# 《数理语言学》

## 图书基本信息

书名：《数理语言学》

13位ISBN编号：9787100083911

10位ISBN编号：7100083915

出版时间：2012-4

出版社：商务印书馆

作者：冯志伟,胡凤国

页数：491

# 《数理语言学》

## 内容概要

本书系统地、全面地、深入浅出地介绍了这三个部分的基本知识和最新成就。为了便于文科读者透彻地理解本书内容，本书专门辟出一章讲述语言学中的离散数学方法。
本书可作为数理语言学的入门教材，著者在写作时尽量考虑到跨学科读者的需要，既可供想了解这门新兴边缘学科而数学准备不够的语言学工作者和其他文科读者阅读，亦可供要求了解语言学方面的现代化知识的理工科读者阅读。

# 《数理语言学》

## 书籍目录

前言
第1章 离散数学与语言
第2章 代数语言学
第3章 统计语言学
第4章 应用数理语言学
结语
附录：胡耀邦同志鼓励我研究数理语言学

# 《数理语言学》

## 精彩短评

1、还没来得及看里头的内容，是老师推荐的一本书
2、内容很好，是大家写的，赞一个
3、瞄一眼的读过……
4、很多知识都已经较为陈旧，未来的数理属于概率推断
5、研一的时候读过，适合学语言学的人读读
6、数理语言学，计算语言学，计量语言学，三者应当分清关系，本书无疑是一本好书。

# 《数理语言学》

## 章节试读

1、《数理语言学》的笔记-语言与信息

语言的使用是一个随机过程，单单将话语中的一个单独成分（比如词）的概率分布拿出来研究，是无法窥探语言使用的全貌的。如果将同一话语中其它成分一起考虑，概率分布就会有显著的不同。这一不同部分原因在于语言有结构性，知道了语言的前面一部分，可以以较大概率推测出后面一小部分。这就像一个链条，从一头开始一点点拉，如果一个语言的各个语言成分出现概率相互独立，且相等，这对使用这一语言的人来说，真是一不小的记忆灾难。更现实的情况是，目前已知的语言，其各个语言成分的出现概率不相等，而且都不是相互独立的。这些思考其实反应了一个更本质的问题，即语言的熵。汉语的信息熵约为9.71比特。
书中举了关于不同阶马尔可夫链的例子。随着马尔可夫链的阶数的增大，语言也越来越接近有意义的文本，也更容易记忆。英文词的一重马尔可夫链如下 ：The head and in frontal attack on an English writer that the character of this point is therefore ancther method for the letters that the time of who ever told the problem for an unexpected
英文词的二重马尔可夫链如下 ：Family was large dark animal came roaring down the middle of my friends love books passionately every kiss is fine
英文词的四重马尔可夫链如下 ：Road in the country was insane especially in dreary rooms where they have some books to buy for studying Greek

2、《数理语言学》的笔记-代数语言学

美国语言学家M. Joos指出，语言这个符号系统在本质上是由一些离散的单元组成的，它不容许与连续性有半点妥协，因此语言可以看成是一个在严格意义上的量子机制，凡是与连续性有关的一切，都得排除在语言的范围之处。不知道Joos在知道Mikolov在发现词向量相互的偏移关系后，会不会改变自己的想法？

[1] F. Harary, H. Paper, Toward a general calculus of phonemic distribution, Language, Vol. 33, No.2, pp. 143-169
[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

3、《数理语言学》的笔记-词的频率分布规律

词频算是自然语言处理中常用的特征。从搜索引擎到垃圾过滤，从分词到抽取评价对象，很多经典的方法都会以词的频率为基础，再经过一系列变换，得到应用所需的特征。但其实最令人们关注的还是词，或者说文本本身有什么更好、更本质的特征，这样就不必为每种文本挖掘应用寻找特征了。目前来说，词向量是一个方向，但会不会有好的结果，还需要时间的检验。书中记载的关于词频的研究成果甚至是在有计算机之前得到的，
事实上关于词频($n_r$)与依词频排序得到的词序号r之间的关系，可以追溯到1916年Estoup的工作，他已经发现词频与序号之间的乘积大体上稳定于一个常数：($$\large n_r \centerdot r = K$$)，Condon对词频和序号取了对数，发现两者在坐标上的分布更加接近一条直线。两者的关系可以如下推导：假设($x=log r,y=log n,OB=log k, OA=\frac{OB}{tan\alpha}$)其中OA和OB为直线与X轴和Y轴的截距。($$\large
\begin{align*}
\frac{x}{OA} + \frac{x}{OB} &= 1\\
\frac{log r}{\frac{log k}{\gamma}} + \frac{log n_r}{log k} &= 1\\
\gamma logr + log n_r &= log k\\
n_r &= k r^{-\gamma}\\
\frac{n_r}{N} &= \frac{k}{N} r^{-\gamma}\\
f_r &= c r^{-\gamma}\\

\end{align*}$$)话说保留一个原始的公式就很好了，但Condon根据实验结果进行了化简，比如直线的角度为45度，并认为c是个常数。Zipf在Condon的基础上，对更多规模的文本数据进行分析，得到了相同的结果，又写了本《Psycho-Biology of Language》，觉得c应该是个参数，这样这一规律就被称为齐夫定律（Zipf's Law）。1936年Joos又认为($\gamma$)也应该是个参数，不一定就是45度，20世纪50年代，Mandelbrot经过苦逼的数学推导，又加了一个参数，这样公式最终变为($$\large f_r = c (r+a)^{-\gamma}$$)但无论怎么样在原始公式上进行改进，都还是不能解决一个问题，也就是频率相同的词的词频与序号的关系，或许当语料足够大时，这一问题就会迎刃而解。红色为Zipf直线，直线末尾的实际分布是破碎的，这是因为相同频率的词的出现

# 《数理语言学》

## 版权说明