

XIANDAI HANZI

TEZHENG FENXI  
YU JISUAN YANJIU

# 现代汉字 特征分析与计算研究

邢红兵 著

 商务印书馆  
THE COMMERCIAL PRESS

# 现代汉字特征分析与 计算研究

邢红兵 著

商 务 印 书 馆

2007 年·北京

**图书在版编目(CIP)数据**

现代汉字特征分析与计算研究/邢红兵著. —北京:  
商务印书馆, 2007

ISBN 7 - 100 - 05310 - 2

I. 现… II. 邢… III. 汉字 - 研究 - 现代  
IV. H12

中国版本图书馆 CIP 数据核字(2006)第 144384 号

**所有权利保留。**

**未经许可,不得以任何方式使用。**

XIÀNDÀI HÀNZÌ TÈZHÈNG FÈNXÍ Yǔ JÌSUÀN YÁANJIŪ

**现代汉字特征分析与计算研究**

邢红兵 著

---

商 务 印 书 馆 出 版

(北京王府井大街 36 号 邮政编码 100710)

商 务 印 书 馆 发 行

北京市白帆印务有限公司印刷

ISBN 7 - 100 - 05310 - 2/H · 1277

---

2007 年 7 月第 1 版

开本 850 × 1168 1/32

2007 年 7 月北京第 1 次印刷

印张 7%

定价:15.00 元

# 目 录

## 上编 汉字的特征分析

|                              |    |
|------------------------------|----|
| 第一章 汉字的使用 .....              | 3  |
| 1.1 汉字的字量 .....              | 3  |
| 1.2 汉字字集 .....               | 4  |
| 1.2.1 “GB 13000.1 字符集” ..... | 4  |
| 1.2.2 现代汉语常用字表 .....         | 5  |
| 1.2.3 “现代汉语研究语料库系统”汉字 .....  | 7  |
| 1.2.4 小学语文教材汉字 .....         | 8  |
| 1.3 汉字的使用 .....              | 10 |
| 1.3.1 从小学语文教材看汉字的使用 .....    | 11 |
| 1.3.2 从语料库看汉字的使用 .....       | 14 |
| 第二章 汉字字形特征 .....             | 17 |
| 2.1 汉字的笔画 .....              | 17 |
| 2.1.1 汉字的笔形 .....            | 17 |
| 2.1.2 汉字的笔画数 .....           | 19 |
| 2.1.3 汉字笔画趋简率 .....          | 22 |
| 2.2 汉字的基础部件 .....            | 25 |
| 2.2.1 部件的类型 .....            | 25 |

## 2 现代汉字特征分析与计算研究

|                        |    |
|------------------------|----|
| 2.2.2 基础部件的数量 .....    | 26 |
| 2.2.3 成字部件和非成字部件 ..... | 30 |
| 2.2.4 基础部件的构字能力 .....  | 33 |
| 2.2.5 基础部件的典型位置 .....  | 35 |
| 2.2.6 基础部件的构形 .....    | 38 |
| 2.3 汉字的构造 .....        | 43 |
| 2.3.1 汉字的结构类型 .....    | 43 |
| 2.3.2 汉字的结构层次 .....    | 51 |
| 2.3.3 汉字的构造 .....      | 56 |

## 第三章 汉字形声字声旁表音研究 .....

|                             |    |
|-----------------------------|----|
| 3.1 汉字的构字法 .....            | 59 |
| 3.2 形声字声旁表音的统计研究 .....      | 61 |
| 3.3 小学语文教材形声字表音情况分析 .....   | 66 |
| 3.3.1 小学语文教材形声字数据库的建立 ..... | 66 |
| 3.3.2 小学语文教材形声字数量 .....     | 68 |
| 3.3.3 小学语文教材形声字的使用 .....    | 70 |
| 3.3.4 小学语文教材形声字的家族 .....    | 72 |
| 3.3.5 小学语文教材形声字的表音特点 .....  | 74 |

## 第四章 形声字的认知研究 .....

|                              |    |
|------------------------------|----|
| 4.1 早期的汉字认知研究 .....          | 83 |
| 4.2 80年代以后的汉字认知研究 .....      | 84 |
| 4.2.1 汉字的构成成分和汉字的正字法规则 ..... | 85 |
| 4.2.2 汉字的亚词汇加工 .....         | 86 |
| 4.2.3 汉字的语音意识 .....          | 87 |
| 4.2.4 儿童发展性阅读障碍 .....        | 88 |

|                           |    |
|---------------------------|----|
| 4.3 形声字字音加工的认知研究 .....    | 89 |
| 4.3.1 形声字的规则性效应及其发展 ..... | 89 |
| 4.3.2 形声字的一致性效应及其发展 ..... | 92 |

## 下编 汉字的计算与模拟研究

|                           |     |
|---------------------------|-----|
| 第五章 联结主义理论和自组织模型 .....    | 99  |
| 5.1 语言习得理论 .....          | 99  |
| 5.2 联结主义理论 .....          | 101 |
| 5.3 神经网络模型 .....          | 104 |
| 5.3.1 主要的神经网络模型 .....     | 104 |
| 5.3.2 神经网络的学习 .....       | 105 |
| 5.3.3 神经网络中的信息表征 .....    | 107 |
| 5.4 自组织理论 .....           | 107 |
| 5.4.1 自组织模型简介 .....       | 107 |
| 5.4.2 赫布学习法 .....         | 110 |
| 5.5 语言习得的自组织模型 .....      | 110 |
| 5.5.1 DISLEX 及其相关研究 ..... | 112 |
| 5.5.2 DevLex 及其相关研究 ..... | 114 |
| 5.6 汉字命名的计算机模拟研究 .....    | 116 |
| 5.6.1 模型的结构 .....         | 117 |
| 5.6.2 汉字字形字音的表征 .....     | 117 |
| 5.6.3 模型的模拟结果 .....       | 119 |
| 第六章 汉字表征研究 .....          | 121 |
| 6.1 汉字字形的表征方法 .....       | 121 |
| 6.1.1 前期研究 .....          | 122 |

#### 4 现代汉字特征分析与计算研究

|                                                     |     |
|-----------------------------------------------------|-----|
| 6.1.2 汉字字形表征 .....                                  | 125 |
| 6.2 汉字字音的表征方法 .....                                 | 136 |
| 6.3 形声字字形的复合表征 .....                                | 140 |
| 第七章 自组织模型的建造 .....                                  | 143 |
| 7.1 模拟软件 DISLEX .....                               | 143 |
| 7.2 汉字表征数据库及表征程序 .....                              | 148 |
| 7.2.1 汉字属性数据库 .....                                 | 148 |
| 7.2.2 基础部件属性数据库 .....                               | 150 |
| 第八章 形声字命名的计算机模拟研究 .....                             | 151 |
| 8.1 汉字输入、输出词典中字形、字音的储存方式 .....                      | 151 |
| 8.1.1 主要术语和参数说明 .....                               | 151 |
| 8.1.2 前期研究 .....                                    | 153 |
| 8.2 形声字命名模拟研究 .....                                 | 156 |
| 8.2.1 模型命名形声字的策略 .....                              | 156 |
| 8.2.2 模型命名形声字的规则性效应及其发展 .....                       | 161 |
| 8.2.3 模型命名形声字的一致性效应及其发展 .....                       | 173 |
| 8.3 计算机模拟研究的本质特征 .....                              | 177 |
| 8.3.1 语言材料库的建设及其在模型中的应用 .....                       | 177 |
| 8.3.2 汉字的字形和字音的表征 .....                             | 181 |
| 8.3.3 模型的知识获得和知识输出 .....                            | 183 |
| 8.3.4 进一步的研究 .....                                  | 186 |
| 参考文献 .....                                          | 189 |
| 附录1:《信息处理用 GB 13000.1 字符集汉字部件规范》<br>中的“基础部件表” ..... | 198 |

|                                      |     |
|--------------------------------------|-----|
| 附录 2:部件视觉属性分组 .....                  | 204 |
| 附录 3:汉语音素表征方法及赋值表.....               | 211 |
| 附录 4:模型文件样例 .....                    | 213 |
| 附录 5:五年级模型命名不熟悉形声字测试材料 .....         | 217 |
| 附录 6:模型命名形声字的规则性效应及其<br>发展测试材料 ..... | 218 |
| 附录 7:模型命名形声字的一致性效应及其<br>发展测试材料 ..... | 221 |
| 后记 .....                             | 223 |

## 上编 汉字的特征分析



# 第一章 汉字的使用

## 1.1 汉字的字量

汉字到底有多少,实际上没有准确的答案。我们从几部大的字典收录的字数来看,至少有5万多字有据可查。其中公元1716年成书的《康熙字典》收47 043字,1915年成书的《中华大字典》收44 908字;1990年完成的《汉语大字典》收54 678字。从这些数据可以看出,汉字的总数量是很大的。

汉字的数量很多,但是汉字的使用情况并不是均衡的,大部分汉字很少用或者罕用。这从汉字字频统计、常用字表的研制等可以看出。1928年出版《语体文应用字汇》,该字汇对554 478字的六种语料进行统计,得出4 261字,排前20位的字是:“的不一了是我上他有人全这来小在们说子可道”。1977年《汉字频度表》包括6 374个字,这些汉字是在2 160多万字材料中统计出来的,这些汉字成为《信息交换用汉字编码字符集·基本集》(6 763汉字)的主要依据。1984年《汉字频度统计》在此基础上重新整理,共有5 991字,排在前10位的是:“的一是在了不和有这”。1986出版的《现代汉语频率词典》,主要对词频、字频、字的组词能力等进行了统

## 4 现代汉字特征分析与计算研究

计,统计的语料为 180 万字,共得到 4 574 字,前 10 位的字是:“的一了是不我在有人这”。1992 年出版的《现代汉语字频统计表》是 1981 年开始由北京航空学院计算机系进行的,该字表统计的语料为 1.38 亿字,得出汉字数为 7 754 个,最后整理出 13 个字频表。

从这些统计数据来看,具有一定使用频度的现代汉字在 8 000 字以下,而通用领域所使用的汉字一般不超过 5 000 字。

### 1.2 汉字字集

我们这里首先介绍一下本书中涉及的几个汉字字集,主要包括:“GB 13000.1 字符集”、“现代汉语常用字表”、“现代汉语研究语料库系统汉字”、“小学语文教材汉字”等。下面分别论述。

#### 1.2.1 “GB 13000.1 字符集”

国际标准化组织(ISO)于 1984 年就开始研究、制定《信息技术通用多八位编码字符集(UCS)》国际标准,即 ISO/IEC 10646。到 1993 年,该标准的第一部分:体系结构与基本多文种平面(即 ISO/IEC 10646.1)正式发布,即《信息技术 通用多八位编码字符集(UCS)第一部分:体系结构与基本多文种平面》(GB 13000.1 - 1993),这个标准规定将其一个区 I 区用于中、日、韩(CJK)统一的表意文字编码,以现有的各标准字符集为源字符集,将其中的汉字按统一的认同原则进行认同和甄别后,生成涵盖各源字符集;按东亚著名的四大字典(《康熙字典》、《大汉和字典》、《汉语大字典》、《大字源》)的页码和字位综合排序的 CJK 汉字统一字汇和字序,构

成 UCS 中的表意文字部;其中共包括20 902个汉字字符,一般简称为“GB 13000.1 字符集”。

在“GB 13000.1 字符集”20 902个汉字中(以下简称“20902 汉字”),中国大陆提出的汉字(G 列)为17 124个;中国台湾提出的汉字(T 列)为17 258个。G 列与 T 列的并集,即中国方面提出的汉字为20 158个,G 列与 T 列的交集为14 224个汉字。交集小于并集的一大原因是台湾地区目前不使用简化字。日本提出的汉字(J 列)为12 157个,其中包括中国未提出的 690 个,这部分暂称 J 列专用字。韩国提出的汉字(K 列)为7 477个,其中包括中国未提出的 90 个,这部分暂称 K 列专用字。J 列和 K 列的并集,即非中国提出的汉字共有 744 个。

### 1.2.2 现代汉语常用字表

“现代汉语常用字表”是国家语言文字工作委员会和国家教育委员会 1988 年 1 月 26 日联合发布的。该字表由国家语言文字工作委员会汉字处负责研制。该字表是在统计了 15 种常用字材料和 5 种通用字材料的基础上制定的(傅永和,1989b)。15 种常用字材料如下:

- (1)陈鹤琴,1928 年 6 月:语体文应用字汇(4 261字);
- (2)四川省教育科学院,1946 年 8 月:常用字选(2 000字);
- (3)中央人民政府教育部社会教育司,1950 年 9 月:常用汉字登记表(1 017字);
- (4)中央人民政府教育部,1952 年 6 月公布:常用字表(2 000字);

## 6 现代汉字特征分析与计算研究

(5)中国大辞典编纂处,1954年1月:汉字正音3500字表(3559字);

(6)山东省教育厅,1958年8月:普通话常用字表(3000字);

(7)北京市教育局中小学教材编审处,1965年3月:常用字表(3100字);

(8)中国文字改革委员会汉字组,1975年:4500字表(4444字);

(9)七四八工程查频组,1976年12月:汉字频度表(常用字部分,4152字);

(10)中国文字改革委员会汉字组,1979年7月:增订2500字表;

(11)国家标准局,1981年5月发布:信息交换用汉字编码字符集·基本集(一级字,3755字);

(12)人民教育出版社,1984年:六年制小学语文统编教材生字表(3189字);

(13)中国人民大学语言文字研究所,1984年3月:常用构词字典(3994字);

(14)北京航空学院计算机科学与工程系、中国文字改革委员会汉字处,1985年3月:社会科学、自然科学综合汉字频度表(3500字);

(15)北京语言学院语言教学研究所,1985年7月:汉字频率表(4574字)。

5种通用字材料是:

(1)中华人民共和国文化部和文字改革委员会1965年1

月联合发布:印刷通用汉字字形表(6 196字)。

(2)中国社会科学院语言研究所,1978年8月:现代汉语词典。

(3)商务印书馆,1979年:新华字典。

(4)中华人民共和国邮电部,1983年:标准电码本(7 292字)。

(5)新华社技术研究所,1987年1月:1986年度新闻信息流通频度(6 001字)

“现代汉语常用字表”(以下简称“常用汉字”)的选字范围是1928年至1986年,在此时段中抽样量按时间越近,抽样量越增的原则,统计了汉字在不同学科的分布和使用情况。汉字抽取的原则以汉字的使用频率为主,考虑了分布度和使用度,也考虑了汉字的构字能力和构词能力等因素等。最后该字表共收录了3 500个汉字,包括常用字2 500个,次常用字1 000个。

“常用汉字”定稿后,通过山西大学计算机系的检测发现:2 500字的覆盖率达到97.97%,1 000次常用字的覆盖率达到1.51%,3 500字的覆盖率达99.48%。

### 1.2.3 “现代汉语研究语料库系统”汉字

国家教育部“八五”人文、社会科学规划项目、北京语言大学(原“北京语言学院”)“八五”重点科研项目——“现代汉语研究语料库系统”于1996年1月29日在北京语言大学通过了专家鉴定。该语料库的核心是200万字规模的精语料。这200万字语料是在第一级的2 000万字语料中按照设定的比例由程序随机抽取出来的。其中包括书面语语料160万字和准口语语料40万字。书面语语料按题材分为10大类:政治和法律类30万字,经济类30万

## 8 现代汉字特征分析与计算研究

字,文学类 37.5 万字,文化教育类 15 万字,社会生活类 15 万字,科技和科普类 12 万字,体育类 8 万字,地理和旅游类 5 万字,历史类 2.5 万字,军事类 5 万字。准口语材料中对话(话剧脚本)部分 25 万字,独白部分 15 万字。这 200 万字的语料全部进行了分词和词性标注。

我们提取了精语料中的全部汉字,并统计了汉字的频度。去除标点、符号、英文字母以及一些错字、“死字”等,该语料库所使用汉字共 4 868 个(以下简称“语料库汉字”)。

### 1.2.4 小学语文教材汉字

“小学语文教材汉字”来源于 5 套小学语文教材的课文正文。这五套教材是:

(1)“五四”学制教材总编委会编著,北京师范大学出版社 1999 出版的《九年义务教育五年制小学试用课本》(以下简称“师大教材”);

(2)北京教育科学研究院基础教育教学中心编著,北京出版社 1998 出版的《九年义务教育六年制小学试用课本》(以下简称“北京教材”);

(3)人民教育出版社小学语文室编著,人民教育出版社 1998 年出版的《九年义务教育六年制小学教科书》(以下简称“人教教材”);

(4)九年义务教育教材(沿海地区)编写委员会编写,广东教育出版社 1991 年出版的《九年义务教育六年制小学试用课本》(第九版,1998 年,以下简称“广东教材”);

(5)张庆、朱家珑主编,江苏教育出版社 2001 年出版的《九年义务教育六年制小学课本(试用修订本)》(以下简称“江苏教材”)。

除了“师大教材”是 10 册以外,其他 4 套教材都是 12 册。5 套教材全部字符数是(包括汉字、标点符号以及阿拉伯数字、英文字符等非汉字符号)825 186 个,去除非汉字符号、标点符号后是 711 219 字,表 1-1 显示了各套教材的汉字总数。

表 1-1 5 套小学语文教材汉字数统计表

| 教材名称 | 总册数 | 总字符数    | 汉字总数    | 平均每册字数    |
|------|-----|---------|---------|-----------|
| 师大教材 | 10  | 170 038 | 146 457 | 14 645.70 |
| 人教教材 | 12  | 174 319 | 149 857 | 12 488.08 |
| 北京教材 | 12  | 185 591 | 160 342 | 13 361.83 |
| 江苏教材 | 12  | 124 964 | 108 261 | 9 021.75  |
| 广东教材 | 12  | 170 274 | 146 302 | 12 191.83 |
| 合计   |     | 825 186 | 711 219 | 12 262.40 |

我们首先生成各套教材的字表,并统计各套教材、各个册数汉字的使用次数,然后再将 5 套教材的字表合并,统计汉字的总频度,并且标注各套教材的频度,最后形成小学语文教材汉字频度表。

从字形是否相同的角度来看,汉字形、音、义的关系包括以下三种情况:(1)形音相同,意义相关,例如“打(~架)”、“打(~球)”,这类字是多义字;(2)形音相同,意义不同,例如“别(告~)”和“别(~来)”,这类字是同形字;(3)字形相同,音义不同,例如“行(háng)”和“行(xíng)”,这类字是多音字。我们统计的汉字数量并没有将以上三类情况分开考虑,比如“行(háng)”和“行(xíng)”在