基于标注语料库的先秦汉语词汇统计*

李斌[1,2] 冯敏萱[1] 陈小荷[1]

1. 南京师范大学语言信息科技研究中心 南京 210097

2. 南京大学计算机软件新技术国家重点实验室 江苏南京 210093

E-mail: lib@nlp.nju.edu.cn

摘 要：先秦汉语在汉语发展史上具有非常重要的作用。由于缺乏深度标注的语料库，先秦汉语的词汇概貌一直难以展现。本文基于 25 种先秦传世文献人工分词和词性标注的语料库，统计得到了先秦汉语字、词频率和分布概貌，计算了先秦汉字熵，详细讨论了学界关心的多音节词数量和词语兼类情况。

关键词：中文信息处理，先秦汉语，词汇统计，词语兼类

Corpus Based Lexical Statistics of Pre-Qin Chinese

Li Bin[1,2] Feng Min-xuan[1] Chen Xiao-he[1]

1. Research Center of Language and Informatics, Nanjing Normal University, Nanjing 210097

2. State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing 210093

E-mail: lib@nlp.nju.edu.cn

Abstract: The Pre-Qin Chinese plays a key role in the history of Chinese. However, for the lack of the tagged corpus, the overview of Pre-Qin Chinese vocabulary is still not clear. This paper introduces the corpus of 25 Pre-Qin classical texts, which are under word segmentation and part-of-speech tagging manually. Then, the character and word frequencies are showed based on the corpus. The character entropy, the syllables of words and the multiple part-of-speech words are also statistically analyzed.

Keywords: Chinese information processing, Pre-Qin Chinese, lexical statistics, multiple part-of-speech word.

1 前言

在汉语的发展史上，先秦汉语具有举足轻重的作用，字词的音形义的考证都需要从先秦汉语找寻最早的用例。先秦汉语的词汇研究已经有了很多专著和论文加以论述，如《吕氏春秋》、《韩非子》等专书的词汇研究（张双棣 1989；车淑娅 2008）。但是这些研究大都是针对专书或者领域词汇的，注重字词在音、形、义方面的考证和整理，而对先秦词汇的概貌研究极为少见（叶正渤 2007）。台湾中央研究院近年来开发了以十三经为主体的先秦汉语语料库和先秦金文简牍词汇数据库[1]，对重要的传世文献进行了词语切分和词性标注，提供了在线检索程序和词典。该语料本可以作为先秦汉语词汇全貌的研究基础，可惜该数据库为检索系统，不公开全文数据库，难以为广大研究者直接使用。

因此，通过近四年的努力，我们甄选了 25 部先秦传世文献，进行全面的分词和词性标校工作，形成了先秦汉语精加工语料库（陈小荷 2008；石民 2010）。借助计算机和语料库技术，本文针对以下两个问题进行了统计分析工作：（1）先秦语料的基本字汇和汉字熵；（2）先秦语料的基本词汇和词类分布情况，词语兼类的比例情况。通过这两个方面的统计，希望对以先秦文献为代表的先秦汉语有一个概貌性的描写。

2 语料来源和说明

本文的数据来源是南京师范大学文学院语言科技实验中心按照自行设计的分词和标注规范（参见表 2），半自动地手工标校的 25 部先秦文献语料库，按篇幅大小依次为：左传、管子、韩非子、吕氏春秋、礼记、墨子、荀子、国语、仪礼、庄子、周礼、公羊传、晏子春秋、谷梁传、孟子、诗经、尚书、楚辞、周易、商君书、论语、老子、孙子兵法、吴子、孝经，共计 133 万汉

*本文承国家社科基金（10CYY021、10&ZD117）、江苏省哲社重点研究基地课题（2010JDXM023）、南京大学计算机软件新技术国家重点实验室开放课题（KFKT2011B03）、中国博士后基金（2012M510178）、江苏省博士后基金（1101065C）、江苏高校优势学科建设工程的资助。

[1] 网址：http://inscription.sinica.edu.tw/c_menu.html。

字。为了统计字频和词频的各种信息，我们把每部文献的字频和词频汇总形成 2 张数据表：字频信息表和词频信息表。通过这两张表，可以方便地查询每个字词在不同文献中的分布情况、每部文献的词型数和词例数。

本文的统计主要区分“型”和“例”。“型（type）”可以是字型、词型等，表示一个词语的形式；“例（token）”则是某一个型在语料中的用例。如汉字型“之”在《左传》中出现了 7260 例。利用数据库查询技术，可以进一步得到字词的概率信息、每本书的最高频或特有字词以及单音节、多音节词的分布情况。本文的统计数据大都基于这两张表展开。为了和现代汉语做比较，我们将北京大学计算语言学研究所公开的 1998 年 1 月人民日报语料（后简称 199801）作为对比语料。

本文的研究设想是基于先秦标注语料库，尝试给出先秦汉语先秦文献的基本字汇和词汇。这在理论上来说，存在一定的困难。“基本词汇”是由孙伏园于 1947 年的《基本词汇研究述要》中提出的。潘允中（1959）指出，基本词汇是语言中最本质的东西，它具有三个特征：历时稳固性、全民性、构词能力强。周荐（1987）认为必须符合稳固性、全民性和能产性三条标准的才算基本词汇的成员。这些定性标准基本上得到了学界的认同，但是操作起来却相当困难。首先，古代汉语的语料库大都没有经过分词和词性标注，无法进行词频统计，只能在纯文本上进行字频统计或用传统的卡片笔记。其次，没有给出完整的定量的分析手段。历时稳固性可以根据《汉语大词典》等大型历时词典来判断，但是这些词典给出的是某字词各个义项最早的用例，并不能体现这些字词在每个时代的全民性和构词能力。而学界共知，古汉语以单音节词为主，在中古以后的古白话或佛经等文本中才谈得上构词能力。全民性则更是现代的网络时代才有可能定量获取的。在古代，能利用的基本上只有文献材料，也只有在文本材料上才有字和词之分，在口语中则为语素和词。因此，我们并不讨论“先秦时代”的基本词汇，研究只限于先秦传世文献的基本字汇和词汇。

那么基本字汇和词汇该如何界定？考虑到先秦文献的历史特殊性，我们在“稳固性、全民性、构词能力强”基础上抽象出两点定量依据，即**“频率高、分布广”**。频率高，是指一个字词在文献中出现的频率高，我们将这些字词称为“高频字词”。使用次数多，则很可能是基本字词。但是，很多词只在某部文献中出现，一两部文献中的高频词也未必在其他文献中出现。所以，还必须考虑一个字词在不同文献中的分布情况。分布文献数量广的字词，我们称为**“通用字词”**。如果一个词，既出现在所有的先秦文献中，频率又很高，那么基本可以认定为基本词汇。不过，这样的界定并非完美，有些文献字数过少、语域窄，一味追求广度，也可能会有缺漏。我们将兼顾这两个方面，尽可能地提供先秦文献的基本字汇和词汇概貌。

3 先秦文献的汉字统计

在了解词汇概貌之前，我们先给出先秦文献语料库的用字情况。25 部文献共有 1334780 个字例，7049 个字型。篇幅最大的是《左传》，共有 179814 个字例，3312 个字型。

3.1 高频汉字和通用汉字

先秦文献中的高频汉字和通用汉字的情况，要依靠频率信息来统计。我们从频度和广度两个角度来进行观察。首先，统计出 25 部文献中频率最高的汉字，然后统计分布度最广即在 25 部文献中均出现的汉字。下面是 25 部文献频率最高的前 100 个汉字：

之、不、也、而、以、其、子、曰、人、者、有、為、則、於、公、君、無、大、故、王、天、所、于、夫、可、是、國、下、矣、民、何、與、乎、上、事、使、行、三、知、一、此、能、侯、言、必、如、得、若、謂、臣、主、道、二、焉、非、齊、十、將、吾、用、諸、然、我、月、在、至、士、禮、命、自、中、師、晉、出、見、五、四、日、皆、生、欲、乃、死、今、成、及、先、從、明、治、相、亦、令、利、入、後、德、食、小、正

不考虑繁简体问题，这些汉字仍然是现代汉语的常用字。几乎每部文献的最高频汉字均为“之”，少数例外体现出文献特色。“之”也是 25 部文献整体上频率最高的汉字，和现代汉语语料中最高频的助词“的”一样。可见助词在先秦汉语和现代汉语中的重要性是一脉相承的。这 100 个汉字存在两个非常有趣的现象：

（1）**在最高频的前 100 个汉字中，有 25 个并没有出现在全部 25 部文献中。**这个现象令人非常吃惊。我们分析了一下这些词语，并没有我们想象得那么奇妙，大多数词都是易于解释的，比如“于”字排名第 23 位，在篇幅较小的《商君书》、《孙子兵法》和《吴子》中没有出现，比较容易理解。但是，《论语》中居然没有出现“此”这个非常常用的汉字，不得不让人有些费解。

我们检索了多个电子版和大型数据库，《论语》确无此字，但是否真的没有这个字，或者为什么没有这个字，还得靠文献学家的考证。我们想强调的是，在语言学的研究中有句俗语“说有易，说无难”，说有和说无，都需要学者多年的积累和扎实的卡片功夫。而基于标注语料库的统计，却可以帮助我们快速地去发现这困难的“无”，更为方便地去考证“无”。

（2）**通用字一般是高频字，但频度未必那么高，且在每部书中的分布也未必均衡。**对于最高频的“之”、“不”等字，在每部书中也基本是最高频的。但是对于通用词的排名较靠后的词语并不尽然，例如“要”这个字排在频率表的 700 位，频次为 296。在《吕氏春秋》和《荀子》中就分别出现了 48 和 38 次，而在《老子》、《孙子兵法》中仅出现 1 次。这种不均衡性也比较容易理解，每部书的领域、年代也各有特点，用字上会有一定的差异。《楚辞》的语气词“兮”频率最高，体现出地域性诗歌特点；《公羊传》、《谷梁传》的判断助词“也”频率最高，体现出“传”这种文体的特色；《论语》的“子”频率最高，更是“子曰”量大和孔子诸多门生的表现。

我们把 25 部文献共用的汉字统计出来，共有的汉字型总计 132 个，按频次由高到低排列如下：

之、不、也、而、以、其、子、曰、人、者、有、為、則、於、公、君、無、大、故、王、天、所、夫、可、是、國、下、矣、民、與、乎、上、事、行、三、知、一、能、侯、言、必、如、得、若、謂、道、非、將、用、然、在、至、士、命、自、中、師、見、五、四、日、皆、生、死、成、及、先、從、明、治、相、利、後、食、小 || 年、未、百、立、聞、地、義、善、心、時、受、文、服、長、重、來、惡、取、足、敢、眾、雖、又、亡、周、已、親、政、失、名、寡、家、陳、安、萬、舉、過、復、居、易、止、始、進、致、高、和、興、退、觀、姓、加、順、厚、離、畏、深、要

“之”不仅频率最高，通用度也高。“||”之前的字表示属于频率最高的 100 个汉字，共计 75 个。而“||”后面的 57 个汉字在频率上排名更为靠后，最后一个“要”字，仅排在频率表的第 700 位。这 132 字都为常用的单字词，各大词类均有分布。由于《老子》、《孙子兵法》、《吴子》、《孝经》的篇幅都不足一万汉字，使得 25 部文献的通用汉字数量过少。但是反过来说，篇幅小的四部书仍有这 132 个汉字，说明了这些汉字的重要程度。

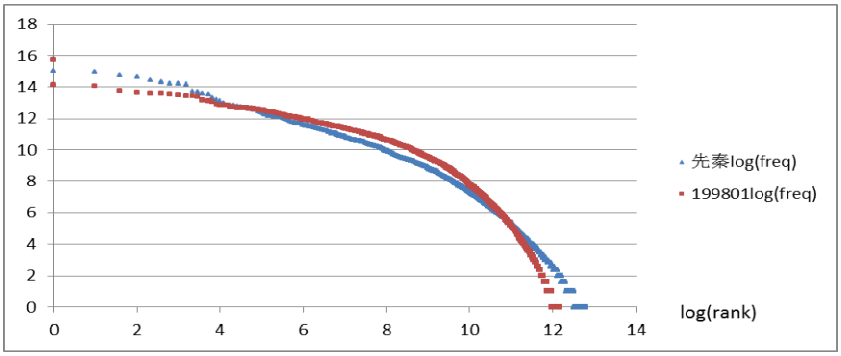
3. 2 先秦文献的汉字熵

汉字的熵，一直是汉语信息处理的基础问题之一。熵在信息论中的意义是指，一个变量的信息量的多少，或不确定性的大小。不确定性越高，则熵越大。冯志伟（1984）测定现代古代汉语的熵值为 9.65。那么先秦汉语的汉字熵，在给定的语料库上，也是比较容易测定的。熵（H）的计算公式如下，

$$H(X) = -\sum_{x \in X} P(x) \log P(x)$$

x 表示任意一个汉字，P(x)表示该汉字在给定语料库中出现的频次/语料库总汉字数，log 的底为 2。25 部文献得到的汉字熵为 9.227047203。作为对比，199801 汉字熵为 9.655485861。这个结果比较耐人寻味。我们的预想是，从概率分布上看，先秦文献和 199801 都应符合 Zipf 定律（Zipf 1935），两者的差异应较小。但是字型数量上先秦文献是 7049 个，大大多于 199801 的 4574 个。因此，熵值应该比 199801 略高才对。而实际测得的值却比 199801 略低，说明先秦文献在汉字频率分布上差异大。

图 1：先秦语料库和 199801 的齐夫定律



和 199801 做对比，我们绘制了两个语料库上齐夫定律的图示。汉字的频率（freq）及其排位（rank），都取了 log（底为 2）。曲线上的差异，也印证了先秦汉语和现代汉语的差别。

4 先秦文献的词汇统计

4.1 词频和分布

我们按词性统计了 25 部文献的词频分布情况，最高频的词仍为助词“之”。25 部文献中均出现的词语共计 89 个，比共同出现的汉字 132 个少了 43 个。这 89 个词，除“天下/n”外，全部为单字词：

之/u、不/d、也/y、而/c、之/r、其/r、以/p、曰/v、者/r、有/v、則/c、為/v、於/p、人/n、可/v、無/v、矣/y、所/r、君/n、民/n、是/r、國/n、以/c、必/d、乎/y、能/v、得/v、謂/v、知/v、一/m、如/v、三/m、行/v、事/n、大/a、道/n、天下/n、子/n、用/v、皆/d、見/v、至/v、天/n、從/v、言/v、無/d、聞/v、立/v、未/d、在/v、死/v、受/v、治/v、士/n、四/m、五/m、日/n、生/v、心/n、取/v、地/n、成/v、雖/c、及/v、來/v、相/d、命/n、又/d、失/v、時/n、亡/v、食/v、舉/v、行/n、後/d、家/n、進/v、居/v、致/v、服/v、退/v、興/v、過/v、觀/v、加/v、親/v、死/n、和/v、陳/v

动词的词型数量最多，名词和副词占了较大的比重。如果按文献出现的个数加以扩展观察，则同时出现在 20 部文献的词语为 654 个，同时出现在 10 部文献中的词语为 28780 个。可见文献的同质性是比较低的。

4.2 词长统计

下面我们来关心一个传统话题，即先秦汉语中单音节词（单字词）和多音节词（多字词）的比例和分布情况。按照一般的观点，认为古汉语特别是先秦汉语中单音节词为主，即单音节词占了绝对的优势。不过这种看法，没有讲清楚词型和词例的关系。现在基于分词和标注词性的语料库，我们可以通过频率统计来看看单音节词和多音节词的分布情况到底如何，和现代汉语 199801 语料相比，有哪些共性和差异。

表 1：25 部先秦文献和 199801 语料词长统计

词长	25 部先秦文献		199801 语料	
	词型数	词例数	词型数	词例数
1	15867	1074180	4547	354253
2	24929	104299	35494	497496
3	2879	8305	10687	48672
4	713	1373	5682	20401
5	19	26	729	2463
>5	0	0	583	926
平均	1.74	1.11	2.39	1.73

统计表明：

①**先秦汉语中的多字词，特别是二字词在词型数量上超过单字词**，只是在使用的词例上远远小于单字词。这表明，先秦汉语中的多字词已占了相当比例。多字词去掉数词（m）、人名（nr）、地名（ns）、专名（nx）和时间词（t）后，词长缩短至 2~4 个汉字，依然剩余 17503 个词型，占总词型的一半以上。如果再保守一些估算，频率在 2 次以上的词型为 16049 个，其中 2 字以上的词型为 6282，占到 39.15%。可见多字词的出现频率大部分为 1，在频率上单字词的出现比较占优。

②**仅以词型计算平均词长的话，先秦汉语接近 2，单字词并不占优，仅为 35.8%**。但是从使用频次上看，单字词占了绝大部分用例（90.4%），动态平均词长为 1.1 字。单字词的优势是相当明显的。

现代汉语 199801 语料的平均词型长度为 2.39，略大于 2。长度大于 5 个字的多字词型也有不少。而从词例上看，多字词的比例占了绝对优势，词例的平均词长达到了 1.73 个字。这些都体现出多字词特别是二字词在现代汉语中的优势地位。因此，先秦汉语单字词占优、现代汉语多字词占优的传统说法大体上还是正确的，只要在先秦汉语方面需要指出词型和词例的差别即可。

4.3 词类统计

下面我们所要关注的是词类问题。在古汉语研究中，词类并不是一个研究重点。一般认为古

汉语中实词兼类是较为常见的。我们在制订分词和词性标注规范的时候是无法回避词类问题的，为了考察词语兼类的情况，没有采用朱德熙先生（1983）的“汉语词类多功能说”，而是使用了黎锦熙先生（1924）的“依句辨品”法，直接标注词语在上下文中的词类。如，名词“雨”做动词时，直接标注为“雨/v”。对于使动、意动、为动、名作状、形作状、动作状，分别使用 vs、vy、vw、zn、za、zv 等六个词类标记。

表 2：25 部文献词类频率

标记	名称	词型数	比例(%)	词例数	比例(%)	标记	名称	词型数	比例(%)	词例数	比例(%)
a	形容词	3226	7.6	39293	3.3	**r**	**代词**	**288**	**0.7**	**104505**	**8.7**
c	连词	223	0.5	64887	5.4	s	拟声词	95	0.2	134	0.0
d	副词	1403	3.3	86768	7.2	t	时间词	396	0.9	9655	0.8
f	方位词	108	0.3	11080	0.9	u	助词	104	0.2	37120	3.1
i	词缀	25	0.1	425	0.0	**v**	**动词**	**7324**	**17.3**	**346036**	**28.7**
j	兼语词	9	0.0	1522	0.1	vs	使动	348	0.8	1016	0.1
m	数词	379	0.9	23114	1.9	vw	为动	32	0.1	89	0.0
n	**名词**	**17228**	**40.7**	**312898**	**26.0**	vy	意动	122	0.3	555	0.0
nr	**人名**	**7320**	**17.3**	**41860**	**3.5**	x	非语素字	13	0.0	19	0.0
ns	地名	2418	5.7	21145	1.8	y	语气词	163	0.4	52214	4.3
nx	其他专名	481	1.1	1410	0.1	za	形作状	71	0.2	162	0.0
p	介词	111	0.3	45158	3.7	zn	名作状	215	0.5	1298	0.1
q	量词	160	0.4	2156	0.2	zv	动作状	64	0.2	313	0.0
--	--	--	--	--	--	**合计**		42326	100	1204832	100

表 3：25 部先秦文献词语兼类统计

兼类数	词型数	带词性词型数	词例数	例词
11	2	22	4299	然
10	2	20	5652	若
9	5	45	29600	上
8	17	136	112892	如
7	43	301	72514	小
6	72	432	137173	止
5	175	875	163248	故
4	301	1204	171679	夜
3	634	1902	182650	罪
2	1479	2958	195486	雷
合计	2730	7895	1075193	

表 4：25 部先秦文献最高频的十个兼类词

词型	兼类数	词例数	兼有的词类（按词频降序排列）
然	11	2797	r、c、v、i、a、n、d、u、x、y、nr
重	11	1502	a、v、n、d、q、nr、vy、vs、ns、za、m
若	10	3755	v、c、p、r、d、nr、n、i、y、a
後	10	1897	d、n、f、v、a、t、c、p、vs、zn
上	9	3960	f、n、v、a、zn、d、m、vs、r
是	9	5559	r、v、u、n、a、d、c、i、p
為	9	12920	v、p、n、c、u、a、r、y、d
于	9	6694	p、u、v、i、y、n、c、d、a
厥	9	467	r、i、u、c、v、d、n、nr、y
如	8	4057	v、p、c、i、y、r、d、n

表 2 给出了这 25 部先秦文献基于 25 个词类标记（去掉标点 w）的词频统计情况，按照词类标记字母序排列。①从词型上看，名词 n、动词 v 和人名 nr 是数量最多的词类。在人名、地名和其他专名中，也是人名的比例最高。介词、语气词特别是量词超过了 100 个，说明先秦汉语这三个词类已经发展得比较成熟了。②从词例上看，动词 v、名词 n、代词 r 的数量最大。这说明动词和名词是先秦语料中最常用的词类，动词词型相对偏少但使用最多。人名 nr 的型虽多，但出现的总比例并不是很高。相反地，少量的代词却占据了文本词例的 8.7%。量词和介词、语气词的使用

频次差距较大，说明量词虽然出现，但用例还相对较少。拟声词的出现则更具有偶发性。兼语词j（如“诸”）词型只有9个，词例多达1522个。使动用法等六种词类活用的型和例都比较低。

虽然语料库经过多次校对，但为了减少人工标注的少量错误带来的影响，我们将每个类别出现3次及以上的词语才算作严格的兼类。表3给出了25部先秦文献的词语严格兼类统计信息。兼类的词型达到了7895个，词例1075193个，分别占总词型数和词例数的17.6%和89.2%。这一点印证了兼类词的一个特点，即兼类词往往都是高频词。兼类词的数量也随着兼有词类的减少而递增。兼两类的词数量最大。表4进一步给出了最高频的十个兼类词的兼类情况。这些词语最高频出现的词类都是他们典型的词类。

5 总结与未来工作

基于25种先秦文献的标注语料库，我们统计了字词的频率分布情况，给出了最高频和分布最广的字词，测定了汉字熵，对先秦汉语的多字词也进行了统计分析。通过词型和词例的统计，我们看到先秦语料库的字汇和词汇有如下特点：（1）各文献用字和用词差异大，同时出现在25部文献中的汉字和词分别仅为132个和89个；（2）字频和词频最高的都为“之”，但并非每本书最高频的汉字均为“之”；（3）先秦语料库汉字分布也符合齐夫定律，汉字熵为9.227，略低于现代汉语语料；（4）多音节词在词型上占优势，单音节词在词例上占优势；（5）词频最高的三大词类分别是名词、动词和人名，但在动态使用中，动词、名词和代词数量最多；（6）具有汉语特色的语气词、量词已得到发展，语气词使用量大而量词较少；（7）词类活用的情况并不太多，词类兼类的情况则非常普遍，兼类词的词例数占到了全部词例的89.2%。

这七个特点，初步勾勒了先秦汉语的字词概貌。下面我们将进一步进行研究，一是进一步挖掘先秦词汇的深层次特点，服务于汉语史数据库的构建；二是对现有语料继续查错校对，核实多字词的标注精度，早日发布给学界使用；三是增补文献，对先秦时代较可靠的其他文献进行标注；最后是做更深入的全文义项标注，为汉语词汇史的研究提供更重要的资源。

6 致谢

感谢参与先秦语料库建设的南京师范大学文学院几十位本科生和研究生的辛勤劳动，感谢匿名审稿专家的修改意见！

参 考 文 献

[1] Zipf, G. K., The Psycho-Biology of Language[M], Houghton Mifflin, Boston, 1935.
[2] 车淑娅. 《韩非子》词汇研究[M]. 成都：巴蜀书社, 2008.
[3] 陈克炯. 春秋左传详解词典[M]. 河南:中州古籍出版社, 2004.
[4] 陈小荷. 先秦文献的信息处理[A]. 中国中文信息学会成立二十七周年学术会议[C], 2008.
[5] 冯志伟. 汉字的熵[J]. 文字改革,1984(4):12-17.
[6] 黎锦熙. 新著国语文法[M]. 1924 初版, 北京：商务印书馆, 2001（重印）.
[7] 潘允中. 汉语基本词汇的形成及其发展[J]. 中山大学学报,1959,(1,2):98-121.
[8] 石民,李斌,陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报, 2010,2(24):39-45.
[9] 叶正渤. 上古汉语词汇研究[M]. 北京：中央文献出版社, 2007.
[10] 张双棣. 吕氏春秋词汇研究[M]. 济南：山东教育出版社, 1989.
[11] 周荐. 基本词汇与一般词汇划分刍议[J]. 南开学报,1987,(3).
[12] 朱德熙. 语法讲义[M]. 北京：商务印书馆, 1983.