当前主流语言理论研究的做法,目前的人工智能的知识基础是从大量现实世界的数据里提取出来的。

这对语言学理论研究而言意义重大。语言数据就是人使用语言的数据,就

是把很多人说的话收集到一起。假如收集 10 亿说普通话的中国人的语料，我们就能得到有史以来最好的汉语普通话的平均数，那个时候就可以说没有人比机器更懂普通话。但是作为语言学家，我们如果不用真实语言中的语料去发现规律，我们对这个世界的进展将一无所知，也帮助不了别人。我们发现另外一个领域的人对我们最熟悉的对象懂得更多；他可能也懂得不多，但是他做出了更好的产品，他需要我们的解释，但是我们却脱离了时代，解释不了。100 多年前，索绪尔提出了语言平均数的概念，但是他计算不了。现在已经有计算平均数的可能了，我们可能需要研究的是在这个平均数的计算过程中，语言的规律是怎么涌现出来的。只有这样，当我们的研究对象和别人的研究对象一样的时候，我们发现的语言规律才能够服务于那些想解决真实世界的人类语言问题的领域（刘海涛 2021）。这就是今天我们所说的要开展数据驱动研究，实际上是大规模基于语言使用的语言理论研究的主要原因。

郑国锋：语言学理论研究曾经为政治、经济、哲学等领域的研究提供了重要的分析思路和研究范式，而如今的语言学理论研究却逐渐边缘化，语言学家们更多的是引用符号学、心理学、计算技术等来打造自己的研究阵地，语言本体研究理论则持续衰减，更不用说输出理论与研究范式，造成这一现象的原因是什么呢？

刘海涛：曾经的语言学理论研究影响了很多学科，但是社会是变化的，人对于社会、世界、自身的认识也在变化，而语言学这么多年来的变化并不大。乔姆斯基在 20 世纪 50 年代提出了先进的形式语法理论，甚至催生了计算机程序设计语言。那是一个语言学家的时代，但是时代在变化，层出不穷的新方法，特别是计算机的出现使我们能够处理过去那些我们想都不敢想的问题。我们甚至都可以算出言语平均数来，而这个平均数能更逼近真实的人类语言系统。在这种情况下，主流语言学理论研究依旧从静态的、形式的、符号的、逻辑的角度去研究语言现象，而所研究的对象，大多是靠内省法在大脑里培育出来的句子。这样产生出来的规律，我们不能说没有任何价值，但是对于那些需要处理真实语言、需要把握一个语言系统运行的真实规律的学科和学者来讲，这些规律已被实践证明用处有限。在这种情况下，他们不得不自己想办法，像语言教学、语言习得、语言规划等应用语言学研究的主要领域，都开始另起炉灶。这也可能是应用语言学家或其他领域的学者对所谓的理论语言学家敬而远之的原因之一。

Jelinek 说，解雇语言学家后，计算机处理语言的效果会更好一些（冯志伟 2018），原因是语言学家没有用处。作为一个基础学科，语言学家有权在研究的过程中不考虑用处，坚守自己的领地，但那就需要重新定义语言学，就不宜说语言学探求的是人类语言系统的运作规律。通过内省的方法构拟出来的各种规律属于个人，而不是我们所研究的语言作为一个系统的运作规律。因此，我们

要反思，要回到我们的初心：语言学是研究语言结构和演化规律的学科。如果我们努力的目标是把“学科”变成“科学”，就应当采用科学家通常使用的科学方法来研究语言。如果想让我们发现的规律被那些需要解决实际语言使用问题的学科或学者使用，我们的研究对象也应该是那些人类实际使用的语言。科姆利(2010：vii)在《语言类型学和语言普遍性》第一版序言中说：“语言学家研究语言，而语言是民众实际所讲的语言。”现实是残酷的，只有像维特根斯坦一样回到我们的日常语言，从大量的语言使用数据中寻找言语的平均数，并采用与时俱进的方法探求这个平均数产生的机理，语言学才有可能成为一个受人尊敬的学科，因为这个平均数可能才是人工智能时代所期待的语言规律。

郑国锋：现在智能化的浪潮席卷了社会的每个角落，面对喷薄而出的语言资源，大数据时代的语言学理论研究，应该如何因应这样的时代机遇，走上科学化与国际化的道路，早日成为一门显学？

刘海涛：讨论至此，这个问题的答案已经呼之欲出。首先，我们应该承认，语言学如果是一门科学，它的研究方法或者研究问题，尤其是研究方法应该与我们今天公认的其他科学领域，特别是实证科学领域不能有太大的不同。正如我们讨论的，我们不能死守着索绪尔及其之后的主张，比如语言和言语的区分，然后心安理得地只研究“我一把把把把住”这样的句子。德国人研究德国的“把”，英国人研究英国的“把”，我们最后就能“把住”语言的本质吗？我们应该梳理一下这么多年来语言学取得了哪些进步，我们是不是与时俱进地使用了这个时代的资源或者方法。语言是一个符号系统，语言也是一个复杂系统，还是一个复杂适应系统，更是一个人驱复杂适应系统。无论如何，语言是一个系统，这个世界上最懂系统的是系统科学家，所以我们需要采用系统科学家提出来的方法研究语言，只有这样发现的语言规律和有关语言的知识才可能是更可靠的科学知识。徐烈炯(1988：2)说学语言是造福自己，学语言学是造福人类。这意味着学语言使自己获得了一种能力，但是语言学研究发现的是知识，这个知识可以造福人类。但什么是知识，科学家有公认的标准，这要求我们采用一般科学家认可的方法研究语言，而不是我们自己虽然声称语言学是科学，但采用的方法却如此与众不同。科学家们研究蚂蚁，是通过观察和其他技术手段研究蚂蚁。而我们自己呢？则是变成蚂蚁，不仅如此，研究上海蚂蚁，还要变成上海蚂蚁，研究纽约蚂蚁，又得变成纽约蚂蚁。只有采用科学的方法，才能产生科学的成果，这个成果才能转化为一种知识。语言学要与时俱进，要想成为一个像格林伯格(Greenberg 1973)好多年前期待的领先的或者引领性的科学，首先我们要做的是让它成为科学，而成为科学，就必须采用科学的方法。

第二，我们今天在这里所讨论的语言学是一种普遍的语言学研究，或普通语言学的领域，其目的是发现人类语言系统的运作规律，这决定了它的研究价值和意义不仅仅限于所研究的语种。这意味着，中国人在这个领域所取得的研

究成果，也有必要让全世界都知道，这就是成果的国际化。关于这个问题，我在2018年第1期《语言战略研究》上专门写过文章，讨论中国语言学的国际化和语言研究的科学化。这是相互关联的两个问题，只有用科学的方法研究出来的语言学成果，才有可能更容易被更广泛的读者和同行接受。“百年未有之大变局”，也为语言学理论研究创造了前所未有的机遇。如果我们没有把握好这次机遇，那么在未来的50年里，我们在语言学领域将继续落后于世界。但现在机会就在面前，我们是继续跟在那些外国人的后面，给他们添加几个汉语的例证，还是回归语言研究者的初心，基于真实的语言材料，发现语言系统真正的运作规律？这是摆在全体中国语言学家面前的紧迫任务，也可能是未来五六十年里，中国语言学唯一一次超越或者引领世界语言学的机会。

参考文献

Behaghel, O. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern [J]. *Indogermanische Forschungen* 25:110-142.

Eco, U. 1995. *The Search for the Perfect Language* [M]. Oxford: Blackwell.

Futrell, R., O. L. Roger & E. Gibson. 2020. Dependency locality as an explanatory principle for word order [J]. *Language* 96 (2): 371-412.

Greenberg, J. H. 1973. Linguistics as a pilot science [A]. In E. P. Hamp (ed.). *Themes in Linguistics: The 1970s* [C]. The Hague: Mouton. 45-60.

Liu, H. 2008. Dependency distance as a metric of language comprehension difficulty [J]. *J Cogn Sci* 9(2):159-191.

Liu, H. 2014. Language is more a human-driven system than a semiotic system [J]. *Physics of Life Reviews* 11(2): 309-310.

Liu, H. 2018. Language as a human-driven complex adaptive system [J]. *Physics of Life Reviews* 26-27 (6) :149-151.

Liu, H., C. Xu & J. Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages [J]. *Physics of Life Reviews* 21(7): 171-193.

Sarangi, S. & P. Sharma. 2020. *BIG DATA: A Beginner's Introduction* [M]. London & New York: Routledge.

Sullivan, L. H. 1896. The tall office building artistically considered [J]. *Lippincott's* 28 (4): 403-409.

伯纳德·科姆里. 2010. 语言共性和语言类型(第二版)(沈家煊、罗天华译)[M]. 北京:北京大学出版社.

费尔迪南·德·索绪尔. 1999. 普通语言学教程(高名凯译)[M]. 北京:商务印书馆.

冯志伟. 2018. 信息时代需要文理兼通的语言学家[N/OL]. 光明网. [2018-10-21]. https://epaper.gmw.cn/gmrb/html/2018-10/21/nw.D110000gmrb_20181021_3-12.htm.

李宇明. 2018. 语言学是一个学科群[J]. 语言战略研究(1):15-24.

刘海涛、林燕妮. 2018. 大数据时代语言研究的方法和趋向[J]. 新疆师范大学学报(哲学社会科学版)(1):72-83.

(下转第31页)