

# 目 录

总 序	文秋芳	vii
前 言	许家金	x
第一章 概论		1
1.1 引言		1
1.2 语料库研究方法中的文本		1
1.3 语料库研究方法中的工具		3
1.3.1 经典时代的语料库研究方法简述		5
1.3.2 后经典时代语料库研究方法的特点		6
1.3.3 两个时代的语料库研究方法评述		9
1.4 语料库研究方法的理论贡献		12
1.5 结语		13
第二章 语料库建设的一般方法		15
2.1 建库原则		15
2.2 文本规格		19
2.3 语料标注		33
第三章 语料库建设的网页爬取方法		39
3.1 R 语言静态网页爬取		39
3.2 R 语言动态网页爬取		44
3.2.1 微博搜索话题博文的爬取		44
3.2.2 微博博主博文的爬取		49
3.3 基于爬虫软件的爬取		54

<b>第四章 语料检索及语料清洗</b> .....	60
4.1 检索语料 .....	64
4.2 语料清洗 .....	77
<b>第五章 数据可视化</b> .....	90
5.1 数据分布 .....	90
5.1.1 箱线图 .....	92
5.1.2 小提琴图 .....	93
5.1.3 密度图 .....	94
5.1.4 峰峦图 .....	95
5.2 数据比较 .....	97
5.2.1 条形图 .....	99
5.2.2 棒棒糖图 .....	100
5.2.3 雷达图 .....	101
5.3 数据演变 .....	102
5.3.1 折线图 .....	104
5.3.2 面积图 .....	105
5.3.3 动态图 .....	108
5.4 数据关系 .....	109
5.4.1 散点图 .....	110
5.4.2 气泡图 .....	112
5.4.3 热力图 .....	113
5.5 其他数据 .....	114
5.5.1 饼图 .....	114
5.5.2 网络图 .....	117
5.5.3 词云图 .....	119
<b>第六章 短语学分析</b> .....	122
6.1 语料库短语学研究实践 .....	122

---

6.1.1 搭配基础研究	123
6.1.2 搭配扩展研究	125
6.1.3 局部语法研究	128
6.2 语料库短语学的分析方法	133
6.2.1 搭配分析	133
6.2.2 扩展意义单位描写	136
6.2.3 局部语法构型描写	141
<b>第七章 搭配构式分析</b>	<b>149</b>
7.1 共现词分析	149
7.2 显著共现词分析	156
7.3 共变共现词分析	161
<b>第八章 对应分析</b>	<b>166</b>
8.1 简单对应分析	166
8.2 多重对应分析	170
<b>第九章 主成分分析和因子分析</b>	<b>184</b>
9.1 主成分分析和因子分析的基本概念	184
9.2 主成分分析和因子分析在语言学领域的应用	184
9.3 主成分分析和因子分析的相似性和差异性	187
9.4 案例分析：一百多年间汉语书面语的语域演变研究	188
<b>第十章 多维尺度分析</b>	<b>198</b>
10.1 多维尺度分析的基本概念	198
10.2 多维尺度分析在语言学领域的应用	198
10.3 多维尺度分析、因子分析和聚类分析之间的相似性和差异性	201
10.4 案例分析：现代原创汉语和翻译汉语的分期和历时演变	201

<b>第十一章 聚类分析</b> .....	209
11.1 聚类分析综述.....	209
11.2 聚类分析种类.....	211
11.3 案例实操.....	215
11.3.1 层次聚类分析：数值数据.....	215
11.3.1.1 案例背景介绍.....	215
11.3.1.2 操作及代码.....	216
11.3.2 层次聚类分析：分类数据.....	222
11.3.2.1 案例背景介绍.....	222
11.3.2.2 操作及代码.....	225
11.3.3 划分聚类分析：分类数据.....	234
11.3.3.1 设定聚类数量.....	235
11.3.3.2 选择划分方式.....	235
11.3.3.3 选择距离度量.....	236
11.3.3.4 确定最优分类.....	236
11.3.3.5 操作及代码.....	236
<b>第十二章 决策树和随机森林</b> .....	240
12.1 决策树.....	241
12.1.1 经典决策树.....	243
12.1.2 条件推断树.....	244
12.1.2.1 算法介绍.....	244
12.1.2.2 语言研究中的运用.....	245
12.1.2.3 案例实操.....	246
12.2 随机森林.....	251
12.2.1 基于 CART 的（经典）随机森林.....	254
12.2.2 基于 CTREE 的条件推断森林.....	260

---

<b>第十三章 逻辑斯蒂回归</b> .....	267
13.1 逻辑斯蒂回归模型基本介绍.....	267
13.1.1 模型原理.....	267
13.1.2 前提假设和数据要求.....	269
13.1.3 两种推广.....	273
13.1.3.1 加入随机效应的逻辑斯蒂回归.....	273
13.1.3.2 多分类逻辑斯蒂回归.....	275
13.2 逻辑斯蒂回归在语言学领域的应用.....	275
13.3 二分类任务：英语中的与格交替.....	280
13.3.1 语料获取与处理.....	280
13.3.2 固定效应模型.....	286
13.3.2.1 模型拟合与解读.....	286
13.3.2.2 模型优度评估.....	294
13.3.2.3 模型诊断.....	299
13.3.3 混合效应逻辑斯蒂回归模型.....	303
13.4 多分类任务：德语中的三种回指形式选择.....	308
13.4.1 语料获取与处理.....	309
13.4.2 两种方式拟合多分类模型.....	313
13.4.2.1 使用 <code>nnet</code> 程序包.....	313
13.4.2.2 使用 <code>polytomous</code> 程序包.....	321
<b>第十四章 语料库研究方法展望</b> .....	327
14.1 语料库建设展望.....	327
14.2 语料库分析技术展望.....	329
14.3 语料库语言学理论建构展望.....	331
<b>参考文献</b> .....	333
<b>附录 A：英汉双语语料库语言学学科术语表</b> .....	363
<b>附录 B：常用正则表达式小结</b> .....	379