

《当代外语研究》2011 年第 1 期



## 我与语言学割舍不断的缘分 My love to linguistics research

冯志伟  
(教育部语言文字应用研究所)

我是一个普通的语言学者，《当代外语研究》主编杨枫老师要我写一篇文章介绍自己的治学经验，我很愿意与广大读者交流自己学习和研究语言学的心得，因此就欣然同意了。在这里，我想讲一讲自己弃理学文、弃文从理，最后又弃理从文的曲折过程，谈一谈 50 多年来自己与语言学之间割舍不断的缘分。

### 弃理学文

我于 1939 年 4 月 15 日出生于云南昆明。1946 年考入昆明市长春路东升小学读书，1951 年以全昆明市会考第一名的高分考入昆明一中就读。昆明一中是云南省著名的重点学校，曾培育了无数的英才。获诺贝尔奖的著名物理学家杨振宁、著名哲学家艾思奇、著名出版家黄洛峰，等等，多年前都曾经是这个学校的学生。入学后，我下决心追赶这些曾经给昆明一中带来声誉的前辈的老校友，努力地学习，从初一到高三，我每年的总平均分都名列全校第一，成为了昆明一中的好学生。

1957 年高中毕业时，我以云南省理科第一名的成绩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。当时我的兴趣主要是在稀有元素上，

它们在元素周期表上是排在比较后的元素，是国家很需要的自然资源。我非常热爱地球化学专业，当时也没有任何想从事其他学科的想法，这个学科确实也很有意思。地球化学在上世纪 50 年代属于国家要重点发展的尖端学科之一，在地球科学里面，地球化学也是属于最先进的学科。

我在入学后曾经对于五光十色的矿物发生了浓厚的兴趣，研究这些矿物的晶体结构，如醉如痴地观察着不同结晶形状的各种矿物，六方晶系的金刚石、方斜晶系的石墨……，这些立体结构不同的矿物有着差异很大的物理和化学性质。我深深地被大自然的奥秘吸引住了。

就在我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，并且在一些大学中开设了数理语言学的课程，从而使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，日益芬芳灿烂。

1956 年，我国开始注意到国外数理语言学的兴起和发展，在我国科学研究的发展规划中，确立了名称叫做“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”的课题。这个课题包括两部分：一部分是机器翻译，另一部分是自然语言的数学理论，也就是今天我们所说的“数理语言学”（mathematical linguistics）。

一个偶然的时机使我了解到数理语言学这个新兴的语言学科。

1957 年冬天，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，无意中看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) 这篇文章，被乔姆斯基在语言研究中的新思想深深地吸引了。乔姆斯基追求语言描写的简单性原则，为了使用有限的手段描述变化无穷的自然语言，在他的文章中，建立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，达到了“有限手段的无限运用”的目标。

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。乔姆斯基当时未满 30 岁，还是一个名不见经传的青年语言学家。他的文章中闪耀着智慧的光芒，我完全被乔姆斯基的卓越智慧征服了。

经过反复考虑，我下决心来研究数学方法在语言中的应用这个问题，并经学校同意，我弃理学文，从理科转到中文系语言学专业从事语言学的学习。

### 胡耀邦鼓励我学习数理语言学

转入语言学专业之后，情况并不像我原来预想的那样顺利。

当时的中文系语言学专业要求学生大量的传统语言学课程，如“汉语史”、“文字学”、“音韵学”、“训诂学”等，根本没有开设任何与数理语言学有关系的课程，而我的志向是用数学方法研究语言，与学校的课程安排有很大的出入。因此，我一面要

学习这些传统语言学的课程，一面还要利用课余时间，继续研究我有兴趣的数理语言学问题，我需要同时在两条战线上作战，感到时间很不够用。我终日埋头读书，不怎么关心政治。尽管我努力学习学校规定的这些传统语言学课程，成绩总是名列前茅，而且还学会了4门外语，但是，同学们对于我这个理科转过来的学生仍然不理解，有的同学发现我能够解一些非常繁难的数学问题，感到十分奇怪。他们觉得，数学这样好的人居然改行来中文系学语言学，简直是南辕北辙！我在班上显得很孤立。

1961年秋天，团中央机关建立了这样一个制度：团中央书记处的每一位书记至少直接联系一个团支部，作为了解情况和结交青年朋友的一个渠道。1961年11月，北京市团市委为团中央第一书记胡耀邦选定北京大学59级语言专业团支部作为联系点。胡耀邦首先找这个班的团支部书记和宣传委员了解情况，问他们：“你们同学中有学习特别专心的吗？”他们回答介绍说：“我们班有个同学叫做冯志伟的学习特别好，他已经学了英语、俄语、德语和日语，而且达到相当水平，但是好像不特别关心政治”。胡耀邦表示，“我希望找冯志伟同学亲自谈一谈”。

团中央第一书记邀请的消息传给了我，我感到非常激动。1961年11月11日，北京大学团委安排我和其他4名同学一起到住在富强胡同的胡耀邦家做客。晚饭后我们乘公共汽车进城，当时北京的公交车数量严重不足，乘车的人很多，我们没有挤上从颐和园路过北大开往西直门的32路汽车，急中生智，干脆从北大乘车到起点站颐和园，再从颐和园乘车直奔北京市内，当我们赶到富强胡同时已经是晚上9点多钟了。胡耀邦还在一直等待着我们，他也等得有些着急了。

我们在会客室坐下，胡耀邦给我们每个同学递上了一个苹果，依次询问我们每个人的姓名、籍贯。

当胡耀邦问到我的时候，他说：“你就是那个学了4种外语的同学冯志伟吗？你学习那么努力，挨批了没有？”

我回答说：“其实我学习只是出于对语言学的兴趣，自己只是想多学点东西而已。”

当时的社会风气不主张学生学习外语，认为那是“崇洋媚外”，胡耀邦洞察秋毫，所以才一见面就关切地问我，“挨批了没有？”

我坦率地向胡耀邦汇报了自己的想法，讲述了自己学习数理语言学的动机和过程。

我向胡耀邦作了如下的自我介绍：

我原是昆明一中的学生，1957年考入北京大学地球化学专业学习，比同班同学早两年进入北大。1958年，我在一本英文的信息论杂志上，读到了一篇关于运用数学方法研究语言的文章，顿时灵感火花四溅，觉得这样的研究有可能为语言在计算机上的处理产生革命性的影响。我想，我的数学基础很好，何不投身到这个领域做进一步的探索？于是，我要求转到语言专业学习，在学校的支持下，我在1959年转入语言专业，一面学习语言学课程，一面学习数学，同时关注国际上运用数学方法研究语言问题的最新进展，当时，国际上把这样的研究叫做“数理语言学”。我对于外语的领悟比较灵敏，到1961年底的时候，已经学会了4门外语，而且能够使用这4种外语阅读数理语言学的外文文献了。由于我对于数理语言学有强烈的兴趣，数理语言学是交叉学科，我除了学好语言学的课程之外，还要自学数学和外语等不同的学科，时间比别的同学紧，没有很多的时间

间来关心政治。而当时学校的政治气氛特别浓，不太主张学生读书，我显得有些古怪：明明是学中文的文科学生，一有空就做些数学题，经常还读点外文书，这在当时是很不合拍的。所以，有的同学认为我是在走“只专不红”的道路，对我颇有微词；有的同学还说我是“孔子学生继承牛顿事业”，认为我的学习方向特别怪异。尽管我自己还没有受到批判，但是，思想压力很大，心里不十分痛快。

胡耀邦带着关注的神色耐心地听了我的这些介绍之后，正色地对我说：“事实将证明你的道路是正确的！”他的话斩钉截铁，掷地有声。

胡耀邦还严肃地回过头来对我们大家说：“外语学习是很重要的，我们需要对外交流，语言是很好的交流工具呀，懂了外语可以扩大眼界。”我们专心地聆听着，默默地思考着，会客室的气氛显得特别肃穆。

接着胡耀邦换了语气，开始和大家轻松地聊天。他告诉大家：“学生的主要任务是学习知识。我在高中的孩子写了篇作文，老师出题目说什么是学生的主要任务？我的孩子写道：学生的主要任务是提高政治水平。”他笑着对我们说，现在不少人对学生的主要任务的认识不很清楚，其实，道理很简单：“学生的主要任务是学习。”

谈话结束时已经很晚了。我们告别了胡耀邦，一路谈论着他的教导，总算赶上了末班车顺利地回到了北京大学。

从这次谈话后，我学习数理语言学更加理直气壮了。

1964年，我考上了北京大学理论语言学专业的研究生，我的毕业论文题目就是：《数学方法在语言研究中的应用》，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。

这样，我国的数理语言学研究便首先在北京大学正式地开展起来。现在媒体报道，北京大学的计算语言学研究是从1985年开始，恐怕与事实不符，我觉得似乎应当是从1964年开始的。

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。

王力先生曾对我说：“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、文学家、音乐家，然后再做一个合格的语言学家。”

朱德熙先生曾对我说：“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，他们的鼓励给了我巨大的力量。

但是这时候发生了一件事情，就是1966年的5月25日，第一张马列主义的大字报贴到了北大饭厅的门口。我记得很清楚那一天是5月25日，因为那一天我要去买一本法文词典，当时的《法汉词典》编得很不好，很简单，单词太少了。我学过日文，可以阅读日文文献，我的导师岑麒祥教授说：“你去买本《仏和词典》<sup>1</sup>吧！”，于是，我就到五道口的外文书店买了一本《仏和词典》。中午时分，我刚刚在五道口外文书店旁

---

<sup>1</sup> 《仏和词典》是《法日词典》的日语写法。

边的小饭馆吃完中饭回到北京大学，看到学校的大饭厅前人头攒动。我伸头一看，大饭厅前面的墙上贴着大字报呢。上面写着：“陆平、彭佩云你们要走往何方？”，言词很激烈，陆平是北大的校长，彭佩云是北大的党委书记，她现在是全国妇联的领导，他们俩当时被认为是北京市委的黑线人物，当时北京市长彭真已被揪出来了。我一看到大字报，就知道我正在准备答辩的毕业论文泡汤了，一场很大的革命就要来临了。

果然，过了几天《人民日报》就发表了社论说，“这是一张马列主义的大字报”，一下把火点起来了。北大进入“文化大革命”的混乱状态，王力先生和朱德熙先生等等，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，这个新兴学科的研究被这场“革命”扼杀在襁褓之中。我的数理语言学之梦破灭了。我弃理学文，意在用数学方法研究语言，现在，我既不能学理，也不能学文，我成为了所谓的“三品学生”<sup>2</sup>，随之离开了北京大学，到云南边疆的一所中学里当一名物理教员，又只好弃文从理了！

### 手工查频估测汉字熵值

在云南边疆当物理教员的这段是时间里，我除了认认真真地教好学生，努力搞好本职工作外，仍然利用一切业余时间，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，我设法利用业余时间，潜心研究数理语言学的问题，在信息不足、资料缺乏的困难条件下，阅读了我所能搜集到的各种关于数理语言学的资料，当时我已经掌握了英、法、德、俄、日等5种外国语，可以阅读了散见于各种外文书刊中的数理语言学文献，紧跟着世界上数理语言学发展的步伐。就在“读书无用论”甚嚣尘上的时候，我总结了当时国外数理语言学的成果，于1975年，以昆明五中教师的名义，写成了《数理语言学简介》的长篇文章，在重庆的一家自然科学杂志《计算机应用与应用数学》上发表，向国内计算机界和数学界详尽地介绍了数理语言学的最新情况，这一篇文章犹如空谷之足音，使当时被文化大革命封闭了世界学术进展的中国学术界了解到国外信息时代已经到来的最新动态。我在这篇文章中兴奋地告诉广大读者：“信息时代的到来，使得语言学、数学和计算机科学结下了不解之缘，语言研究和计算机技术已经到了非结合不可的地步了！”

在云南边疆的中学教物理学期间，我还有机会阅读了一些物理学的经典著作，例如，伽利略的《关于两个世界体系的对话》，牛顿的《自然哲学之数学原理》等。这些经典著作给了我很多启示。

伽利略认为，人们正在构建的理论体系是确实的真理，由于存在过多的因素和各种各样的事物，现象序列往往是对于真理的某种歪曲。所以，在科学研究中，最有意义的不是去考虑现象，而应当去寻求那些看起来确实能够给予人们深刻见解的原则。伽利略告诫人们，如果事实驳斥理论的话，那么，事实可能是错误的。伽利略忽视或无视那些有悖于理论的事实。伽利略举例说，人们看到每天太阳从东方升起，从西方落下，都误

<sup>2</sup> 当时把文革中找不到工作的大学生叫做“旧教育制度的牺牲品，新教育制度的实验品，社会上的处理品”，简称为“三品学生”。

以为太阳是围绕地球旋转的，而实际上却是地球围绕太阳旋转。因此，现象序列往往是对真理的某种歪曲，科学研究应当揭示那些隐藏在现象序列后面的真理，千万不要被表面的现象所迷惑。

牛顿认为，在他那个时代的科学水平下，世界本身还是不可理解的，科学研究所要做的最好的事情就是努力构建可以被理解的理论，牛顿关注的是理论的可理解性，而不是世界本身的可理解性，科学理论不是为了满足常识理解而构建的，常识和直觉不足以理解科学的理论。牛顿摒弃那些无助于理论构建的常识和直觉。

通过阅读这些博大精深的物理学经典著作，我认识到，在语言学研究中，我们应当探索和发现那些在语言事实和现象后面掩藏着本质和原则，不要只是总是停留在现象的观察和描写上，语言学研究的目的在于通过语言的现象揭示语言的本质。在这样的思想的启示之下，我下决心模仿 Shannon 研究英语字母的熵的做法，通过汉字频度的手工统计来探测隐藏在字频的表面现象之后掩藏着汉字的熵值(entropy)，也就是汉字中包含的信息量。从此，我利用业余时间潜心研究汉字熵值的测定问题。

汉字熵值的测定首先需要统计汉字的频度，通过频度来计算汉字的熵值。这显然是一个通过现象揭示本质的典型的科学问题，正好与伽利略和牛顿的科学方法不谋而合。

为了进行语言文字的信息处理，必须知道文字的信息量，因此，也就必须测定文字的熵。这是信息时代语言文字处理应该研究的基础性问题。汉字的“熵”是汉字所含信息量大小的数学度量，是汉字的一个重要的本质属性，一旦进入信息时代，我国必定要用计算机来处理汉字，首先就会遇到汉字信息量的问题。汉字熵的研究可以为汉字进入信息时代做好理论上的准备。

近几十年来，国外学者已陆续测出一些拼音文字字母中的熵，而汉字数量太大，各个汉字的出现概率各不相同，因此，要计算包含在一个汉字中的熵是一个十分复杂和繁难的问题。

为了计算汉字的熵，首先需要统计汉字在文本中的出现频度，由于上世纪 70 年代我们还没有机器可读的汉语语料库，哪怕小规模汉语语料库也没有，我是一个中学物理老师，也没有计算机，我只得根据书面文本进行手工查频，请了几个志同道合的朋友，用手工帮助我进行汉字频度的调查。我给这些朋友每个人发了一箱卡片，请他们帮助统计在选定样本资料中的汉字出现的频度，并且把这些频度记录在卡片上。在朋友们的帮助下，我用了将近 10 年的时间，对数百万字的现代汉语文本（占 70%）和古代汉语文本（占 30%）进行手工查频，从小到大地逐步扩大统计的规模，建立了 6 个不同容量的汉字频度表，最后根据这些不同的汉字频度表，逐步地扩大汉字的容量，终于计算出了汉字的熵。

通过汉字熵值的测定，我认识到了科学方法论的重要性，语言学研究不能总是停留在对于语言表面现象的描述上，而应当通过语言的表面现象深入地揭示语言的根本性的属性。汉字熵值的测定正好体现了这样的科学方法论原则：通过汉字频度的手工统计出来的数据来揭示隐藏在这些数据后面的汉字的信息量的大小 -- 汉字的熵值。

为了给汉字熵的测定建立一个坚实的理论基础，我还提出了“汉字容量极限定律”，我用数学方法证明：当统计样本中汉字的容量不大时，包含在一个汉字中的熵将随着汉

字容量的增加而增加，当统计样本中的汉字容量达到 12366 字时，包含在一个汉字中的熵就不再增加了，这意味着，在测定汉字的熵的时候，统计样本中汉字的容量是有极限的。这个极限值就是 12366 字，超出这个极限值，测出的汉字的熵再也不会增加了。在“汉字容量极限定律”的基础上，我在包含 12370 个不同汉字的统计样本的范围内，初步测出了在考虑语言符号出现概率差异的情况下，包含在一个汉字中的熵为 9.65 比特。由此得出结论：从汉语书面语总体来考虑，在现代汉语和古代汉语的全部汉语书面语中，包含在一个汉字中的熵是 9.65 比特。由于我采用的是手工查频的方法，尽管工作十分繁重，准确性还是难以得到保证，我一直认为，我测定出的汉字熵值只是一种初步的猜测，还需要更加精密的手段来进一步检验这样的猜测。

20 世纪 80 年代，我国北京航空学院计算机系刘源教授使用计算机统计汉字的频度，并计算出汉字的熵为 9.71 比特。刘源教授使用计算机计算的结果与我通过手工测定的结果相差不大，这说明我在 70 年代对于汉字熵的测定是十分认真的。

这项科学研究的结果说明，由于汉字的熵大于 8 比特，所以，汉字不能使用 8 比特的单字节编码，而要使用 16 比特的双字节编码。这项研究为汉字信息的计算机处理提供了基本的数据，对于汉字编码、汉字改革和汉语的规范化都有一定的指导意义。

汉字熵值的测定还使我更加深入地理解了通过表面现象揭示隐藏在现象后面的本质的科学研究方法。这些都是我认真地阅读伽利略和牛顿的物理学经典著作而得到的收获。

## 研制世界上第一个汉语到多种外语的机器翻译系统

粉碎四人帮之后，迎来了科学的春天。高等学校开始招生。毛泽东主席生前对于大学招生做过指示：“大学还是要办的”，但接着他又指示：“我这里主要说的是理工科大学还要办”。毛泽东在他的指示中没有说文科大学还要办。这样，大学招生时，首先恢复的是理工科大学招生，而文科没有招生。我渴望着早日回到科学研究的岗位上去，因此决定，既然文科不招生，那就报考理工科，于是，我报考了中国科学技术大学研究生院，毅然参加理工科大学的入学考试。1978 年，我通过了理科的入学考试，考上了中国科学技术大学研究生院，成为了这所全国一流的理工科大学的研究生。于是，我在弃理学文 20 年之后，又反过来弃文学理，重新开始了理科的学习，从云南边疆回到了北京。

在中国科学技术大学研究生院学习期间，我很快就在理工科的杂志上发表论文。1979 年，《计算机科学》杂志创刊，我就在该杂志创刊号上发表了《形式语言理论》的长篇文章，用严格的数学表达方式向计算机科学界说明数理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：在数理语言学研究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。许多人认为这篇文章一定是资深的计算机科学家写的，后来，当计算机界的一些专家了解到，这篇论文的作者竟然是文革前北京大学中文系的一个文科研究生的时候，感到非常惊讶。

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，专门研究自动翻译和数理语言学问题。

沃古瓦教授是国际计算语言学委员会的创始人，是当时国际计算语言学的领军人物，他领导的 GETA 在机器翻译的理论和实践上都做出了出色的成绩，我在 GETA 良好的学习环境中，可以了解到机器翻译发展的最新情况，可以学习到当代机器翻译最前沿的技术。我自幼就喜欢数学，而沃古瓦教授是数学家，我们一拍即合，都深知自然语言的形式理论对于构建机器翻译系统的重要性。从此，我的研究重点逐渐由数理语言学转到了计算语言学（computational linguistics）。

在法国留学期间，我的主要工作是进行汉语与不同外语的机器翻译研究。开始时，我使用的自然语言形式理论是乔姆斯基的短语结构语法（phrase structure grammar），我试图使用短语结构语法来进行汉语的自动分析。

早在 1957 年，我就接触到乔姆斯基的形式语言理论，对于乔姆斯基的理论是有深入了解的。乔姆斯基根据形式语法的原理，提出了短语结构语法来作自然语言形式描述的一种手段，这种语法在自然语言处理中得到了广泛的使用。国内外的许多机器翻译系统都采用乔姆斯基的短语结构语法作为系统设计的基本理论依据。根据乔姆斯基的短语结构语法，表示句子结构的树形图中的每一个结点只有一个相应的标记，结点与标记之间的这种关系是一种单值标记函数，会出现大量的歧义问题，难于区分句法结构相同而语义结构不同的汉语句子，这种分析法是短语结构语法在分析汉语时一个致命的缺点。

当时我在法国研制开发机器翻译系统的实践中，就更加具体地认识到短语结构语（mono-label function）的缺陷。这种单值标记函数表示的语言特征是十分有限的，因而在机器翻译中进行汉语的自动分析时会显得左支右绌。

有一天，沃古瓦教授和我讨论汉语自动分析的问题。我坦率地向沃古瓦教授说：“乔姆斯基的短语结构语法对于法语和英语的分析可能没有多大问题，可是，用这种语法来分析汉语，几乎寸步难行”。

沃古瓦教授用好奇的目光看着我，他希望我进一步阐述自己的看法。于是，我举例对沃古瓦教授作了如下的说明：

在汉语中可以说“点心吃了”，实际上是“点心被吃了”，但汉语一般不用“被”字；汉语中还可以说“张三吃了”，实际上是“张三把点心吃了”。“张三”是个名词短语 NP (Noun Phrase)，“点心”也是个 NP，“吃了”是个动词短语 VP (Verb Phrase)，这两个句子的规则都是： $S \rightarrow NP+VP$ ，其中，S(Sentence)表示句子，它们的层次相同，词序相同，词性也相同，但它们却有截然不同的含义，一个是被动句，一个是主动句。我们怎么来解释这样的差异呢？如果我们使用短语结构语法，用计算机来分析这两个不同的句子，计算机最后做出来的肯定是一样的树形图，它们的差别只是在叶子结点上的词不一样，整个树形图的上层都是同样的  $S \rightarrow NP+VP$ ，这样在结构上相同的句子为什么会有不同的语义解释，从而产生不同的含义？使用短语结构语法显然是解释不了的，而中文里到处都是这样的句子，因为中文里的被动关系有不同的表示方法，有时主动和被动在形式上没有明显的区别，可以从句子的上下文和意念上来加以区分。在这种进退两

难的局面下，唯一的出路就是根据汉语语法的特点来改进乔姆斯基的短语结构语法，设法使用一种新的方法来描述汉语。

沃古瓦教授耐心地听完了我的说明，他从沙发上站起来惊叹地说：“汉语真是一种 *langue terrible*（法语：糟糕的语言）”。他说：“哪种语言能够不分主动和被动，人吃了和被人吃了怎么能是一样？怎么这么乱？”

我向沃古瓦教授解释道：其实中国人一点儿也不感觉到乱，我们中国人在说话时是分辨得很清楚的，因为我们中国人知道，在一般的情况下，人是不能被吃的。所以“小王吃了”的语义不能是“小王被吃了”，而点心不吃东西，所以“点心吃了”必定是“点心被吃了”。汉语是靠词汇的固有语义来解决语法问题的，但是对于你们法国人来讲，并不存在这样的问题。所以，我们不能按照法语的思考方法来处理这个汉语的问题，我们必须另辟蹊径！

沃古瓦教授是一个知识广博、眼界开阔的学者，他鼓励我沿着这个思路继续探索。他对我说：“乔姆斯基的短语结构语法也不一定永远正确嘛！”

在我告别时，沃古瓦教授兴奋地说：“我相信，你一定能找出一种汉语自动分析的新方法。”

这次和沃古瓦教授的谈话使我深刻地认识到，乔姆斯基的短语结构语法在汉语自动分析时确实出现了极大的困难。这种困难甚至连沃古瓦教授这样世界第一流的计算语言学家也承认了。作为中国的科学工作者，我必须想出一种新的办法，来克服短语结构语法的缺点。不然，我现在进行的汉语自动分析就很难搞下去了。

这一天夜里我很不平静，翻来覆去总在思考这个问题。第二天清早，我走到沃古瓦教授的办公室，明确地向沃古瓦教授提出：我们正面临一个新的挑战，我们必须思考一种新的语法理论来解决这个问题。沃古瓦教授完全同意我的意见，他进一步鼓励我探索新的理论和方法来解决汉语自动分析中出现的这个困难问题。

在沃古瓦教授的鼓励下，我对这个问题反复进行了思考。我观察到：“小王吃了”和“点心吃了”这两个貌似相同的句子在词汇的语义上有很大的不同，“小王”在语义上是一个“人”，在一般情况下，“人”是“吃了”这个行为的主动者（agent），而“点心”在语义上是“食品”，在一般情况下，“食品”是“吃了”这个行为的被动者（patient），是“吃了”的对象。在短语结构规则  $S \rightarrow NP+VP$  中，如果我们不要把 NP 看成一个不可分割的单元，而把 NP 进一步加以分割，使用若干个特征来代替 NP 这个单一的特征。例如，在“小王吃了”中，我们把 NP 分解为“NP|人”两个特征，在“点心吃了”中，我们把 NP 分解为“NP|食品”两个特征，这样一来，就有可能在计算上把它们分解开来了。在计算机处理语言时，特征也就是“标记”，因此，我提出，如果我们使用“多标记”（multiple label）来代替短语结构语法中的“单标记”（mono label），就有可能大大地提高短语结构语法描述语言的能力，我们就可以使用改进后的这种语法来描述汉语，实现汉语的自动分析。这就是我关于“多标记”的设想。

我对于短语结构语法的另一个改进是使用多叉树代替短语结构语法的二叉树。乔姆斯基曾经提出乔姆斯基范式，他认为自然语言的结构具有二分特性，因此他主张在自然语言处理中使用“二叉树”（binary-tree）。我认为，在汉语中存在着“兼语式”和

“连动式”等特殊句式，它们都不具备二分的特性，因此，我主张使用“多叉树”来代替“二叉树”，从而提高短语结构语法描述汉语的能力。例如，“请小王吃饭”是一个兼语式的句子，其中的“小王”做前一个动词“请”的宾语，又做后一个动词“吃饭”的主语，在计算机处理时，究竟是分析为“请 / 小王吃饭”，还是“请小王 / 吃饭”，我们会感到举棋不定，处于进退维谷的境地，如果勉强分析，只会得到一棵交叉的分析树，违反了句法树的“非交特性”。如果我们采取三分，把这个句子分析为“请 / 小王 / 吃饭”，可以避免分析树的交叉，得到唯一的分析结果。

经过在计算机上编写程序进行潜心的钻研和反复的试验，我提出了“多又多标记树模型”（Multiple-labeled and Multiple-branched Tree Model，简称 MMT 模型），在 MMT 模型中，我采用多值标记函数（multiple-label function）来代替短语结构语法的单值标记函数（mono-label function），使得树形图中的一个结点，不再仅仅对应于一个标记，而是对应于若干个标记，我还使用多叉树来代替二叉树，这样便大大地提高了树形图的标记能力，使得树形图的各个结点上，都能记录足够多的语法语义信息，把句子中所蕴含的丰富多采的信息充分地表示出来，这种多值标记函数的理论，从根本上克服了乔姆斯基的短语结构语法在描述自然语言时的严重缺点，提高了其有限的分析能力，限制了其过强的生成能力。显而易见，MMT 模型是对乔姆斯基短语结构语法的一个带有实质意义的重要改进，这个模型提出后，立即引起了国际语言学界的高度重视，在 1982 年于布拉格召开的国际计算语言学会议(COLING'82) 上，在 1983 年于北京召开的国际中文信息处理会议(ICCIP'83)上，在 1984 年于香港召开的东南亚电脑会议(SEARCC'84)上，我都介绍了 MMT 模型。沃古瓦教授在国际计算语言学会议 COLING'82 的大会发言中，也满腔热情地赞扬了我的研究工作。

就在我提出 MMT 模型的同时，国外一些计算语言学家也看到了短语结构语法的局限性，分别提出了各种手段来改进它。例如 1983 年卡普兰(R. M. Kaplan)和布列斯南(J. Bresnan)提出的“词汇功能语法”、1983 年马丁·凯依(Martin Kay)提出的“功能合一语法”、1985 年盖兹达(G. Gazdar)等提出的“广义短语结构语法”、1985 年珀拉德(C. Pollard)提出的“中心语驱动的短语结构语法”等，都采用了“复杂特征”（complex features）来描述自然语言，他们所谓的“复杂特征”实际上也就是我提出的“多值标记”（multiple labels），名异而实同。所以，我当时提出的 MMT 模型，是全世界计算语言学者对乔姆斯基的短语结构语法进行改进的一个重要方面和不可分割的组成部分，MMT 模型是 20 世纪 80 年代较早提出的一个旨在改进短语结构语法的形式化模型，当时我国学者在这方面的研究在国际上是处于前沿地位的。

1984 年荷兰阿姆斯特丹北荷兰出版社出版的多卷专著《计算机科学基础研究》第 9 卷《自然语言处理的计算机模型》一书（由意大利米兰大学主编）中，曾详细介绍了 MMT 模型，并评论说：“冯氏关于独立分析—独立生成的主张，关于尽可能地从源语言分析中获取多方面信息的主张，是当前自然语言处理研究中的一个重要进展”。

我还结合汉语的特点需要，研究了采用 MMT 模型来解决汉语自动分析的各种问题。我认为，在汉语的自动分析中，采用“多值标记”的必要性更加明显。这是因为汉语的句子不能只用词类或词组类型等简单特征来描述，汉语句子各个成分的词类、词组类型、

句法功能、语义关系、逻辑关系之间，存在着极为错综复杂的关系，如果只采用简单特征，就无法区分各种歧义现象，达不到汉语自动处理的目的。具体地说，这是由于：1. 汉语句子中的词组类型（或词类）与句法功能之间不存在简单的一一对应关系；2. 汉语句子中词组类型（或词类）和句法功能相同的成分，它们与句子中其它成分的语义关系还可能不同，句法功能和语义关系之间也不是简单地一一对应的；3. 汉语中单词所固有的语法特征和语义特征，对于判别词组结构的性质，往往有很大的参考价值，除了词组类型这样的简单特征之外，再加上单词固有的语法特征和语义特征，采用多值标记来描述，就可以判断词组结构的性质。

我还提出了用于多值标记的汉语“特征—值”系统，特征可分为静态特征（static feature）和动态特征（dynamic feature）两大类。其中，静态特征有：词类特征、单词的固有语义特征和它的值、词的固有语法特征和它的值，动态特征有：词组类型特征和它的值、句法功能特征、语义关系特征、逻辑关系特征。在自动句法语义分析中，静态特征是计算机进行运算的基础，计算机依赖于这些预先在词典中给出的静态特征，通过有穷步骤的运算，逐渐计算出各种动态特征，从而逐步弄清楚汉语句子中各个语言成分之间的关系，达到句法语义分析的目的。这就是我的“双态理论”（bi-states theory）。

我在法国留学期间，了解到法国语言学家泰尼埃(L. Tesnière)的从属关系语法和语法“价”的概念，我用这种语法来研究汉外机器翻译问题，首次把“价”（valence）的概念引入我国的机器翻译研究中，我把动词和形容词的行动元（actant）分为主体者、对象者、受益者三个，把状态元（circonstant）分为时刻、时段、时间起点、时间终点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、目的、工具、范围、条件、作用、内容、论题、比较、伴随、程度、判断、陈述、附加、修饰等 27 个，以此来建立多语言的自动句法分析系统，对于一些表示观念、感情的名词，也分别给出了它们的价。我还把从属关系语法和短语结构语法结合起来，在表示结构关系的多叉多标记树形图中，明确地指出中心语的位置，并用核心(GOV)、枢轴(PIVOT)等结点来表示中心词。这是我国学者最早利用从属关系语法和配价语法来进行自然语言计算机处理的尝试。

我根据机器翻译的实践，提出了表示从属关系语法的从属树(Dependence Tree)应该满足如下 5 个条件：1. 单纯结点条件：从属树中，只有终极结点，没有非终极结点，从属树中的所有结点所代表的都是句子中实际出现的具体的单词；2. 单一父结点条件：在从属树中，除了根结点没有父结点之外，所有的结点都只有一个父结点；3. 独根结点条件：一个从属树只能有一个根结点，这个根结点，就是从属树中唯一没有父结点的结点，这个根结点支配着其他的所有的结点，4. 非交条件：从属树中的树枝不能彼此相交；5. 互斥条件：从属树中的结点之间，从上到下的支配关系和从左到右的前于关系之间是互相排斥的，如果两个结点之间存在着支配关系，它们之间就不能存在前于关系。我提出的这 5 个条件比 1970 年美国计算语言学家罗宾孙(J. Robinson)提出的从属关系语法的 4 条公理更加直观，更加便于在机器翻译中使用。

我在法国研究的另一个问题是生成语法的公理化方法。我从公理化方法的角度来研究乔姆斯基的形式文法，把乔姆斯基的形式文法同数学中的半图厄系统(semi-Thue

system)相比较,指出了乔姆斯基的形式文法,实际上是数学中的公理系统理论在语言分析中的一种应用,语言就是由文法这一公理系统从初始符号出发推导出的无限句子的集合;文法的规则是有限的,文法中的终极符号和非终极符号的数目也是有限的,可是,由于语言符号具有递归性,文法这一公理系统就能够根据有限的符号,通过有限的重写规则,递归地推导出无限的句子来。这样的研究,从数学的基础理论方面揭示了形式文法的实质。

根据 MMT 模型,我于 1981 年完成了汉-法/英/日/俄/德多语言机器翻译试验,建立了 F A J R A 系统(F A J R A 是法语、英语、日语、俄语、德语的法文首字母缩写)。在 IBM-4341 大型计算机上,把二十多篇汉语的文章自动地翻译成英文、法文、日文、俄文、德文。这是世界上第一个汉语到多种外语的机器翻译系统,开创了多语言机器翻译系统之先河。

我的研究从理论和实践上都改进了短语结构语法,受到了导师沃古瓦教授的赞赏。我急着想把这些成果应用到中国的科技信息文献的大规模翻译方面,建立一个实用的机器翻译系统,因此,实验报告一写完,我就马上告别沃古瓦教授,离开法国回到了祖国。

### 立志做文理兼通的语言学家

回到北京,我想到的第一件事情就是到北京大学拜见著名语言学家王力先生,向王力先生汇报我在法国学习的收获。早年在北京大学中文系开始研究数理语言学的时候,王力先生就支持过我的研究,在北京大学求学期间,我曾经认真地听过王力先生讲授的《古代汉语》《汉语史》《中国语言学史》《清代古音学》等课程,这些课程,为我后来的计算语言学研究奠定了坚实的基础,我永远忘不了我的恩师王力先生。

1982 年春天,为我和老同学吴坤定(现为北京出版社编审)一起到北京大学燕南园去看望王力先生,一进门,王力先生就高兴地请我们坐下,王力先生对我说:“听说你到法国之后已经改行学习自然科学了,现在,你有了很好的数理化基础,因此也就有了科学的头脑,这些都是很宝贵的财富,在语言学研究随时用得着”。我向王力教授汇报了自己在法国研究多语言机器翻译的收获。王力先生细心地听着,他对我说:“我前年在武汉开的中国语言学会成立大会上曾经说,我一辈子吃亏就吃亏在我不懂数理化。现在你懂得数理化,就不会像我这样吃亏了,我相信你今后一定会做出更好的成绩”。接着,王力先生又说:“20 多年前我曾经对你说过,我希望你学习赵元任先生。当然,这是很难的。赵元任先生由哲学家、物理学家、数学家、文学家、音乐家做底子,最后才成为世界著名的语言学家的。我一辈子都想学他,但是,我的数理化基础差,没有学好。你现在到法国学习了自然科学,已经具备学习赵元任先生的条件了,我再一次提醒你,你要向赵元任先生学习,而且一定要学得比我好”。王力先生这些语重心长的话,极大地鼓励了我,我决心按照王力先生的教导,把数理化的知识和语言学的知识结合起来,做一个信息时代的文理兼通的语言学家。

从法国回国之后,我在中国科技信息研究所计算中心担任机器翻译研究组的组长,在王力先生的鼓励之下,我利用当时北京遥感技术研究所的 IBM-4361 计算机,于 1985

年进行了德—汉机器翻译试验和法—汉机器翻译试验，建立了G C A T德—汉机器翻译系统和F C A T法—汉机器翻译系统，检验了M M T模型生成汉语的能力，试验结果良好。可惜当时由于国内的科研资金缺乏，不能提供足够的财力和人力来开展更大规模的实验，我要建立实用性机器翻译系统的愿望没有马上实现。

1982年秋天，我应北京大学的邀请，在北京大学中文系汉语专业开设了“语言学中的数学问题”的选修课。这是国内首次在高等学校全面地、系统地讲述数理语言学的课程，受到学生们的欢迎。北京大学前任校长、著名数学家丁石孙教授在他的专著《数学与教育》一书中，对这门课程作了如下的评价：“1982年，北京大学中文系开设了《语言学中的数学问题》，这是给汉语专业学生开的选修课程，许多同学对这门学科产生了很大的兴趣，经过一个学期的学习，同学们初步认识了现代数学的发展给语言学注入了生机，觉得获益匪浅，对语言学这门古老的学科分支的发展充满了信心，而且这一举动冲击了相当多的人的旧概念，使闭塞的中国学术界认识到，即使在人文科学教育中，数学也在逐渐起作用。”<sup>3</sup>

在北京大学讲稿的基础之上，我写出了我国第一部数理语言学的专著，书名就叫做《数理语言学》，于1985年8月由上海知识出版社出版。接着，我又出版了《自动翻译》的专著，深入地探讨自然语言机器翻译的理论和实践问题。这两本专著的出版，受到了我国计算语言学界的欢迎。不少出国学习计算语言学的留学生，出国时都带着这两本书，作为入门的向导。

## 研制世界上第一个中文术语数据库

1985年，原文字改革委员会改名为国家语言文字工作委员会，需要计算语言学方面的人材，我调入了国家语言文字工作委员会语言文字应用研究所担任计算语言学研究室主任，得以专门从事计算语言学的研究工作，这是我1978年弃文学理之后又一次弃理从文，我又重新回到了语言学的怀抱。与此同时，由于工作的需要，我还在中国科学院软件研究所担任兼职研究员的工作。

根据中德科技合作协定，我受中国科学院软件研究所的派遣，于1986年至1988年到德国夫琅禾费研究院新信息技术与通讯系统研究所(Fraunhofer Gesellschaft, 简称FhG)担任客座研究员，从事术语数据库的开发。

术语是人类科学技术知识在自然语言中的结晶。术语数据库是在计算机上建立的人类科学技术的知识库，这项研究属于知识工程的研究，具有重要的意义。

当时世界上还没有很好的汉字输入输出软件，我国自己开发的CCDOS还很不成熟，我克服了重重困难，在FhG使用UNIX操作系统和INGRES软件，建立了数据处理领域的中文术语数据库GLOT-C，并且把这个数据库与FhG的其他语言的术语数据库相连接，可以快速地进行多语言术语的查询和检索，而且还可以处理简繁体汉字。这是世界上第一个中文术语数据库，具有开创作用。

<sup>3</sup> 丁石孙，《数学与教育》，湖南教育出版社，长沙，1991年版。

在 FhG 研究术语数据库的过程中,我还接触到多种语言的大量术语,我惊异地发现,几乎在每一种语言中,词组型术语的数量都大大地超过了单词型术语的数量。根据多年前我学习过的伽利略和牛顿的科学方法论,我试图揭示出语言事实后面隐藏的本质,从理论上对这样的语言事实进行解释。

为此,我把数理语言学的理论应用到术语数据库的研究中,提出了“术语形成的经济律”。

我根据大量的实验数据证明了:在一个术语系统中,术语系统的经济指数与术语平均长度的乘积恰恰等于单词的术语构成频度之值,并提出了“F E L 公式”来描述这个定律。根据 F E L 公式可知,在一个术语系统中,提高术语系统经济指数的最好方法是在尽量不过大地改变术语平均长度的前提下,增加单词的术语构成频度。这样,在术语形成的过程中,将会产生大量的词组型术语,使得词组型术语的数量大大地超过单词型术语的数量,而成为术语系统中的大多数。F E L 公式从数理语言学的角度,正确地解释了为甚么术语系统中词组型术语的数目总是远远大于单词型术语的数目的数学机理,它反映了语言中的省力原则和经济原则,这是我国学者对于数理语言学中著名的齐夫定律 (Zipf's law) 的新发展,并从术语的角度说明了语言中的省力原则和经济原则是具有普遍意义的原则<sup>4</sup>。

“术语形成的经济律”提出之后,国内外的术语学研究者根据术语数据库的事实进行检验,检验证明,在各种语言的术语数据库中,词组型术语的数目都大于单词型术语的数目。因此,“术语形成的经济律”是适应于各种语言的一条普遍规律,是现代术语学的一条重要的基本定律。

语言是现实的编码体系,术语形成的经济律反映了用词作为语言材料进行单词型术语和词组型术语的编码时的经济律,这一经济律也可适用于语言编码的其他领域。汉语中在用单字组成多字词的时候,有限数目的单字组成了为数可观的多字词,多字词以增加自身的长度为代价来保持汉语中原有单字的个数或者尽量不增加原有单字的个数,体现了组字成词这个编码过程的经济律。多字词也就是双音词或多音词,著名语言学家吕叔湘先生指出,“北方话的语音面貌在最近几百年里没有多大变化,可是双音词的增加以近百年为甚,而且大部分是与经济、政治和文化生活有关的所谓‘新名词’。可见同音词在现代主要是起消极作用,就是说,要创造新的单音词是极其困难的了。”吕叔湘先生在这里一方面指出了要创造新的单音词(即单字)极其困难,一方面又指出了双音词(即双字词)的大量增加的现象,这正是组字成词的经济律的生动体现。

对汉字结构及其构成成分的统计与分析表明,在《辞海》(1979年版)所收的 16295 个字和 GB2312-80 国家标准《信息交换用汉字编码字符集·基本集》收入而《辞海》未收的 43 个字中,简化字和被简化的繁体字(包括被淘汰的异体字和计量用字)以及未简化的汉字共有 16339 个,它们是由 675 个不能再分解的末级部件构成的,简化字和未简化的汉字(不包括被简化的繁体字、被淘汰的异体字和计量用字)共 11837 个,它们

<sup>4</sup> Feng Zhiwei, Analysis of Formation of Chinese Terms in Data Processing, Fraunhofer-Gesellschaft, Stuttgart, Germany, 1988.

是由 648 个不能再分解的末级部件构成的。由少量的部件构成大量的汉字，体现了部件构成汉字这一编码过程的经济律。

所以，术语形成经济律实际上乃是“语言编码的经济律”，这是语言学中的一个普遍规律，它支配着语言编码的所有过程。

在研究 F E L 公式的同时，我还提出了“生词增幅递减律”，我指出，在一个术语系统中，每个单词的绝对频度是不同的，经常使用的单词是高频词，不经常使用的单词是低频词，随着术语条目的增加，高频词的数目也相应地增加，而生词出现的可能性越来越小，这时，尽管术语的条数还继续增加，生词总数增加的速率却越来越慢，而高频词则反复地出现，生词的增幅有递减的趋势。这个“生词增幅递减律”不仅适用于术语系统，也适用于阅读书面文本的过程，人们在阅读一种用自己不熟悉的语言写的文本时，开始总有大量不认识的生词，随着阅读数量的增加，生词增加的幅度会逐渐减少，如果阅读者能够掌握好已经阅读过的生词，阅读将会变得越来越容易。

我还与上海交通大学博士生李晶洁合作，基于布朗语料库 (Brown corpus) 的证据，考察科技英语的篇际词汇增长模型，以篇章为计量单位，描述科技英语文本中词汇量与累积文本容量之间的函数关系。我们注意到，国外现有的词汇增长模型不能够精确地描述科技英语的词汇增长曲线，因此，我们通过对幂函数和对数函数的比较分析，构建了新的词汇增长模型，并应用此模型推导出科技英语的理论词汇增长曲线及其 95% 双向置信区间。

在术语研究中，我还提出了“潜在歧义论” (Potential Ambiguity Theory, 简称 P A 论)，指出了中文术语的歧义格式中，包含着歧义性的一面，也包含着非歧义性的一面，因而这样的歧义格式是潜在的，它只是具有歧义的可能性，而并非现实的歧义，潜在的歧义能否专转化成现实的歧义，要通过潜在歧义结构的“实例化” (instantiation) 过程来实现，“实例化”之后，有的歧义结构会变成真正的歧义结构，有的歧义结构则不然。这一理论是对传统语言学中“类型—实例” (type-token) 观念的冲击，深化了对于歧义格式本质的认识，近年来，我又把 P A 论进一步推广到日常语言的领域，促进了自然语言处理中的歧义消解的研究。

术语是记录科学技术知识的基本单元，因此，术语的研究对于人类知识的系统处理，对于科学技术交流都有着重要的价值。1977 年，我把这些研究术语的成果写成《现代术语学引论》一书出版了，这是我国第一本关于术语学理论的专著。

## 用德语讲授中国语言文学课程

1990 年至 1993 年，我被德国特里尔大学文学院聘任为客座教授。特里尔是一座有 2000 年历史的古城，又是马克思的故乡，我有机会经常到马克思的故居了解这位无产阶级革命导师的光辉业绩。

在特里尔大学文学院任教期间，我用德语给德国学生讲授《汉魏六朝散文》《唐诗宋词》《中国现代散文》《汉字的发展与结构》《汉语拼音正词法》《汉语词汇史》《机器翻译的理论和方法》等课程。

我学过德语，有一定的德语口语交流经验，可是，用德语在高等学校的课堂上讲课，与日常生活中用德语口语交流大不一样，课堂是学术的殿堂，课堂上的语言不能有很多差错，特别是不能在语法上出错，而德语语法十分复杂，需要我严肃对待。为了讲好课，我苦练德语口语，认真用德语备好每一节课，在上每一节课之前，我都要先用德语把讲课的内容自己对自己叙述一遍或多遍，直到能够熟练地背诵为止，我把“备课”当作了“背课”。由于备课特别认真，我的课堂教学效果越来越好，我的讲课受到德国学生们的一致好评。当时我的一些德国学生现在已经成为德国知名的语言学家了。

在教学中，我发现德国学生学习汉语时，学讲话并不困难，最困难的是学汉字。汉字数量多，结构复杂，因此，我开始研究如何教德国学生学习汉字的问题。

我经过反复的思考，把自己在法国留学时提出的MMT模型运用到汉字结构的教学中，提出了汉字结构的括号式表示法，用这种方法可以把一个汉字按层次分解为若干个部件，构成一个树形结构，再把这样的树形结构用括号表示出来。学生只要掌握了基本的汉字部件，就可以进一步学会由这些部件构成的整个汉字，以简驭繁，使汉字便于理解和记忆。这样的方法受到德国学生的欢迎。

我把这样的研究成果写成了《汉字的历史和现状》一书用德文在特里尔科学出版社出版。德国特里尔大学韦荷雅(Dorothea Wippermann)博士1996年在《评冯志伟新著〈汉字的历史和现状〉(德文版)》一文中指出，冯志伟“在汉字研究中引入了现代的成分分析法。对于这种方法，直到现在为止，许多在专家圈子之外的普通人还很不熟悉，所知极少。这种分析法认为，汉字是由不同的图形成分组合而成的一个封闭的集合，其中的每一个较大的成分都可以进一步被拆分为较小的成分，一直被拆分到单独的笔画为止。汉字结构的这种多层次的构造图形可以用树形图来表示，这样一来，便为揭示汉字总体结构的研究提供了一种系统性的理论和方法。这种在中文信息处理中行之有效的成分分析法，对于汉字的研究和学习，也提供了一种新的记忆手段”。

汉字的计算机处理一直是我关注的一个重要的应用问题。近年来，我与旅居加拿大的青年学者欧阳贵林合作，把汉字的基本字根归纳为25个，我们在这25个字根基础上提出的“机写汉字学习法”(简称“和码”)，这是一种以简驭繁的汉字学习的方法。我们在加拿大和九江的儿童识字教学中进行试验，效果良好。

目前，汉字输入计算机主要使用拼音输入，拼音输入是一种简捷而方便的输入法，为群众喜闻乐见。但是，由于拼音与汉字的字形之间没有明确关系，长期使用拼音输入，往往会忘记汉字的字形，写字时出现“提笔忘字”的情况，有人把这种情况叫做“汉字失写症”。我认为，除了继续使用和推广拼音输入法之外，我们还需要在计算机上根据汉字的结构使用键盘来书写汉字，从而避免“汉字失写症”，继承汉字的文化传统。“机写汉字学习法”使用键盘来书写汉字，有助于克服由于长期使用拼音输入汉字而导致的“汉字失写症”这种文化病。

我们还开发出针对外国学生学习汉字的相关的软件，在北京语言大学的部分外国学生中进行过初步的试验，效果良好，“机写汉字学习法”软件让外国学生在学习“听说”汉语的同时，也能够“读写”汉语，达到“听说读写”四会的要求。

“机写汉字学习法”为汉字的键盘“机写”提供了一种方便而适用的手段，使我

们在计算机上输入汉字的时候，永远也不会忘记怎样书写汉字。这对于发扬我国汉字文化的优秀传统是大有好处的。

## 用英语讲授自然语言处理课程

2001年，我应邀到韩国科学技术院（Korean Advanced Institute of Science and Technology, 简称 KAIST）电子工程与计算机科学系担任教授。KAIST 是韩国著名的理工科大学，大部分学生都是通过严格的考试和数学物理竞赛选出来的精英。我不会韩国语，因此，只能用英语给该系博士研究生开设“自然语言处理-II”（Natural Language Processing -II, 简称 NLP-II）的课程。在这门课程中，我系统地讲授了词汇自动分析、形态自动分析、句法自动分析、语义自动分析、语用自动分析等自然语言处理中的各种方法，受到韩国学生的欢迎，韩国科学技术院还特别出版了文集来纪念我的这次讲学<sup>5</sup>。

在用英语备课的过程中，我发现美国 Colorado 大学的 Daniel Jurafsky 和 James Martin 的新著《Speech and Language Processing -- An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition》（《语音和语言处理 – 自然语言处理，计算语言学和语音识别导论》）是一本很优秀的自然语言处理的教材，这本教材覆盖面非常广泛，理论分析十分深入，而且强调实用性和注重评测技术，几乎所有的例子都来自真实的语料库。我想，如果能够把这本优秀的教材翻译成中文，让国内的年轻学子们也能学习本书，那该是多么好的事情！

2002年，我回国参加机器翻译的学术讨论会，电子工业出版社的一位编辑找到我，说他们打算翻译出版此书。这位编辑说，电子工业出版社已经进行过调查，目前国外绝大多数大学的计算机科学系都采用此书作为“自然语言处理”课程的研究生教材，他们希望我亲自来翻译这本书，与电子工业出版社配合，推出高质量的中文译本。电子工业出版社的意见与我原来的想法不谋而合，于是，我欣然接受了这本长达 600 多页的英文专著的翻译任务，于 2003 年开始进行翻译。

我虽然已经通读过这本书两遍，对于这本书应该说是有一定的理解了，但是，亲自动手翻译起来，却不像原来想象的那样容易，要把英文的意思表达为确切的中文，下起笔来，总有汲深绠短之感，大量的新术语如何用中文来表达，也是颇费周折和令人踌躇的难题。

在韩国教书期间，我利用了全部的业余时间来进行翻译，晚上加班到深夜，连续工作了 11 个月，当翻译完 14 章（全书的三分之二）的时候，不幸患了黄斑前膜的眼病，视力出现障碍，难于继续翻译工作，还剩下 7 章（全书的三分之一）没有翻译，“行百里者半九十”，这 7 章的翻译工作究竟如何来完成呢？正当我束手无策、一筹莫展的时候，中国科学院软件研究所的一位年轻的副研究员孙乐表示愿意继续我的工作，协助我完成本书的翻译。孙乐把剩下的 7 章逐一翻译成中文，通过计算机网络一章一章地传到

---

<sup>5</sup> KORTERM, 2001-2002 Collection of FORTERM Publication – in Honor of Professor Feng Zhiwei -, KAIST, Korea, 2002.

韩国，我使用语音合成装置，让计算机把书面的文本读出来，通过读出来的语音进行译文的校正，语音合成技术克服了我视力不济的困扰，帮助我迈过了重重的难关。2004年，在我们两人的通力合作下，全书的翻译总算大功告成了，由电子工业出版社以《自然语言处理综论》的书名出版。

这本书的出版受到广大读者的欢迎，而我为此却损害了自己的视力，不得不借助于语音合成装置来阅读了。

现在我已经进入古稀之年，不能再做很多具体的开发和研究工作了，我的视力不济，难于长时间看书，所以，我近来主要做一些介绍和引进外国优秀计算语言学英文原著的工作，为这些著作写导读，以便帮助年轻学子尽快地接触到当代计算语言学的前沿问题。我写的导读有：《应用语言学中的语料库》（世界图书出版公司&剑桥大学出版社，2006年版），《译者的电子工具》（外语教育与研究出版社，2006年版），《人工智能在第二语言教学中的应用》（世界图书出版公司，2007年版），《语言学中的数学方法》（世界图书出版公司，2009年版），《**牛津计算语言学手册**》（**外语教育与研究出版社，2009年版**），《自然语言生成系统的建造》（北京大学出版社，2010年版）。

## 学海无涯苦作舟

2006年6月30日，联合国教科文组织奥地利委员会（Austrian Commission for UNESCO）、维也纳市（City of Vienna）和国际术语信息中心（INFOTERM）给我颁发了维斯特奖（Wüster Special Prize），表彰我在术语学理论和术语学方法研究方面做出的突出贡献。维斯特（Eugen Wüster, 1898-1977）是奥地利著名科学家，是术语学和术语标准化工作的奠基人。维斯特奖是专门为那些对于术语学和术语标准化工作有出色成就的科学家而设置的。

可惜的是，我的视力越来越差，当我接受维斯特奖的时候，已经不能看清奖章上面的图案了。

我自己从事语言学研究已经50多年了，在这50年中，我始而弃理学文，继而弃文从理，后来又弃理从文，最后还是回到了语言学的队伍，看来我与语言学之间，确实有着割舍不断的缘分。

1957年我第一次阅读乔姆斯基的文章的时候，我还是一个不谙世事的19岁的小青年，乔姆斯基还是一个不满30岁的年轻学者，现在，我已经是年过70岁的白发苍苍的古稀老人了，而乔姆斯基已经82岁了。2010年8月，乔姆斯基应邀访问北京，我和乔姆斯基见了面，我们这两个老人一起合影留念。

我在乔姆斯基的影响下步入语言学的殿堂，曲曲折折地走了50年，乔姆斯基可以说是我学习语言学的最早的启蒙老师。我把我们合影的照片复制在这里，作为永远的纪念。



乔姆斯基与冯志伟合影留念（2010年8月14日）

语言学是一门历史久而博大精深的学问，50多年来，我主要是在数理语言学和计算语言学领域中研究和学习。尽管我现在已经年逾古稀，并且一天天地变老，但是，我50年来一直如痴如醉地锺爱着的数理语言学和计算语言学还是一门新兴的学科，她还非常年青，充满了青春的活力，尽管她还比较幼稚娇嫩，还不够成熟，但是她无疑地有着光辉的发展前景。我们个人的生命是有限的，而科学知识的探讨和研究却是无限的。我们个人渺小的生命与科学事业这棵常青的参天大树相比较，显得多么地微不足道，有如沧海之一粟。想到这些，怎不令我们感慨万千！

“书山有路勤为径，学海无涯苦作舟”，我们应当勤苦地工作，把个人的有限的生命投入到无限的科学知识的探讨和研究中去，从而实现人生的价值。