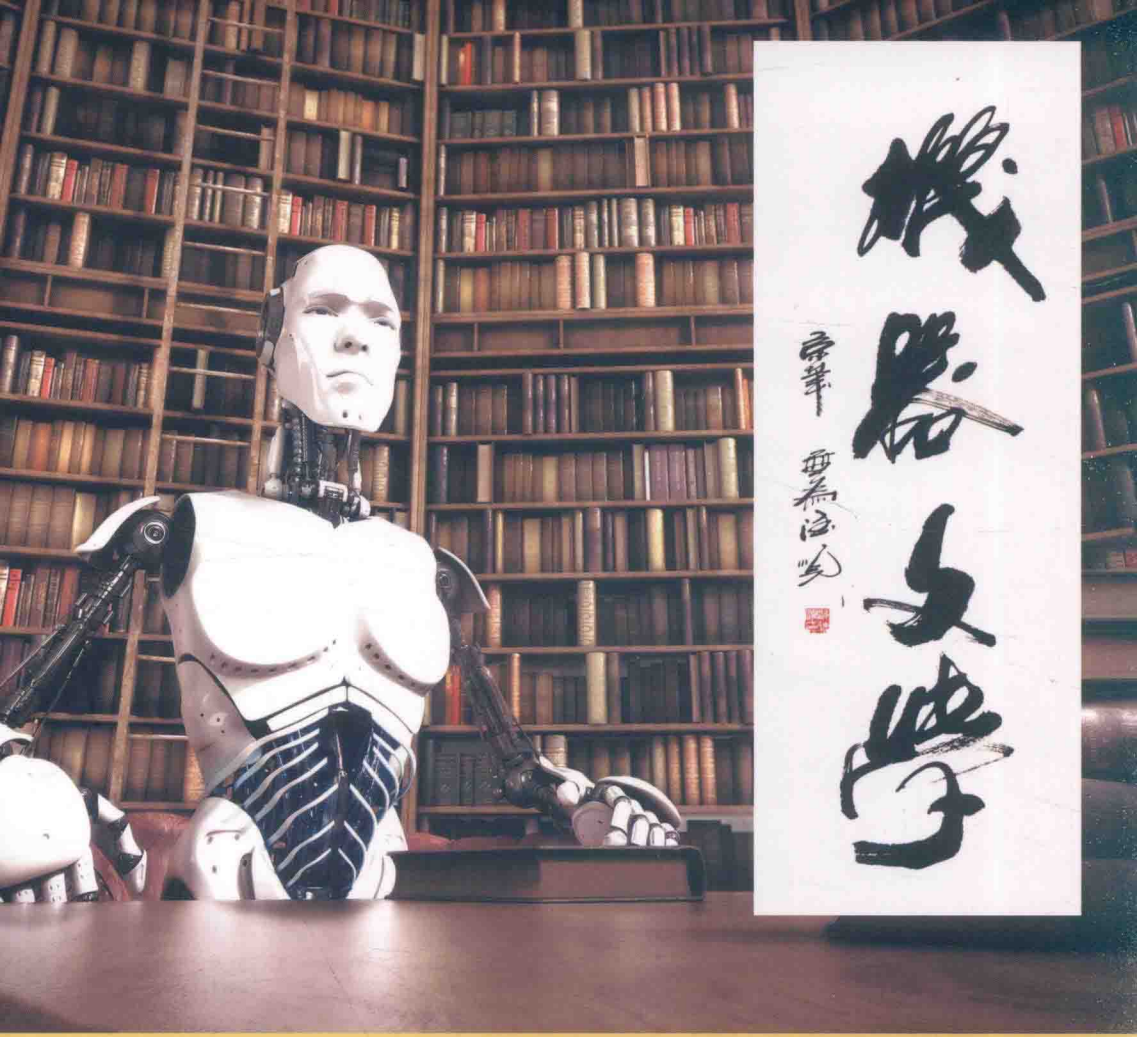

杨义先 著

机器文学就是研究“机器文”的学问，它是典型的文理结合的全新学科。由于几乎不可能完全由机器来代替人类进行文学创作，因此，退一步，人们开始研究由机器来代替人类进行某些特殊类型的文种的创作。到目前为止，机器文学所用的主要工具包括：计算机科学与技术、密码分析与破译理论、逻辑分析与综合、数学等。至今，人们在机器文学研究方面取得的突破性成果主要有：易道文类的机器文、璇玑图类的机器文、语音类的机器文。

——作者推荐

北京邮电大学出版社
www.buptpress.com

机器人学

杨义先 著

北京邮电大学出版社
www.buptpress.com

内容简介

“机器文学”既可理解为“用机器创作文学”，又可理解为研究“机器文”的学问。谦虚其心，宏大其量，融会贯通，方能大成。用科技琢磨辞采，将中华民族上下五千年的文字魅力表现的淋漓尽致，妙趣横生。如何利用人工智能和计算机等技术，创作出高水平的文学作品？这将是一个理论和技术难度很大，但是，应用前景十分美妙的领域。《机器文学》是老少皆宜的休闲知识类书籍，适合于所有背景的读者，甚至中学生都能享受其大部分乐趣。

图书在版编目（CIP）数据

机器文学/ 杨义先著. --北京 ：北京邮电大学出版社，2016.11

ISBN 978-7-5635-4898-9

Ⅰ. ①机… Ⅱ. ①杨… Ⅲ. ①计算机应用－文学创作－研究 Ⅳ. ①I04－39

中国版本图书馆 CIP 数据核字（2016）第 192829 号

书　　名：机器文学

著作责任者：杨义先　著

责 任 编 辑：付兆华

出 版 发 行：北京邮电大学出版社

社　　址：北京市海淀区西土城路 10 号(邮编：100876)

发 行 部：电话：010-62282185　传真：010-62283578

E-mail：publish@bupt.edu.cn

经　　销：各地新华书店

印　　刷：北京宝昌彩色印刷有限公司

开　　本：710 mm×1 000 mm　1/16

印　　张：11.25

字　　数：176 千字

版　　次：2016 年 11 月第 1 版　2016 年 11 月第 1 次印刷

ISBN 978-7-5635-4898-9　　定　价：35.00 元

序

最近因一事，如鲠在喉，寝食难安。

邢育森，从1998年起，我们就在一起合作，迄今，快二十年了。他写剧本，我做导演。育森想象力和词汇量丰富，逻辑紧密，做人低调，颇合我的胃口，于是，我们也成了生活中的挚友。但是，因为这样一本书，让我对这位多年的搭档和伙伴产生了怀疑。

是啊，谁能不怀疑呢。一个电信博士，精通密码学，还在国际权威学术刊物上发表过论著，居然弃理从文，做了网络文学三驾马车的一驾，还混入影视编剧圈儿兴风作浪。原来对他由衷钦佩——跨行业，居然也做得这么深度“垂直”，代表作和影响力让职业编剧汗颜艳羡——可是现在，看了这本《机器文学》，谁能不深深地怀疑！

数日前，育森拿来一本书，说是他博士导师的著作，要我写个序。我笑喷了：你导师不是国内密码泰斗吗，我能看懂他的书?！别逗了，还能不能好好创作啦！

育森幽幽地看了我一眼，幽幽地说：你还是认真看看的好……说完就幽幽地走了。

看着他意味深长的背影，我打开了这册本来距离我甚为遥远的书。多年喜剧的创作，已经让我笑点极高，国产某几部高票房的喜剧我能哭着看，可这本书开始就让我嘴角微扬——不愧是育森的师傅，杨义先教授，国内密码学泰斗，居然琢磨着让计算机搞文学！而且写得颇有板眼！哈哈……可是，看着看着，该轮到我“幽幽”了！我幽幽地想，万一今后机器真能随意创作了，那我们咋办？……咦！不对呀，这样的研究肯定非一日之功，今天能出书，一定是卓有成效了，那之前肯定有无数次实验和论证，那育森也肯定协助导师做了不少工作，那育森是不是用机器完成了自己的文学创作来验证导师的科研……我瞬间明白了育森“幽幽”的眼神。

你别说，《机器文学》这本书还真出人意料！其人工撰写部分，能够出自一位理工科教授，真的少见，其文字水平完全不亚于文科生；其机器创作的部分，

更是有点让人目瞪口呆：

2014 央视春节联欢晚会上，成龙大哥在《剑心书韵》节目中领诵的《千字文》我也懂，可是，万万没想到的是，《机器文学》却把这么高深的文学作品，玩得如此轻松。不但纠正了古人的错误，删补了重字部分；而且，还从数学上找到了一般性的存在定理，把奇迹变成当然。书中的“史上最长千字文”，长达七千余字，而且四言、五言、全韵、转韵，想怎么玩，就怎么玩。把人看得眼花缭乱，只剩下一堆震撼。这算恐怖剧?! ——那育森写的剧本……

见过百家姓，听过千家姓；还真没碰到过，由机器创作的“百家姓”和“千家姓”。著名的“赵钱孙李、周吴郑王……”，在计算机面前，竟然变成“死的百家姓”了，真的还有活的，每句话都有含义的“百家姓”呢！“千家姓”的众多隐私秘密，在《机器文学》里，已经被剥得几乎一丝不挂了。这算“儿童不宜”剧?!

坦率地说，“机器训诂学”和第一章“用数学研究语文”我没看懂。什么甲骨文啦，猜想啦，命题啦等等，我就不瞎评了，否则，我又不是机器……

关于单音文，我还想再啰嗦几句。这章的内容，除了让我头晕，还是头晕！一篇文章中的所有字，都只发一个音！虽然在网上偶尔看到过什么“唧唧鸡，鸡唧唧。几鸡挤挤集矶脊。机极疾，鸡饥极，鸡冀己技击及鲫。……”，但是，真没想到，机器能够创作出上百篇这种怪文，而且，还声称：汉字的每个音也许都存在着这样的文章。可能不会有任何一个人能把这章一字不漏地读完。本章不但能让你惊讶，而且，还可以给你催眠——想想育森创作的喜剧中某些人物的大贯口……不寒而栗啊！不行，我得去问问邢育森，见见他导师，我要弄明白这二十年来，育森给我们写的剧本和出版的书是他写的还是机器创作……但是，我还得发自内心地“幽幽”地说：《机器文学》还真值得认真研读，而且，确实老少皆宜！也许有一天，你们都能通过机器，成为邢育森……

吕小品

作者序

浪漫的文学与冷冰冰的机器能结婚生子吗？

逻辑思维的数学与形象思维的语文能谈恋爱吗？

左脑支撑的理工科人员与右脑支撑的文科学者水火不容吗？

文学家、科学家、艺术家、音乐家、工程师等职业人员能彼此融会贯通吗？

读大学之前，就非要把学生分为文科生和理工科学生吗？

以上问题一直困惑着学生、家长、老师、官员，甚至每一个社会成员。许多人都会以“那不是我的专业”为借口，拒绝了无数天赐良机。虽然跨界成功者比比皆是，虽然大部分学生毕业后其实都会转换专业，都会“用非所学”，但是，“专业框框”仍然根深蒂固，牢牢地卡住了创新的脖子。这也许就是“钱学森之问”的答案之一吧。

本书将以出人意料的事实证明：**人为划分的所谓专业、学科等真的不重要，它山之石也许更可攻玉！**特别是进入大数据时代后，一方面大量的博闻强记已经越来越不重要了（即，右脑优势被冲淡了），许多精密的逻辑推理也不需要人工了（即，左脑的优势也被冲淡了）；另一方面超凡的想像力在解决理工难题中越来越重要了（即，右脑在理工科中可发挥更大作用了），抽象思维在社会科学的建模研究中也越来越重要了（即，左脑在文科中也能发挥更大作用了）。总之，左脑与右脑的融合已经势不可当了！

为了让所有背景的读者都能够轻松、愉快地享受此书，认可我们的文理融合观点，我们主要论述了如何用理工科思路去解决文科中的一些难题，并将其命名为《机器文学》。

机器文学有两层含义：其一，利用机器（计算机）来研究多种文学创作问题；其二，研究“机器文”的学问。

关于第一层含义，目前国际上已经取得了许多重大成果，比如，计算机已经能够快速地自动创作许多特殊的新闻稿，几乎可以与记者的作品相提并论；能

写藏头诗的计算机软件已经随处可见。但是，关于这部分成果的描述，绝对逃脱不出众多高科技知识范畴，因此，它不是本书的重点。

本书重点研究机器文学的第二层含义。那么，何为“机器文”呢？粗略地说，**所谓“机器文”就是历史和现代文学家们一直在研究，但其实更加适合于机器来研究的文学体裁**。其实，历史上各种“机器文”的地位相当高，甚至数百年来，国人首先接触到的就是“机器文”。例如，

《百家姓(千家姓)》：几乎每个中国人都喜闻乐见的一种文章，细节见第二章；

《千字文》：历史上与《三字经》齐名的蒙童必读书籍，细节见第三章；

单音文：如果文章中的所有“字”都发同一个“音”，那么，这样的文章就称为“单音文”。历史上，最著名的单音文作者可能要数“中国语言学之父”赵元任(1892,11,3—1982,2,24)老先生了！他一生创作了五篇单音文，比如，最具代表性的单音文之一便是《施氏食狮史》：石室诗士施氏，嗜狮，誓食十狮。施氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，氏始试食是十狮。食时，始识是十狮，实十石狮尸。试释是事。

不难看出，这篇《施氏食狮史》完全仰仗赵元任老先生无与伦比的文字功底，一般人很难企及！但是，如果借助计算机，那么，单音文的创作难度将大幅度降低，因为，从理论上讲，只需要将某音的所有同音字放入一个库中，然后，让机器来自动排序便可产生相应的“同音句子”。当然，其核心难点是：机器如何判断一串同音字组成的内容是“人话”！根据《新华字典》，当代汉字共有400余个音，事实证明，几乎每个音都能够产生一篇单音文。至今，本书作者已经创作了150余篇单音文(见第五章)。比如，根据北京堵车的事实，我们写出了如单音文。堵都：嘟，……，嘟，嘟……！堵，毒堵，都堵，堵都，渎都，黩都。独堵，独都堵，都督堵，妒堵督，睹都堵，读堵都。都督笃堵，毒渎独都；都堵肚堵，肚妒都督；独犊杜堵，赌椟杜堵；堵堵都督，督督堵度。杜堵赌杜牍，都督黩堵都；笃犊督堵都，堵都堵堵堵！嘟，嘟，……

同音文：两篇文章称为同音文，如果它们的发音完全相同，但是，内容和含义又完全不同！虽然对同音文的研究不多，但是，同音字和同音词绝对是现在网上的潮语，比如，“同学”与“童鞋”、“有才华”与“油菜花”等。同音短句的例子

是“分久必合，合久必分”与“汾酒必喝，喝酒必汾”等。关于一般的同音文，本书将在第一章中提出有趣的“影文猜想”。

单字文：即由单独一个字的不同读音写成的文章。至今，最著名的“单字文”可能要算下述三副对联了。1）上联：长长长长长长长（读法：Chang Zhang Chang Zhang Chang Chang Zhang ）；下联：长长长长长长长（读法：Zhang Chang Zhang Chang Zhang Zhang Chang）；横批：长长长长（读法：Chang Zhang Zhang Chang）。2）上联：朝（zhao）朝（chao）朝（zhao）朝（chao）朝（zhao）朝（zhao）朝（chao）；下联：朝（chao）朝（zhao）朝（chao）朝（zhao）朝（chao）朝（chao）朝（zhao）；横批：朝（zhao）朝（chao）朝（chao）朝（zhao）。3）上联：行（hang）行（xing）行（hang）行（xing）行（hang）行（hang）行（xing）；下联：行（xing）行（hang）行（xing）行（hang）行（xing）行（xing）行（hang）；横批：行（xing）行（hang）行（hang）行（xing）。

单字单音文：它由单独一个同音字的不同音调写成。至今，最著名的“单字单音文”也是这样两副对联。1）上联：好（hào）好（hǎo）好（hào）好（hǎo）好（hào）好（hào）好（hǎo）；下联：好（hǎo）好（hào）好（hǎo）好（hào）好（hǎo）好（hǎo）好（hào）；横批：好（hào）好（hǎo）好（hǎo）好（hào）。2）上联：种（zhǒng）种（zhòng）种（zhǒng）种（zhòng）种（zhǒng）种（zhǒng）种（zhòng）；下联：种（zhòng）种（zhǒng）种（zhòng）种（zhǒng）种（zhòng）种（zhòng）种（zhǒng）；横批：种（zhǒng）种（zhòng）种（zhòng）种（zhǒng）。

当然，“机器文”绝不仅限于上述的几类，本书中将对历史上最著名的几类“机器文”进行系统而深入的研究，解决多个千百年来，文学家们前赴后继，想解决但却又始终未能解决的难题。

无论从选题内容、著述思路、笔法运用、读者对象等方面来看，本书都属于异类。感谢研究生庞林源同学，在整理此书过程中所付出的辛劳。感谢雷敏博士的众多编辑和协调工作。更要感谢夫人钮心忻教授和儿子杨牧龙的支持和理解，因为，作为一名理工科教授的我，花费大量的时间和精力，动笔创作这本“不伦不类的异端邪说”确实有点不可思议。

虽然我已经出版了数十本中英文学术专著和大学教材，涉及数学、密码、编码、安全等领域；但是，本书是我的第一本，但肯定不是最后一本“莫名其妙”的

作品。我坚信，中国太缺少本书这样的“胡思乱想”了。真心盼望本书能够激发更多的专家和学者，拿起笔来，打破学科界限，创作出更多、更好的跨界作品，以此激活国人创新能力，早日解决钱学森之问。

作者于北京

目　录

第一章　用数学研究语文

数学是科学之母，语文是科学之父。可是，国人非要棒打鸳鸯，活生生地把这个幸福家庭拆散，让数学和语文比邻若天涯：文科专家们一听到数学，就头皮发麻，恨不能只学点《九九口诀表》；理科专家则对语文不屑一顾，即使贵为博士，也仍然是错别字连篇，甚至许多高大上科研成果的文字描述也令人汗颜。

那么，数学和语文真的就不共戴天吗？如果数学和语文都能『相亲相爱』，以数学为核心的理工科和以语文为核心的文科，还有天壤之别吗？

本章是全书的理论基础，主要利用数学思路，从全新的角度，提出有关语文的若干重要猜想。这些猜想将在解决后面各章的许多重要问题中，扮演关键角色。

字距猜想

人类的全部内涵可概括为两个要素：“言”与“行”。并且，成立如下定律。

定律 1：“言”与“行”其实是基本一致的。虽然确有“言行不一”的情况，但是，从整体统计规律看，长期生活在谎言中的人不多，而且也很痛苦。因此，可通过对“言社会”的分析，来了解“行社会”。比如，在“言社会”中，若“政府”与“贪腐”这两个词经常“碰面”的话，那么，很有可能“行社会”就“亚健康”了；

定律 2：“言”与“行”是相互影响的。人类通过各种“行”，获得若干经验，然后，以文字、图表、音视频、物品等“言”（或可以转化为“言”）的方式，把“经验”记录下来并（异地）传承给后人，以此影响后人的“行”。

定律 3：“言”是可以继承的。“行”却不能继承，至少“行”无法异地直接继承，即，必须以“言”为媒介。因此，人与动物的根本区别在“言”而不在“行”，当然，就更不在“劳动”这种特别的“行”了。

定律 4：“言”的稳定性远远好于“行”。甚至几千年前的经文、遗物等“言”，至今都还在（对“行”和“言”）发挥影响重要作用，当然，也在不断地产生新“言”。特别是在当今“大数据时代”，每天产生的新“言”量，大大地超过了人类早期数百年的“言”量总和。

关于“行”的社会，过去人们认为完全杂乱无章，但是，现在最新的科研成果发现，“行”的社会其实是一个紧凑的“小世界”。即：

6度社交空间猜想：任何两个人，都可以经过至多六次引荐，便能够相互认识。

虽然作为一个世界著名的数学猜想，“6度社交空间猜想”的表述非常不严谨，但是，事实证明该猜想在指导诸如Facebook、新浪微博、Twitter、微信等社交网络的建设和发展过程中扮演着非常重要的角色。而且，该猜想表明，至少在某一点（“相互认识”这一点）上，“行”社会是小尺度社会。

互联网是“言”社会中的第一大“国”，此外，诸如档案、影视、文艺等也都是“言”社会中的不同“国”。既然，根据上述定律1，“言”的社会与“行”的社会基本一致，那么，在“言”社会中也应该有类似的“6度空间猜想”，即：

字距2度猜想：任何两个字A、B，要么，它们在同一个词中（此时称为A与B的距离为1）；要么，可以找到第三个字C，使得A与C在同一个词中，同时，B与C也在一个词中（此时，称为A与B的距离为2）。

与“行”的“6度社交空间猜想”相比，此处“言”的“字距猜想”显然更加清晰。虽然至今仍然未能证明其正确性，但是，也没能找到反例，即，没找到某两个特殊字，使得它们之间的距离既非1，也非2！

由于汉语语法研究中，其实没有“字”的概念，代之却是“语素”“词”“短语”“句子”等“似曾相识而又生”的概念。虽然直观含义最清楚的是“字距2度猜想”，但是，为吸引更多研究者们的注意，上述猜想分解为如下几种情况：

语素级2度猜想：任意两个语素A、B，要么它们在同一个词中（此时称为A与B的距离为1）；要么，可以找到第三个语素C，使得A与C在同一个词中，同时，B与C也在一个词中（此时，称为A与B的距离为2）。

词级2度猜想：任意两个词A、B，要么它们在同一个短语中（此时称为A与B的距离为1）；要么，可以找到第三个短语C，使得A与C在同一个短语中，同时，B与C也在一个短语中（此时，称为A与B的距离为2）。

短语级2度猜想：任意两个短语A、B，要么它们在同一个句子中（此时称为A与B的距离为1）；要么，可以找到第三个短语C，使得A与C在同一

个句子中，同时，B与C也在一个句子中（此时，称为A与B的距离为2）。

如果上述“2度猜想”正确，那么：

（1）“言社会”将比“行社会”更紧凑。而且，在“言社会”中各种概念更确定，相关数学工具和建模理论将更有用武之地，当然，必须承认，至今对“言社会”的动力学理论几乎是一无所知，但是，只要有足够强大的需求驱动，“语言动力学复杂性理论”的诞生一定不会太遥远了。

（2）由于“言”的继承性和“言”对“行”的影响性将导致“行”的可预测性。换句话说，虽然“个人命运”不一定能“算”出来，但是，从统计学观点来看，人群的命运是“可算”的。当然，这绝对不是在宣扬“封建迷信”。

（3）直接改变“行社会”难度较大，甚至基本上不可能；但是，相比而言，“言社会”的改变就容易多了。对“言社会”的篡改，其影响将肯定蔓延到“行社会”中，并最终改变“行社会”，虽然有一定的时滞；同样，如果融入到全人类的统一“言社会”之中，那么，若干年后，“行社会”也就真正“与国际接轨”了。

（4）人们对“行社会”的“6度社会空间猜想”已经做了多年研究，并取得了不少成果，相信其中某些成果可以应用于研究“言空间”的“字距2度猜想”；同时，由于“言社会”的确定性更好，相信今后在“言社会”中的成果将更加深刻，而这些“更深刻”的成果，又将有助于“行社会”的研究。

在本小节结束前，还有如下几点说明：

（1）虽然上节是以中文为例来表述“字距2度猜想”的，但是，该猜想与语种无关，因为，各语种之间是可以翻译的，即用数学术语来说，它们是“同构的”，所以，只需考虑一种语言的“言社会”就行了。

（2）“字距2度猜想”还处于相当幼稚的阶段，理论基础、模型等都是空白，但是，随着大数据时代的来临，对它的研究将越来越必要。相信在“语言动力学”研究方面，在不远的将来，一定会发表一批高水平的学术成果，甚至可能登上《Nature》或《Science》这样的世界顶级刊物。

（3）给出“字距2度猜想”的严格数学证明其实并不重要，重要的是用它来揭示若干不为人知的重要秘密。当然，证明该猜想的可能思路有如下几种，其一，语文方法，比如，找反例来否定该猜想；其二，数学方法，仿照“6度社

交空间猜想”的数理统计法；其三，生物学方法，从人类的智能水平来考虑，比如，众所周知的“言不达意，词不达言”这个事实就表明，当今人类“言”的表述水平还不高，也许再经过若干世纪的进化后，人类的“言”水平将大幅度提高，“言社会”将更加复杂，到那个时候，“言社会”的维度数将有所增加；同理反推，也许人类早期（比如，甲骨文或更早的时期）的“言社会”是一个很简单的0度孤立空间呢，这时只有“语素”，压根就还没有“词”。

字典猜想

字典猜想：对任何一个自然字库，即，没有人工有意干扰的字库（比如，《新华字典》中的所有汉字或其中某些汉字组成的字库等），都可以撰写出至少一篇满足“字不重叠”且“有含义”两个条件的文章。其实，创作于一千多年前的《千字文》和经典的《百家姓》就是字典猜想的典型案例。为突出要点，本书中将满足“字不重复”和“有含义”这两个条件的文章也称为“千字文”，虽然它们的字数其实不足一千字。

其实从前面的字距猜想，我们可以看出：字与字之间的关系非常紧密。而此处的字典猜想，又使得我们从另一个角度体会了“字与字之间的紧密程度”。同样，本书不去努力证明该猜想，而是要揭示它给我们带来的若干惊奇。

现以大家喜闻乐见的《百家姓》为例，来说明字典猜想。

提起《百家姓》，大家立即就会想起那篇家喻户晓的文字：“赵钱孙李，周吴郑王；冯陈褚卫，蒋沈韩杨；朱秦尤许，何吕施张……”但是，这类文章不是本文要研究的“千字文”，虽然在此文中每个字也只出现一次（并未重复），但是，每句话却都没有含义，仅仅是简单的堆叠，内容是“死的”。实际上，我们研究的“千字文”必须满足两个条件：其一，每个字都不重复出现；其二，文章是“活的”，即，每句话都是“有内容”的。