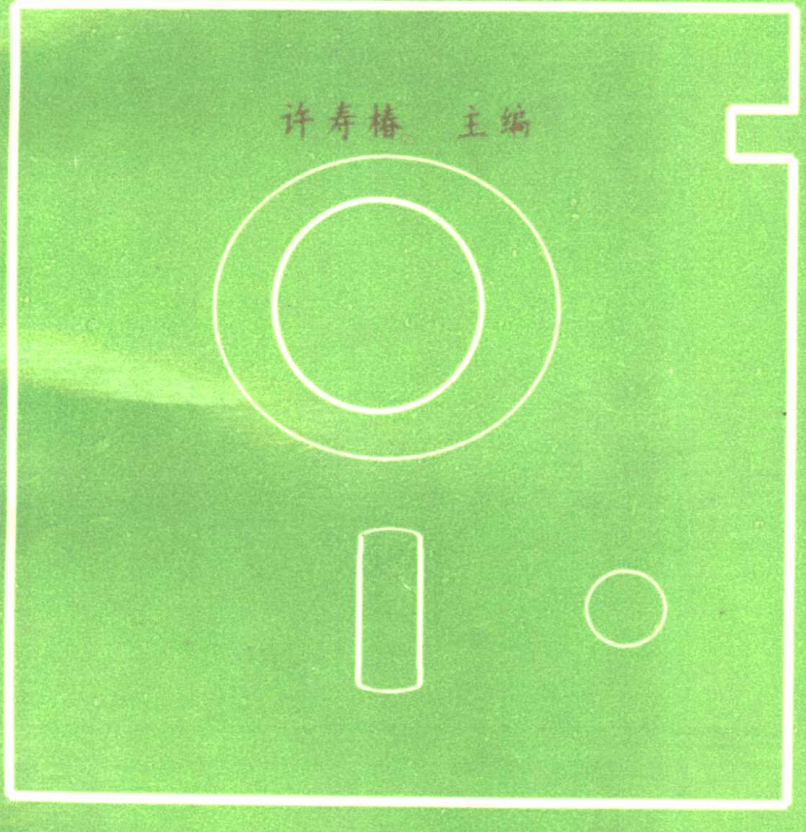
国家自然科学基金资助项目
(NO. 6883027)著作成果之三

文字比较研究散论

——电脑时代的新观察

许寿椿 主编



中央民族学院出版社

国家自然科学基金资助项目
(No. 6883027)著作成果之三

文字比较研究散论

——电脑时代的新观察

许寿椿　主编

中央民族学院计算机系
中国中文信息学会　民族语言文字信息专委会

中央民族学院出版社
1993年·北京

京新登字 184 号

责任编辑：葛小冲

封面设计：李全文

文字比较研究散论

——电脑时代的新观察

许寿椿　主编

*

中央民族学院出版社出版

（北京白石桥路 27 号）

（邮政编码：100081）

新华书店北京发行所发行

飞达印刷厂印刷

787×1092　毫米　32 开　11　印张　234 千字

1993 年 5 月第 1 版　1993 年 5 月第 1 次印刷

印数：01—2000 册

ISBN 7-81001-575-3/H·39　定价：8.80 元

内容提要

本书收入47篇论文，撰稿人包括语言学家和电脑专家两部分人。论文主题是文字类型及属性比较研究。定性与定量结合，文理结合，多文种比较、多侧面观察，是编写者力求突出的特点。本书面对语言学与电脑信息处理两个领域的广大读者，包括关心文字学现代发展的语言学家、文科研究生、大学生，也包括从事语言文字信息处理的电脑专家、有关研究生、大学生。本书为国家自然科学基金资助项目(No. 6883027)成果之一。*

---

* 项目名称"多语种微机语料库及语言比较"，申请人为许寿椿。项目有关其它成果情况见本书末尾。

编　者　的　话

本书编辑缘起　1991年和1992年中国中文信息学会民族语言文字专委会会同中央民族学院科研处、计算机系、民族语言研究所就“文字类型与属性研讨会”两次进行征文。1991年曾油印《文字类型与属性论文集(一)》进行通讯交流。1992年征文后，两次征文计收论文约60篇。现选约20篇，另加入编者近年来为计算语言学方向研究生讲授“现代文字学(电脑文字学)”中所得相关札记编成此书，以飨读者。本书中清格尔泰先生的一篇转自民族语文，因为收到四篇论文是直接讨论清氏方案的。

需要说明的是，未编入本书的征文，部分文稿是由于和本书主题不符，大部分实在是由于论文涉及太多特殊文字符号，排印将极大增加成本，而不得不割爱了。未编入本书的征文，目录附书后，对有关作者表示歉意和感谢。

本书的主题和特点　本书收入的论文都围绕一个主题，那就是文字类型和属性的比较。编者力图突出的特点是多文种、多侧面的比较观察，是定性与定量的结合，是语言文字学与电脑文字信息处理技术的结合。

80年代，语言文字信息的计算机处理技术获得迅猛发展，实现了普及应用。特别是多文种处理技术极为引人注目。世界上主要国家和民族的文字，重要的专业技术文字及某些特种文字，都实现了计算机处理。民族文字中，包括了与英、俄等拼音文字类型迥异的东方语文：汉文、日文、朝文、藏文。专业技术文字中包括数学表达式，化学方程式和结构式，五线

谱，简谱等。特种文字如盲文。文字信息处理，不仅在电脑打字、排版印刷、情报检索等方面获得成功，在文——音转换（机器朗读、语音合成）和音——文转换（语音打字、语音识别）方面也获得实质性进展。电脑文字处理技术多方面的成功和日益广泛的应用，促进了语言文字学家和电脑信息技术专家的了解、合作和结合，拓宽了他们的视野。这也推动了、激励了编著者从更广大的范围、更众多的侧面来观察思索。

本书的编著者　本书一共包括17名作者，其中半数稍多的人本专业为语言文字学，半数稍少的人一直是理工科技工作者。编著者中多数人有某种利用电子计算机从事文字信息处理的经验；那没有什么实际经验的一部分，对电脑文字信息处理至少抱有可贵的兴趣和热情。两部分人意外的结合，根本上说来，是80年代以来文字信息电脑处理技术的发展浪潮促成的。

本书的编者　本书编者早年就读北京大学计算数学专业，30年来长期从事计算机应用及软件的教学和研究。近六、七年内涉足语言文字信息处理。在筹建、主持中央民族学院计算机系的同时，积极地参加了中国中文信息学会及所属民族语言文字信息专委会的学术活动，并以专委秘书的身份参与了中文信息学会民族语言文字专委学术活动的筹划和组织。中文信息学会中突出的文理结合的特点和中央民族学院及民族专委内部多民族语言的环境，逼迫或推动编者不得不学习、思索一些相关的语言文字学问题，也使得自己较多地了解国内外多文种计算机技术的动态、问题和发展。这种条件和环境，这种逼迫或推动是编者有勇气出面主编这本小书的缘由。这里所以说明这点，编者以为是一种必要的交待。用意之

一是告诉读者本书编辑的背景，提醒读者：在语言文字学王国面前，编者实在是个门外汉，顶多不过是个小学生。论述之中也难免有些无知妄说的。用意之二是感谢一切给编者以影响、鼓励、帮助的那些同行的和隔行的朋友。

本书的读者（编著者心中设想的读者） 两类专业编著者心里的读者群也都包括两类：从事或关心语言文字信息计算机处理的语言文字学家、文科大学生、研究生和理工科电脑信息技术专家及相关专业大学生和研究生。编写者撰稿时心里都想着要使隔行的读者可能读懂和感兴趣。

致谢 本书编者自 1989 年起获得国家自然科学基金资助，主持“多语种微机语料库及语言比较”项目。本书论文的撰写和征集都是在项目过程中进行的。这是本书注明“国家自然科学基金资助项目”的缘由。另外，感谢中国中文信息学会民族语言文字专委会有关领导、同仁，感谢中央民族学院科研处领导及同志们的鼓励支持。

目　录

三

四

五

六

七

八

九

拉丁文字圈概况

许寿椿

1. 拉丁字母，因其早期用于书写拉丁文而得名。又因拉丁文是古罗马的文字，现今意大利首都罗马是拉丁字母的故乡，所以也称之为罗马字母。拉丁字母，无论从使用人口还是从使用地域来看，都是现今世界上应用最广的字母，详见本书“世界上文字的一些综合情况”中的附图。

全世界使用拉丁字母文字为全国性正式文字的，全世界计有 117 个国家。其具体分布如下：欧洲 27.5 个国家。这里的半个指南斯拉夫。非拉丁字母文字的仅 3.5 个国家。美洲 33 国，大洋洲 10 国，全部使用拉丁字母文字，没有其他文字圈的范围。非洲 51 个国家中，44 个国家使用拉丁字母文字。另外 6 个用阿拉伯文，1 个用埃塞俄比亚文。亚洲是拉丁字母势力范围最小的洲，是各文字圈并立的洲。使用拉丁字母文字的仅 7 国。另外的分属 4 个文字圈。用阿拉伯字母的 17 国；用印度字母的 10 国；用汉字的 3 国，用斯拉夫字母的 1 国（指蒙古，前苏联列入欧洲），用希伯来文 1 国（以色列）。以上是苏联解体前的统计。苏联解体使苏联分为 15 个独立国家的联合。这使世界上拉丁文字圈国家增到 120 个。斯拉夫文字圈中的国家增至 14.5 个。据报导，原苏联中亚的一些国家可能在未来几年内改用阿拉伯字母。

不以拉丁字母为正式文字字母的，大都规定了拉丁转写表示方法。数理科学或技术科学中，拉丁字母已成为不可缺少

的国际字母。其他四大文字圈中，数理科学论文或技术资料，实际上都是与拉丁字母混用写成的。

2. 拉丁字母形成于公元前 7 世纪。它是希腊字母(形成于公元前 9 世纪)经埃特鲁斯坎字母(形成于公元前 8 世纪)演变形成的。最早的拉丁文遗物是公元前 7 世纪的一枚金扣针。上面的拉丁字母从右向左书写，词之间的间隔符用类似于冒号的双点。这都是埃特鲁斯坎字母的特点。

早期的拉丁字母有 21 个，都从挨特鲁斯坎字母表中借来。公元前 3 世纪从希腊字母引入 Y 和 Z，中世纪由 I 中分化出 J，由 V 分化出 V 和 W，形成现今 26 个元素的拉丁字母。

早期的拉丁字母没有大小写的区别。大写字母是在早期铭刻体的基础上形成的，在四世纪达于完善程度。小写体是适应笔写要求的手书字体。小写体比大写体笔画省略(如 H 变为 h，B 变为 b)，便于手写。通常快速的手写比铭刻容易混淆。为了弥补这点，小写体字形改变了大写体高低一律的格局，形成上伸字母(b,d,h,k……)、下延字母(g,j,p,q,y……)和短字母(a,c,m,n……)。这种高低错落有利于扫读。小写体的成熟在公元 3—8 世纪。今天印刷上广泛使用的正体(罗马体)和斜体(意大利体)，成熟于 15 世纪的威尼斯。

3. 拉丁字母最初的 600 年使用范围主要限于意大利半岛，是书写拉丁文的工具。公元前 30 年建立了罗马帝国，到公元 2 世纪扩张成为版图辽阔的大帝国。拉丁字母随着罗马军队传播到欧洲的广大地区。当时欧洲许多民族还没有自己的民族文字。拉丁文成为欧洲许多民族的文字工具。这些民族的口语和拉丁文是完全不同的。罗马帝国时期正值中国的汉朝。欧洲国家学用与自己口语不同的拉丁文和东方的日本、朝

鲜、越南学用与自己民族语不同的汉语文的情况相似。罗马帝国时期是拉丁字母的第一次大传播。8 世纪，欧洲开始了加罗林王朝的文艺复兴。这时西欧各民族开始有民族文字的萌芽。这个时期的欧洲，《圣经》是主要的、甚至是唯一的文字读物。拉丁字母随《圣经》的传播而传播。有的在拉丁文《圣经》上用拉丁字母书写本民族语言的注释，有的创造拉丁字母的本民族文字来翻译《圣经》。直到 14—15 世纪的文艺复兴时期，欧洲各国的民族文字才成熟。这个时期是拉丁字母的第二次大传播。

1498 年哥伦布发现南美新大陆。不久后欧洲开始产业革命，随着欧洲帝国的殖民扩张，拉丁字母传播到拉美、非、亚广大的地区。这是拉丁字母的第三次大传播。有的地区创造了拉丁字母的民族文字。有的地方语言、文字同化于宗主国，而宗主国绝大多数都使用拉丁字母的文字。

第二次世界大战之后，由于航空和通讯技术的发展，世界在变小，文字信息的交际变得更频繁、更便捷、更重要。科学技术上，由于北美、西欧的优势，使拉丁字母成为国际上最重要的科技文字，电子计算机软件的最初版本，也几乎都是以拉丁字母为基础的。以后的多文种版本也都是与拉丁字母字符集兼容。非拉丁字母的文字，普遍地建立了用拉丁字母表示的转写方法。这一切都使拉丁字母的应用更扩大了。这个时期是拉丁字母的第四次大传播。

上面提到的拉丁字母的四次大传播，第一、第三两次主要是地域的扩大；第二、第四两次更表现为应用领域广度和深度的扩大。其他文字圈，或大或小都有为拉丁字母取代的情况。

4. 拉丁字母是现今世界上应用最广泛的字母。拉丁字母

用于描述世界上众多的语言，而基本保持了字母体系的稳定，不似印度文字圈和阿拉伯文字圈那样内部变化巨大、纷繁。它主要采用了如下变通方法：

①用组合的字母串表示单一的辅音或元音音素；②一符两用或多用；③使用附加符号；④增加新字母。这四种办法中，前三种用得最普遍。这是保持文字体系基本稳定的方法。而增加新字母，给文字机械化造成甚大困难。

机械打字机、电传打字机，在增加新字母时，设计、制造中，容易形成批量多、数量少的难于标准化的局面。也给培训和使用带来不便。但就全球范围看，各拉丁字母系文字中所增加的新字母总量仍然十分可观。1992 年通过的国际编码标准 ISO 10646 中，变形或带附加字符的拉丁字母已收入近千个。电脑化和机械化有很大不同：机械化靠增加机械复杂性解决问题，电脑化可以靠增加软件复杂性解决问题。增加新字符给电脑化系统带来的困难和麻烦不象机械化系统那样严重棘手。

5. 基督教的传播对拉丁字母的传播有过巨大的推动。最早的《圣经》是用希伯来文写成。3 世纪译为希腊文，公元 383 年由希腊文译为拉丁文。圣经和相应的传播活动推动了拉丁字母的传播。在欧洲，12 世纪从中国传入造纸术，1439 年开始出现印刷所，1546 年产生印刷的《圣经》。《圣经》是欧洲古代及中世纪的主要的、甚至是唯一的文字读物。中国造纸和印刷术为《圣经》的广泛传播提供了技术条件。欧洲各国民族文字成熟于 14—16 世纪，这和中国印刷、造纸术的传入及《圣经》的传播有密切的关系。

拉丁字母和斯拉夫字母在东欧的争夺，和基督教与东正

教的宗教斗争密切联系。西欧向拉美、亚、非的殖民扩张是伴随着基督教的精神、文化侵略的。在传播基督教过程中也传播了拉丁字母。

近代，土耳其由阿拉伯字母改为拉丁字母，是由国内解除伊斯兰宗教束缚的政治革命促成的。印度尼西亚历史上经历了梵文字母——阿拉伯字母——拉丁字母的变化，是与佛教、伊斯兰教、基督教在印尼统治地位的更迭相联系的。

本文的具体材料，大多引自周有光先生的专著《世界字母简史》（上海教育出版社，1990 年版）周先生在书中把拉丁字母的世界性传播比作水中波圈的扩散，一圈大于一圈。周先生把历史上的扩散分为六个时期，比作六个波圈。书中还有各国拉丁字母文字的字母表。

斯拉夫文字圈概况

刘晓波

1. 斯拉夫文字圈的地域范围小于拉丁文字圈，也稍小于阿拉伯文字圈。使用斯拉夫字母的国家有东欧的俄罗斯、乌克兰、白俄罗斯和摩尔达维亚，外高加索的阿塞拜疆，中亚的乌兹别克、土库曼、塔吉克、吉尔吉斯和哈萨克（以上10国是原苏联的加盟共和国），保加利亚，塞尔维亚和马其顿（以上两国为原南斯拉夫联邦的共和国）以及亚洲的蒙古。现在有14个国家操60多种语言的各民族人民（人口近二亿八千万）使用着斯拉夫文字。

2. 斯拉夫字母直接传承于希腊字母，它形成于公元9世纪，比拉丁字母晚1600年。保留至今的古斯拉夫语文献曾使用两种字母记载：格拉戈尔字母和基里尔字母。

格拉戈尔字母的名称来源于ГЛАГОЛ一词。在格拉戈尔字母表中，很多字母来自希腊古体小写字母，有些字母是在萨马里亚文字和希伯来文字的基础形成的。这套字母表在字母成分、字母排列的顺序和音值方面与基里尔字母表几乎完全一致，但在书写形式上两者却有极大的差别。字母形状很独特，使人看不出它与当时的哪一种字母有联系。尽管如此，这确能间接地证明它与基里尔字母表的关系。古斯拉夫语文献表明，格拉戈尔字母在西斯拉夫人中间曾广为使用，特别是当时的大摩拉维亚王国，后来又传入保加利亚和克罗地亚，一直使用到18世纪。在古代俄罗斯很少使用。关于格拉戈尔字母