《现代语文》，2007 年第 9 期（总第 282 期），p127-128

我与计算语言学的缘分

冯志伟

计算语言学是一个横跨语言学、数学和计算机科学的交叉学科。语言学和数学都是有着相当长历史的古老学科。语言学历来被看作是典型的人文科学，数学则被许多人看作是最重要的自然科学。在学校教育中，语文和数学被认为是两门最基础的学科，成为任何一个受教育者的必修课。它们似乎成了学校教育的两个极点：一个极点是作为文科代表者的语文，另一个极点是作为理科代表者的数学。很少有人想到，这两门表面上如此不同的学科之间竟然会存在着深刻的学术联系。计算机科学是研究计算机的新兴学科，带有相当强的工程性质和浓烈的技术色彩，属于高科技的范畴，表面上看来，作为典型的人文科学的语言学与属于高科技的计算机科学也不会存在什么瓜葛。因此，在一般人的心目中，这门横跨语言学、数学和计算机科学的计算语言学似乎是不可思议的，甚至是有悖于常识的，就是连“计算语言学”这个名称也几乎是荒谬绝伦的。

然而，一些具有远见卓识的学者却早就洞察了语言学、数学和计算机科学之间的紧密联系。1847 年，俄国数学家布里亚柯夫斯基(В．Я．Буляковский)提出了用概率方法来进行语法、词源及语言历史比较研究的思想。1894 年，瑞士语言学家索绪尔(De Saussure)指出，“在基本性质方面，语言中的量和量之间的关系可以用数学公式有规律地表达出来”，后来，他在其名著《普通语言学教程》（1916 年）中又指出，语言学好比一个几何系统，“它可以归结为一些待证的定理”。1904 年，波兰语言学家博杜恩.德.库尔特内(Baudouin de Courtenay)认为，语言学家不仅应该掌握初等数学，而且还有必要掌握高等数学 。他表示坚信，语言学将日益接近精密科学，语言学将根据数学的模式，一方面“更多地扩展量的概念”，一方面“将发展新的演绎思想的方法”。1933 年，美国语言学家布龙菲尔德(L. Bloomfield)提出了一个著名的论点：“数学只不过是语言所能到达的最高境界”。法国数学家阿达玛(J. Hadamard)说得更好：“语言学是数学和人文科学之间的桥梁”，他一语道破了语言学和数学之间的联系，并清楚地看出了语言学是人文科学中最容易与数学建立联系的学科。

著名俄国数学家马尔可夫(A. A. Марков)更是躬行实践，他在 1913 年把普希金的叙事长诗《欧根·奥涅金》中的连续字母加以分类，研究俄语字母序列内部的关系，提出了马尔可夫随机过程论，后来成为了一个独立的数学分支，对现代数学的发展产生了深远的影响。语言结构中蕴藏着的数学规律，成为了马尔可夫创造性思想的取之不尽的源泉。

1946 年第一台电子计算机 ENIAC 在美国研制成功。就在电子计算机问世的同一年，英国工程师布斯（A.D.Booth）和美国洛克菲勒基金会副总裁韦弗（W.Weaver）在讨论电子计算机的应用范围时，就提出了利用计算机进行语言自动翻译的想法。韦弗在 1947 年 3 月 4 日给控制论学者维纳(N. Wiener）的信中说：“我怀疑是否真的建造不出一部能够作翻译的计算机？即使只能翻译科学性的文章（在语义上问题较少），或是翻译出来的结果不怎么优雅（但能够理解），对我而言都值得一试。”可见，电子计算机一出现，计算机科学家的慧眼就投到了自然语言的自动处理方面。

这样看来，语言学、数学和计算机科学之间确实有着深刻的内在联系，那么，作为一门横跨语言学、数学和计算机科学的计算语言学的存在便应该是合情合理的了。

我是一个普通的凡人，当然不可能有上述学者那样的远见卓识。我是在一个偶然的机会与计算语言学结下了不解之缘的。这里，我愿意说一说我与计算语言学的这种缘分。

我于 1957 年高中毕业后，考入北京大学地球化学专业本科就读，当时我非常崇拜俄罗斯地球化学家费尔斯曼(Felsman)，一心想研究化学元素在地球上的分布规律。就当我在北京大学认真学习地球化学的前后，国外兴起了数理语言学，建立起了完善的理论和方法，并且在许多大学中开设了数理语言学的课程。数理语言学作为一个独立的学科出现在现代语言学的百花园中。在五十年代虽然还没有出现“计算语言学”这个名称，但是，数理语言学与后来出现的计算语言学有着密切的联系。北京大学高举五四“民主”和“科学”的大旗，学术空气非常自由，北京大学的图书馆藏书丰富，学生可以阅读到各种最新的科学杂志，了解到国内外最新的学术发展动向。当时我才十九岁，求知的愿望非常强烈，对于新事物极为敏感，我成为了北京大学图书馆的常客，整天泡在图书馆的书海之中。一个偶然的机会，我在北京大学图书馆馆藏的 1956 年出版的美国《信息论》 (IRE Transaction, Information Theory)杂志上，读到了美国语言学家乔姆斯基(N. Chomsky)的论文《语言描写的三个模型》(Three models for the description of language)，被乔姆斯基在语言研究中的新思想深深地吸引了。乔姆斯基在他的文章中，提出了形式语言和形式文法的新概念，他把自然语言和计算机程序设计语言置于相同的平面上，用统一的数学方法进行解释和定义，提出了语言描写的三个模型。用数学方法描写的这三个模型是这样地抽象，它们既可以用于描写自然语言，又可以描写计算机程序设计语言。我预感到这种语言的数学描写方法，将会把自然语言和程序设计语言紧密地结合起来，在信息的处理和研究中发挥出巨大的威力。于是，我下决心来研究数学方法在语言中的应用这个问题，开始了我从事计算语言学研究的美梦。

我的这个计算语言学的美梦一直做到现在，悠悠岁月，艰辛备尝。

1959 年，经领导批准，我从理科转到中文系语言学专业从事语言学的学习。转入语言学专业之后，我一面学好传统语言学的各门课程，一面利用课余时间，继续研究数理语言学的问题，我尽量充分地利用北京大学图书馆丰富的藏书和最新的杂志，跟踪着国际上数理语言学发展的足迹。1964 年我考上了语言学理论的研究生，经导师同意，我的研究生毕业论文的题目定为《数学方法在语言学中的应用》，在我国语言学研究中，首次系统地、全面地来研究数理语言学这个新兴学科。

北京大学中文系的著名语言学家王力先生和朱德熙先生都支持我的数理语言学研究，王力先生对我说：“语言学不是很简单的学问，我们应该像赵元任先生那样，首先做一个数学家、物理学家，然后再做一个合格的语言学家。”朱德熙先生对我说：“数学和语言学的研究都需要有逻辑抽象的能力，在这一方面，数学和语言学有共同性。”北京大学的这些第一流的学者，总是站在科学的最前沿来看待学术的发展，他们的鼓励给了我以巨大的力量。

可是，不久便发生了文化大革命，王力先生和朱德熙先生都被打成反动学术权威，我的数理语言学研究也随之失去了支持，我被分配到云南边疆的一所中学里教物理课。

在中学任教期间，我除了认认真真地教好学生，努力搞好本职工作外，还利用一切业余时间，密切地关注着国外学术发展的动向。数理语言学仍然像磁石一样强烈地吸引着我，在云南边疆那样闭塞的环境中，在信息不足、资料缺乏的困难条件下，我阅读了当时所能搜集到的各种关于数理语言学的资料，为了阅读散见于各种外文书刊中的数理语言学文献，我学会了英、法、德、俄、日等五种外国语，紧跟着世界上数理语言学发展的步伐。1973 年，我在云南省图书馆看到了美国语言自动处理咨询委员会(Automatic Language Processing Advisory Committee) 于 1966 年 11 月发表的《语言与机器》(Language and Machine)这个关于机器翻译的调查咨询报告，这个报告一方面对于机器翻译采取了消极悲观的态度，一方面强调了继续从计算角度研究自然语言规律的重要性，明确地提出了要研究“计算语言学”(computational linguistics)。这是我第一次接触到“计算语言学”这个名称，从此以后，计算语言学便成为了我终身为之奋斗的事业。

后来我才知道，早在 1962 年美国就成立了计算语言学学会，每年召开一次年会，并且出版学术季刊《美国计算语言学杂志》（American Journal of Computational Linguistics），后改名为《国际计算语言学杂志》（International Journal of Computational Linguistics）。1965 年在美国纽约成立了国际计算语言学委员会（International Committee of Computational Linguistics，简称 ICCL），每两年召开一次国际会议，叫做 COLING。在四人帮横行的那些年代，我们与国际的学术几乎完全隔绝了，所以，我在 1973 年才第一次接触到“计算语言学”这个名称，估计我是中国第一个接触到这个名称的学者，而其他没有机会阅读到《语言与机器》这个报告的中国学者，还根本不知道世界上还存在着“计算语言学”这个学科。可是，1973 年我第一次接触到“计算语言学”这个名称时，离开美国计算语言学学会成立的时间，已经整整 11 个年头了。我们已经远远落后于美国，这是多么可惜呀！

粉碎四人帮之后，迎来了科学的春天，我有了归队的可能。但是，这次我归的队不是文科的队，而是理科的队。为了提高自己的数学和计算机科学的知识水平，我于 1978 年通过理科考试，考上了中国科学技术大学研究生院信息科学系的研究生，弃文学理，又开始了理科的学习，从云南边疆回到了北京。1979 年，《计算机科学》杂志创刊，我在《计算机科学》创刊号上发表了《形式语言理论》的长篇论文，用严格的数学表达方式向计算机科学界说明语言学中的形式化方法如何推动了当代计算机科学的发展，并且指出，在语言学研究中发展起来的形式语言理论事实上已经成为了当代计算机科学不可缺少的一块重要的理论基石，计算机科学绝不可忽视形式语言理论。这样，我便从语言学的领域跨入了计算机科学的领域，开始从计算机科学的角度来研究语言学问题。

不久，我被中国科学技术大学研究生院选送到法国格勒诺布尔理科医科大学应用数学研究所(IMAG)自动翻译中心(GETA)学习，师从当时国际计算语言学委员会主席、法国著名数学家沃古瓦(B. Vauquois)教授专门研究机器翻译和计算语言学问题。1981 年回国后，我在中国科学技术信息研究所计算中心从事机器翻译的研究工作，仍然是从计算机科学的角度来研究语言学问题。直到 1985 年 11 月，由于语言文字工作的需要，我又一次弃理从文，调入国家语言文字工作委员会语言文字应用研究所工作。这是在我第一次弃理学文 26 年之后的又一次改行，最后还是回到了语言学的队伍。这时我已经是 47 岁的人了，斗转星移，人事沧桑，其中的甘苦有谁知道呢？

现在我已经是白发苍苍的 68 岁的老人了，从 1957 年我开始做计算语言学研究之梦到现在，我从事计算语言学研究已经整整五十年了。在这五十年的时间里，我国的计算语言学已经取得了很多成就，机器翻译、信息自动检索、信息自动提取等都已经投入了实用，有的达到了国际水平，这是令人欣慰的。我希望学习语言学的年轻学子注意这门新兴的语言学科，努力学好语文、数学和外语，练好基本功，为将来从事计算语言学的研究打好基础，我相信你们一定会为我国计算语言学的发展做出新的贡献。