# 《俄罗斯计算语言学与机器翻译》

**（语文出版社，2009 年 8 月出版）**

## 中文序言

近年来，我国计算语言学开始注意引进和介绍国外先进的理论和技术，但是，这些引进和介绍主要是针对欧美等发达国家的，对于俄罗斯的介绍寥寥无几，至于引进那就更是微乎其微了。这是一件非常遗憾的事情。

俄罗斯是我国的友好邻邦，在计算语言学和机器翻译研究方面，俄罗斯的起步比我国早得多。

在计算语言学方面，早在 1913 年，俄罗斯著名数学家马尔可夫（А．А．МАРКОВ，1856-1922)就注意到语言符号出现概率之间的相互影响，他试图以语言符号的出现概率为实例，来研究随机过程的数学理论，提出了马尔可夫链的思想，他的这个开创性的成果用法文发表在俄罗斯皇家科学院的通报上[1]。后来马尔可夫的这一思想发展成为在计算语言学中广为使用的马尔可夫模型（Markov model），是当代计算语言学最重要的理论支柱之一。在用数学思想来研究语言的创新性研究中，我们甚至可以追溯到 19 世纪中叶，早在 1847 年的时候，俄罗斯数学家布良柯夫斯基（Б．Буляковский）就提出了用概率论方法来进行语法、词源和语言历史比较研究的卓越见解了。1958 年，苏联数学家库拉金娜（О．С．Кулагина）采用集合论描述了基本的语法概念，为机器翻译研究奠定了坚实的数学基础。

在机器翻译方面，1933 年，苏联发明家特洛扬斯基（П．П．Троянский）设计了用机械方法把一种语言翻译为另一种语言的机器，并在同年 9 月 5 日登记了他的发明。特洛扬斯基认为翻译可以分为三个阶段，第一个阶段由只懂源语言的编辑，将输入的原文分析成特定的逻辑形式，将带有屈折词尾的变形词还原成原形词，并分析出各个单词的句法功能，为此，他创造了一套逻辑分析符号。第二阶段是利用他的翻译机，把源语言的原形词和逻辑符号转换成目标语言的原形词和符号。第三阶段由只懂目标语言的编辑，把目标语言的原形词和符号转换成目标语言。特洛扬斯基认为，他的翻译机只能在第二阶段作为自动词典来使用。不过他相信，只要能够建造出一部专门处理逻辑分析过程的机器，总有一天，上述的整个翻译程序都能够用机器来实现。1939 年，特洛扬斯基在他的翻译机上增加了一个用“光元素”操作的存储装置；1941 年 5 月，这部实验性的翻译机已经可以运作；1948 年，他计划在此基础上研制一部“电子机械机”(electro-mechanical machine)。但是，由于当时苏联的科学家和语言学家对此反映十分冷淡，特洛扬斯基的翻译机没有得到支持，最后以失败

[1] A. A. Markov, Essai d’une recherche statistique sur le texte du roman “Ougene Onegin” illustrant la liaison des epreuve en chain, Bulletin de l’Academie Impériale des Sciences de St-Pétersbourg, 7, 153-162.

告终了。

1954 年美国乔治敦大学进行世界上第一次机器翻译试验，成功地把俄语自动地翻译成英语之后不久，在同一年出版的第 10 期苏联《数学》杂志上就马上刊登了苏联科学院科技信息所所长潘诺夫（Д • Панов）的文章，介绍了美国乔治敦大学俄英机器翻译试验的成就，这篇文章的发表有力地推动了苏联组建自己的科研队伍、开展机器翻译研究的进程。1955 年在别尔斯卡娅（И • Бельская）领导下的苏联科学院精密机械与计算机技术所机器翻译研究组进行了苏联的第一次机器翻译试验，将应用数学文本从英语译成俄语，这是世界上继美国之后进行的第二次机器翻译试验。我们可以说，在计算语言学和机器翻译研究方面，俄罗斯和前苏联的学者做了很多开创性的研究，取得了很大的成绩。

中华人民共和国成立后，俄罗斯和前苏联在计算语言学和机器翻译方面的研究成果传入我国，我国也开展了计算语言学和机器翻译的研究。1956 年，国家便把机器翻译和计算语言学的研究列入了我国科学工作的发展规划，成为其中的一个课题，课题的名称是：“机器翻译、自然语言翻译规则的建立和自然语言的数学理论”。1957 年，中国科学院语言研究所与计算技术研究所合作，开展俄汉机器翻译的研究。1959 年，他们在我国制造的 104 大型通用电子计算机上，进行了俄汉机器翻译试验，翻译了 9 个不同类型的、较为复杂的句子，建立了我国第一个机器翻译系统。我国第一个机器翻译系统是俄汉机器翻译系统，直接受到了俄罗斯和前苏联的计算语言学和机器翻译研究的影响，这个机器翻译系统的主要设计人刘涌泉教授就曾经到当时的苏联专门学习机器翻译的方法，掌握了机器翻译的关键技术，可以说，我国的计算语言学和机器翻译研究首先是在俄罗斯和前苏联计算语言学和机器翻译研究的影响下开始的。

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，一心想研究化学元素在地球上的分布规律。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，北京大学图书馆丰富的藏书吸引了我，我成为了图书馆的常客，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。与此同时，我还在北京大学图书馆中，看到了俄罗斯数学家库拉金娜（О • С • Кулагина）在 1958 年的《控制论问题》上发表的《在集合论基础上确定语法概念的一种方法》（Об одном способе определения грамматических п

онятий на базе теории множеств)[2]，库拉金娜采用集合论描述了基本的语法概念，提出了“族”（семейство）、“域”（окретность）、“构形”（конфигурация）等语法概念的数学模型，为机器翻译研究在计算语言学的基本理论方面奠定了坚实的数学基础，使我的眼界大开。库拉金娜的成功坚定了我使用数学方法来研究自然语言问题的决心，于是我在 1959 年毅然从理科转到中文系语言学专业从事语言学的学习，从此走上了研究计算语言学和机器翻译的道路。

在上世纪 50 年代和 60 年代，由于东西方的隔绝，在学术领域中，很难找到英文的文献，不过，当时要找到俄文文献并不十分困难，为了跟踪国外计算语言学和机器翻译的发展情况，俄文的文献帮了我的大忙，我以俄文的文献作为中介，间接地了解到不少的情况。在当时极为封闭的学术气氛下，俄文成为了我了解国外学术动态的最主要的语言工具，它就像一扇敞亮的窗子，使我有可能艰难地跟踪着国外学术的发展步伐。我从《语言学问题》（вопросы языкознания）、《控制论问题》（провлемы кибернетики）、《语言学中的新事物》（новое в лингвистике）等俄文文献中，了解到国外学术的最新进展，大大地拓广了我的学术视野。

在这个时期，我还阅读了列夫辛（Ревзин）的《语言模型》（модель языка）以及阿赫玛诺娃（О•С•Ахманова）、梅尔楚克（И•А•Мельчук）等的《语言研究中的精密方法(О точных методах исследования языка)》等专著，对于自然语言处理中使用的集合论、数理逻辑、概率论和数理统计等方法有了深入而系统的认识。俄罗斯和前苏联在计算语言学和机器翻译方面的研究以及他们通过俄语对于西方学术的介绍，成为了中国学者间接地了解西方计算语言学和机器翻译进展的一个重要的途径。

改革开放以来，我国打开了国门，很多青年学者掌握了英语，英语文献成为了我们了解国际计算语言学和机器翻译进展的重要来源。但是，由于青年学者中懂得俄语的人很少，不能阅读俄文的文献，我们对于俄罗斯和前苏联的计算语言学和机器翻译的介绍和了解非常不够，这是一个很大的缺憾！

《俄罗斯计算语言学和机器翻译》弥补了这个缺憾。本书全面地介绍了俄罗斯以及前苏联学者在计算语言学和以机器翻译为代表的自然语言信息处理系统领域取得的成就，包括基础理论探讨、语言词汇知识库的建造、形态自动分析、句法自动分析、语义自动分析、句法-语义一体化分析、文本自动处理、语料库语言学、机器翻译、信息检索系统、自动文摘和计算机辅助语言教学软件等实用系统的开发等内容，涉及到了计算语言学的各个方面，使我们对于俄罗斯计算语言学和机器翻译的研究以及他们取得的成就，获得一个鸟瞰式的认识。

俄语是一种典型的屈折语，形态变化非常丰富，名词有 6 个格以及单数与复数的变化，还有阳性、阴性和中性的差别，动词有人称（person）、时态（tense）和体（aspect）的变化，形容词不仅有性、数、格的变化，还有一般级、比较级、最高级的变化。与英语比起来，俄语的形态变化丰富得多，俄语如此丰富的形态变化，为充分揭示形态分析、句法分析、语义分析、篇章分析提供了极为充分的语言数据，可以用更加多样的语言数据，来检验基本上以

[2] О•С•Кулагина,Об одном способе определения грамматических понятий на базе теории множеств,《проблемы кибернетики》, вып1, стр 201-214, 1958.

英语的语言数据为基础而建立的自然语言处理的理论和方法，发现其不足，从而促进自然语言处理研究的发展。

本书是教育部人文社会科学重点研究基地第二批重大项目成果，是作者们多年研究心血的结晶。计算语言学和机器翻译是语言学、数学和计算机科学的交叉学科，而本书的作者都是俄语专家，他们的学术背景是语言学，几年来，他们努力进行知识更新的再学习，逐渐地熟悉了数学和计算机科学，现在，他们在数学和计算机科学方面，虽然还不是精研通达的内行，却也不是似懂非懂的外行，他们已经掌握了数学和计算机科学的基本知识，在知识更新的基础之上，他们把语言学与数学和计算机科学巧妙地结合起来，写出了这样的专著，这是难能可贵的，相信本书的作者们在探索计算语言学和机器翻译的道路上一定会取得更加出色的成绩。

冯志伟
于北京后拐棒胡同寓所
2008 年 8 月 8 日