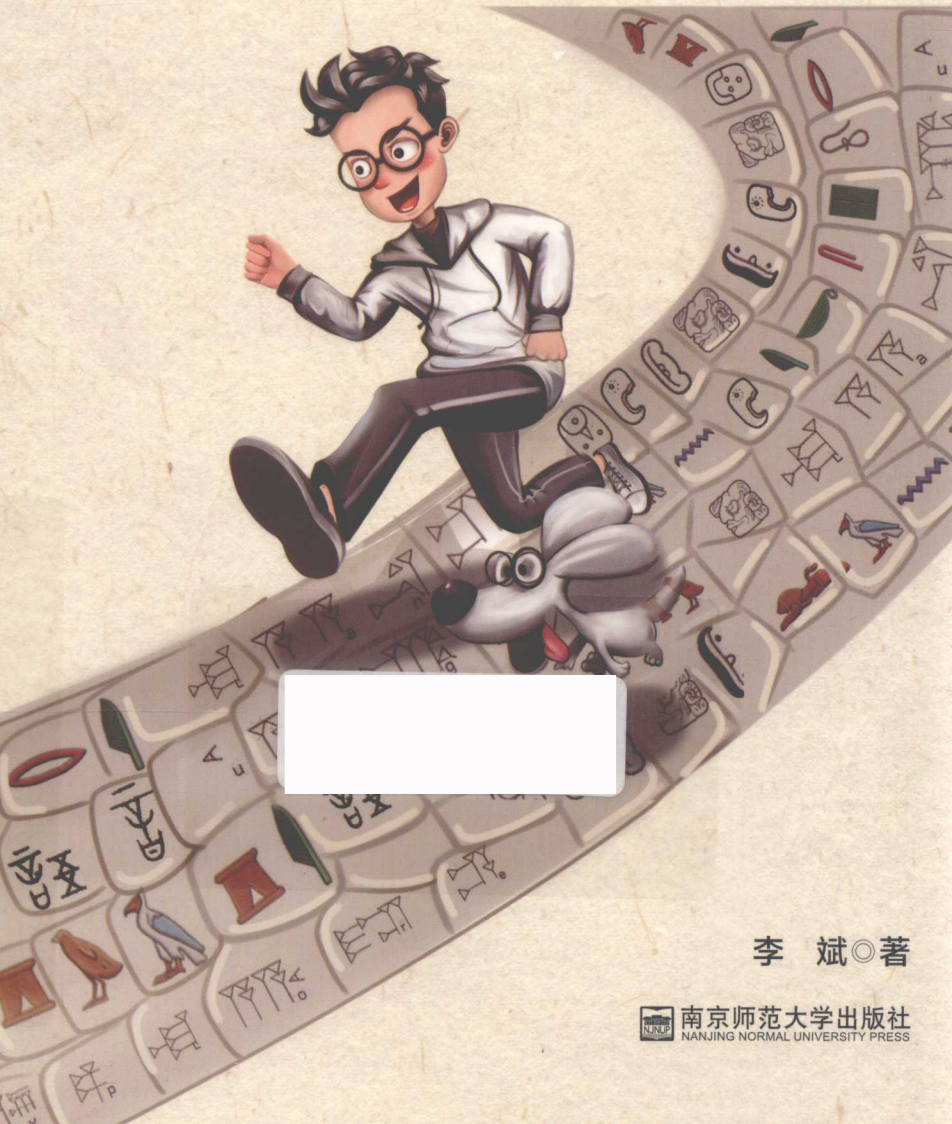
語言探秘



李　斌◎著

南京师范大学出版社
NANJING NORMAL UNIVERSITY PRESS

语言探秘

李斌◎著

南京师范大学出版社
NANJING NORMAL UNIVERSITY PRESS

图书在版编目(CIP)数据

语言探秘 / 李斌著. —
南京：南京师范大学出版社，2018.12
（语言与智能新视野系列）
ISBN 978-7-5651-3968-0

Ⅰ. ①语… Ⅱ. ①李… Ⅲ. ①语言学—研究 Ⅳ. ①H0

中国版本图书馆 CIP 数据核字(2018)第 298238 号

书　　名	语言探秘
作　　者	李　斌
责任编辑	于丽丽
出版发行	南京师范大学出版社
地　　址	江苏省南京市玄武区后宰门西村 9 号(邮编:210016)
电　　话	(025)83598919(总编办)　83598412(营销部) 83598297(邮购部)
网　　址	http://www.njnup.com
电子信箱	nspzbb@163.com
照　　排	南京理工大学资产经营有限公司
印　　刷	南京工大印务有限公司
开　　本	880 毫米×1230 毫米　1/32
印　　张	7.25
字　　数	162 千
版　　次	2018 年 12 月第 1 版　2018 年 12 月第 1 次印刷
书　　号	ISBN 978-7-5651-3968-0
定　　价	35.00 元

出 版 人　彭志斌

谨以此书献给

带我走入语言学殿堂的前辈师长、

令人百读不厌的语言学经典著述、

一起为语言智能而探索的同学们！

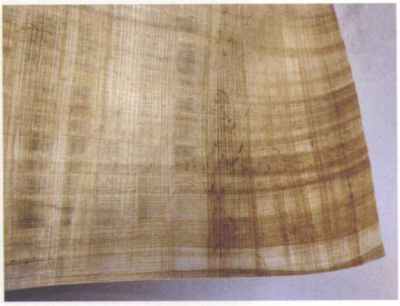
本书在语言大数据和人工智能的知识体系下，将语言学和汉语的基本问题，用对话的形式来讨论，步步推进，深入浅出地揭示语言的神奇与奥妙。对语言学、现代汉语、词汇语义学、计算语言学等方向的科研人员、本科生、研究生具有较高参考价值。

楔形文字印章，形成于公元前 2000–前1000 年，摄于波士顿艺术博物馆

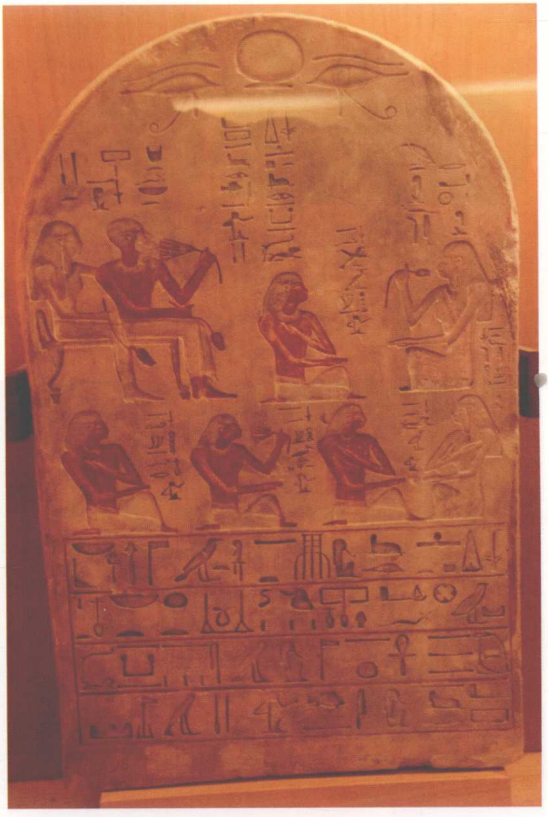
楔形文字泥版，约公元前3000年，取自网站https://cdli.ucla.edu/

玛雅石刻，摄于哈佛博物馆

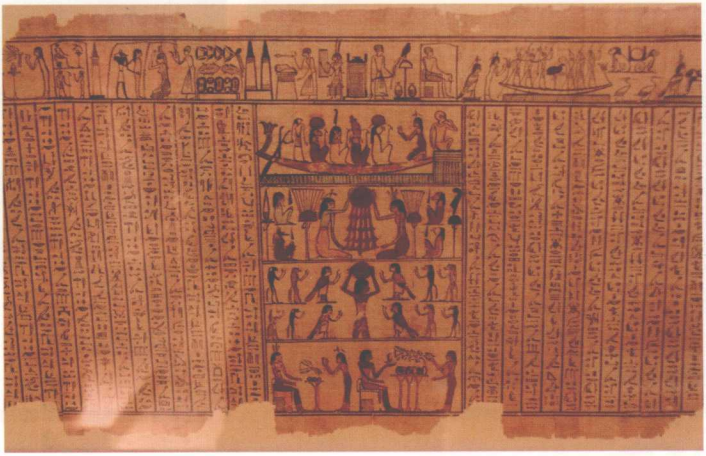
古埃及石碑，摄于波士顿艺术博物馆



现在埃及新制的由纸莎草编织而成的纸，摄于作者办公室

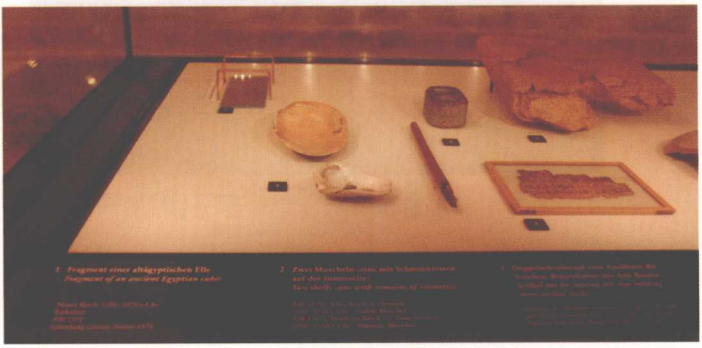


古埃及石碑，摄于意大利佛罗伦萨古埃及博物馆

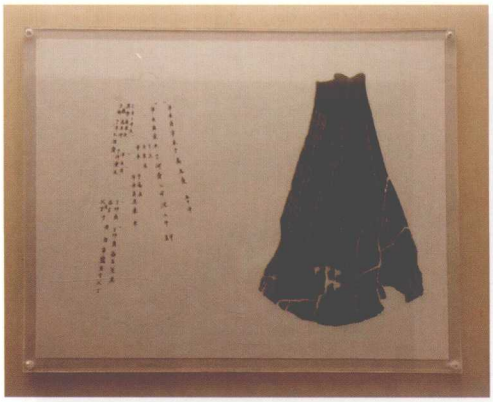


古埃及纸草书，铭文体，摄于柏林博物馆岛

古埃及纸草书，僧俗体，摄于柏林博物馆岛

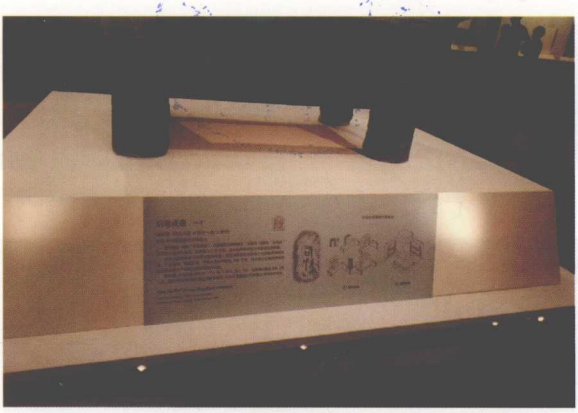


古埃及的笔、砚、纸草书残片，摄于柏林博物馆岛



商朝甲骨文及其拓片，摄于上海博物馆

公元 8 世纪的玛雅文字，摄于美国大都会博物馆



后（司）母戊鼎上的铭文，摄于中国国家博物馆

冯序

我怀着极大的兴趣通读了李斌博士的新著《语言探秘》。语言确实充满了奥秘，值得我们深入地探索。

丹麦哥本哈根学派的叶尔姆斯列夫(L. Hjelmslev)在他的《语言理论导论》(*Prolegomena to a Theory of Language*)一书中，曾经这样赞美语言："语言是人类社会基本的和最不可少的基础……在我们的意识第一次觉醒之前，语言就是我们的回声，它反映我们思想的第一次温柔的喃语，从日常活动一直到最细腻、最甜蜜的时刻，它寸步不离地伴随着我们……语言不是伴随人的外部现象。它十分紧密地跟人的理智联系在一起。它是个人和部族继承下来的财富。"他又说，"语言，即人的话语，是永不枯竭的、方面众多的巨大宝库。语言不可与人分割开来，它伴随着人的一切活动。语言是人们用来构造思想、感情、情绪、抱负、意志和行为的工具，是用来影响别人和受别人影响的工具，是人类社会的最根本、最深刻的基础，同时语言又是每个人的最根本、不可缺少的维持者，是寂寞中的安慰。在十分苦恼时，诗人和思想家是使用独白来解决思维矛盾的。在我们有意识之前，语言就已经在我们耳边回荡，准备环抱

我们最初思想的嫩芽，并伴随我们的一生。不论是平常最简单的活动，还是最崇高的事业，或者私人生活，人们一分一秒也离不开语言。是语言赋予我们记忆，我们又借助于记忆而获得温暖和力量。然而，语言不是外来的伴侣，语言深深地藏在我们的脑海之中，它是个人和家族继承下来的无穷记忆，是有提醒和警告作用的清醒的心智。而且，言语是个人性格的明显标志，不论是何种性格；它又是家族和民族的显性标记，是崇高人性的特殊标志”。叶尔姆斯列夫还说，“语言在个人、家庭、民族、人类及生活本身中扎根如此之深，以致使我们忍不住提出这样的问题：语言是否不仅是现象的反映，而且也是这些现象的体现——也就是产生出这些现象的种子”。①

语言如此美妙，如此有用，按理说，每一个学习和使用语言的人都应当对语言学兴趣盎然。可是，现在大学里的语言学课程却不太受学生的欢迎，不少学生都觉得语言学是一门索然无味的课程。李斌博士的这本《语言探秘》，没有按大学教材的方式来写，而是通过语言学家林贵思博士和小狗罗奇的对话，一步一步地把读者引入语言的殿堂，饶有趣味地揭示出语言的奥秘。学习语言学课程的读者如果同时也读一读这本《语言探秘》，不仅不会再有索然无味的感觉，而且将会产生学习和研究语言的兴趣，还会激起学习和研究语言的热情。

本书共有五个部分。

第一部分“语言的产生与发展”，讲述了词汇和语法的发展，特

① L. Hjelmslev. *Prolegomena to a Theory of Language*[M]. Baltimore: Waverly Press, 1953.

别是介绍了苏美尔文字、古埃及文字、玛雅文字，其中的许多古文字照片都是作者在国内外的博物馆亲自拍摄的，拓展了我们的眼界。

第二部分“信息时代的语言新视野”，讲述了信息时代中语言符号的电子化表达方式、语言与大脑的关系，特别是解释了语言的经济性原理和霍夫曼编码方法。

第三部分“语言与信息论”，讲述了香农的信道理论，介绍了图灵测试和齐夫定律，并分析了活字印刷的原理。

第四部分“语言的数学建模”，讲述了现代语义学的原理，分别介绍了谓词逻辑、比喻、借代和语义选择限制等语义形式描写方法。

第五部分“语言信息处理”，讲述了计算机汉字输入的原理，分别介绍了搜索引擎、自动分词、机器翻译等语言信息处理的技术。

李斌博士是语言学专业出身的，具有文科背景，几年来，他不断地进行更新知识的再学习，又到美国进修计算机科学专业，从而逐渐改变了他自己原来的知识结构，成为兼通语言学和计算机科学的新一代语言学家。这本《语言探秘》，是他近年来在研究实践中对于语言的奥秘进行深入思考的产物。

李斌在《语言探秘》这本书中还给我们讲述了一个饶有趣味的故事。他介绍说，在 1988 年的一次自然语言处理评测讨论会上，美国著名语音识别专家贾里尼克(F. Jelinek)在报告他的语音识别系统研究工作时，说了一段很尖刻的话，贾里尼克说：“每开除一个语言学家，我的系统性能就提高一些。”贾里尼克对于参加语音识别系统研究的语言学家，采取了嗤之以鼻的蔑视态度。

我是研究自然语言处理的，当然也很关注贾里尼克的研究，拜读过他的论文。他曾经使用隐马尔可夫模型（Hidden Markov Model，HMM）等统计方法来研究英语的语音识别，有效地降低了误识率，大大地提高了正确率，一举把英语语音识别提高到实用的水平，他也因此而成为美国工程院院士。我非常钦佩贾里尼克的杰出成就，可是，贾里尼克为什么会说出这样的话呢？

对此，李斌在书中做了这样的分析，他指出，“传统语言学家由于不太了解计算机的算法模型，他们提出的很多解决方案反而拖后了开发的进程，降低了系统的性能”。因此，贾里尼克才说出这样尖刻的话。

我同意李斌的意见，有的传统语言学家确实不太了解计算机的算法模型，他们对语音识别系统和其他的自然语言处理系统提出的很多解决方案只是他们一厢情愿的想法，却又自认为他们的方案很有用，可是实际上这是不可能在计算机上实现的，一旦采用他们的方案，必定会拖了语音识别和其他自然语言处理研制的后腿，降低系统的性能，造成欲益反损的严重后果。因此，这样的语言学家遭到贾里尼克的奚落，也就不足为奇了。

我认为，贾里尼克在他的报告中奚落的是那些不懂计算机算法而且又不愿意更新知识的语言学家，如果语言学家也学习计算机的算法，与时俱进，更新知识，把计算机算法与语言学规则结合起来，就不至于受到贾里尼克的奚落了。

就在贾里尼克发表奚落语言学家言论的五年之后，1993 年 7 月在日本神户召开了第四届机器翻译高层会议，英国著名学者哈钦斯（J. Hutchins）在会议的特约报告中指出，自 1989 年以来，机

器翻译的发展进入了一个新纪元。这个新纪元的重要标志是在基于规则的技术中引入了语料库方法。这种建立在大规模真实文本处理基础上的机器翻译要使用统计技术，叫作统计机器翻译(Statistical Machine Translation，SMT)。统计机器翻译是机器翻译研究史上的一场革命，它把自然语言处理推向一个崭新的阶段。哈钦斯在他的报告中并没有奚落语言学家，而是号召语言学家学习语料库的方法，更新自己的知识。

在统计机器翻译的研究中，由于有语言学家参与语料库的加工，有效地提高了语料库的质量；由于有语言学家在统计方法中导入了可计算的短语规则和句法规则，克服了数据稀疏的缺陷。在参与统计机器翻译研制的过程中，不少语言学家通过努力学习计算机算法的理论和技术，不断地进行更新知识的再学习，成为兼通语言学和计算机科学的新型语言学家。

语言学家更新知识之后，贾里尼克也改变了对于语言学家的偏见，他在2004年发表了一次演讲，演讲的题目是“我的一些最好的朋友是语言学家”，他在演讲的最后说：“物理学家研究物理现象，语言学家研究语言现象。工程师要学会利用物理学家的真知灼见，而我们则要学会利用语言学家的真知灼见。”可见贾里尼克在十六年前奚落的并不是所有的语言学家，而是那些故步自封并且不愿意更新知识的语言学家，我们不应当苛责贾里尼克。为了适应信息时代语言学研究的新发展，语言学家有必要进行更新知识的再学习，努力完善自己的知识结构，这应当是信息时代的语言学家责无旁贷的任务。

目前，基于多层神经网络的、以大数据作为输入的深度学习

(Deep Learning)方法引入机器翻译中。这是一种新型的机器自动学习。深度学习的训练方式是无监督的特征学习,使用多层神经网络的方法。这种多层神经网络是非线性的,可以重复利用中间层的计算单元,减少参数,计算机从海量的大数据中可以自动地产生模型的特征和算法。

词向量(Word Vector)是多层神经网络的一种重要方法,词向量把单词映射为一个固定维度的向量,不同的词向量构成词向量语义空间,在这个词向量语义空间中,语义相似的单词距离会比较近。美国机器学习研究者米克罗夫(T. Mikolov)发现,如果用"意大利"这个词语的属性向量减去"罗马"这个词语的属性向量,再加上"巴黎"这个词语的属性向量,就能得到"法国"这个词语或者相近词语的属性向量。类似地,如果用"国王"的属性向量减去"男人"的属性向量,再加上"女人"的属性向量,就能得到"王后"的属性向量。词向量的计算结果竟然与人们对于语言词汇的理解直觉很接近,这是非常令人振奋的结果,因为米克罗夫事先并没有刻意地做这样的安排。但是,其中的奥秘究竟如何,还有待我们进一步探索。

2007 年以来,采用深度学习的方法,以大规模的双语对齐的口语语料库作为语言知识的来源,从双语对齐的口语语料库中获取翻译知识,统计机器翻译又进一步发展成了神经机器翻译(Neural Machine Translation, NMT)。口语神经机器翻译正确率已经超过了 90%,针对日常口语的神经机器翻译已经基本上可以付诸实用了。

然而,在这种神经机器翻译中,语言之间的翻译细节还是一个

黑箱(black-box),尽管翻译的结果不错,但是研制者对于其中的语言处理机制仍然是不清楚的,在语言学理论上还难以做出科学的解释。探索这个黑箱的奥秘,当然需要语言学家的参与。

在自然语言处理中,类似的语言奥秘数不胜数,需要我们进一步探索,语言学家在自然语言处理的研究中是大有可为的。

李斌的《语言探秘》一书给我们揭示了语言的很多奥秘,语言中还有大量的奥秘等待我们去发现,去研究,去解释。希望读者在阅读了本书之后,积极地投身到语言探秘的研究工作中去,为语言学的新发展贡献出自己的聪明才智。

冯志伟

2017 年 10 月于德国海德堡

前言

这是一本探索语言奥秘的书。不同于一般的语言学教材和专著的地方在于，这本书不仅追求知识的基础性、针对性，而且更像一本旅游探险小说，采用林贵思博士和小狗罗奇两位虚拟主人公对话的形式，以“是什么？为什么？”为主线，剖析我们习以为常的语言。这本书与其说是追求结论或答案的确定性和可靠性，不如说是将各种不确定性用对话的方式组织起来，与读者一起思考和探讨语言的奥秘。

本书是笔者多年来教学、科研与思考的总结。书中相当多的内容，源自作者的科研成果和师生之间的讨论。在这个信息爆炸的年代，国内外语言学著作多如牛毛，穷一生之时也难以卒读，何况初步了解语言学的学生，难以分辨语言学论著的优劣，往往陷入各种不同的理论甚至杂说之中，难以系统地理解与把握语言学的主要问题和发展脉络。于是，我们干脆放弃新时代的论文海读战术，代之以传统的问答推敲，将语言学和汉语的基本问题贯穿起来，从语言、历史、生物、计算、认知等多重视角展开讨论，激发师生的研究和学习热情。不过，海阔天空地讨论，使得本书的内容超出