

# 先秦文献的信息处理刍议

陈小荷，2008年11月23日

我们正在做一个项目“先秦汉语词汇统计与知识检索”，准备对25种最重要的先秦传世文献<sup>1</sup>进行词语切分、词性标注、个别常用词（包括古今字和通假字）的词义标注，建立先秦书面汉语（以下简称先秦汉语）的词汇知识库以及先秦文献的历史知识库并研制相应的检索系统。有关文献选择、版本选择、校勘等方面的问题，我们会请教古文献方面的专家学者。今天想主要就先秦汉语的词汇计算和内容计算两个方面的问题向在座专家汇报我们的初步想法并希望得到指教。

目前先秦文献的信息处理大体还处于字处理阶段，以解决古文字的输入输出、文献逐字索引等问题为主要内容<sup>2</sup>。先秦文献的词处理只有一些尝试性的实践。山西的几位古文献专家对《左传》做了词语切分。他们在《左传》约25万字语料中，切出了7069个多字词条<sup>3</sup>，在总词频中约占12%。《左传》总共3282个不同汉字，即使以平均每个汉字表示3个单字词条来粗略估算，词表中多字词条仍占有相当大的比例。由此可见，虽然先秦汉语以单字词为主，但多字词也是不可忽视的。

先秦汉语词汇研究具有悠久的历史，传统的语文学以经、传为核心，在大量的注、疏中包含了宝贵的词汇研究材料，并逐步形成了“训诂”这样一门专门的学问。现代语言学继承了这一传统，在古代汉语个别词的词义考释方面取得了丰富的研究成果。但是，还缺乏对汉语词汇演变的大规模调查，缺乏对汉语词汇发展脉络的宏观把握。从古代汉语到近代汉语再到现代汉语，词汇系统在社会发展和民族融合过程中已经发生了巨大变化。古代汉语的许多词语已经消失，现代汉语的许多词语来源不明（例如“这”和“那”），一些词语虽然还在使用，但词义已经迥异或微殊。当然，汉语基本词汇的传承性是不容怀疑的，其中的根词仍然是现代汉语复合构词的基本材料。我们想运用计算语言学方法对先秦汉语词汇做

---

<sup>1</sup> 历史7种：尚书、左传、公羊传、谷梁传、国语、晏子春秋、吕氏春秋；韵文2种：诗经、楚辞；诸子11种：管子、论语、孟子、老子、墨子、庄子、孙子、吴子、荀子、商君书、韩非子；其他5种：仪礼、礼记、周礼、孝经、周易。

<sup>2</sup> 国外对中国古籍索引研究最多的当属日本，据不完全统计，日本出版的中国古籍索引约占世界各地所出版的全部中国古籍索引的80%。我国台湾中研院1984年开始着手开发“瀚典全文检索系统”，他们的古籍数字化成果现在基本都能提供网上检索服务。我国香港从事古籍数字化工程的中坚力量是香港中文大学中国文化研究所下属的“汉达古文献数据库中心”。目前，香港商务印书馆正在陆续出版的《先秦两汉古籍逐字索引丛刊》，由香港中文大学中国文化研究所采用电脑整理先秦两汉文献而编成，共输入古籍102部，计800万字。该索引可以详细展示某部古籍中所用单字的使用频率以及在句子中出现的具体情况，甚或某字在古籍中的用例、出处等等。我国大陆地区有湘潭大学研制了古籍索引自动编辑系统，可自动编制古籍的逐字索引、句子索引、人名索引、地名索引及其他专题索引。

<sup>3</sup> 多字词条出现26353词次、56356字次。

近乎穷尽式的断代研究，目的在于追本溯源，然后顺流而下<sup>4</sup>，理清汉语词汇发展的基本脉络。具体来说，我们想搞清楚以下一些问题：

- 先秦汉语有多少个不同的词，分别表示哪些概念？这需要在词语切分的基础上做一个词表，将同形词区分开来释义。在词的认定上可不可以找到某种起辅助作用的统计标准？
- 先秦汉语中单字词和多字词的使用频率是怎样的，哪些多字组合有时候是词、有时候又不是词，如何鉴别？现代汉语自动分词中组合型歧义消解的方法有借鉴作用吗？
- 将先秦汉语跟现代汉语在词汇统计上进行宏观的比较，看看先秦汉语中哪些词还在用，哪些词已经不用或很少用了，重合的部分占多大比例（姑且不深究词义的区别）。

我们对先秦文献中容量最大的《左傳》（25万字次）做了一个词频统计，前10个最高频汉字词是：“之、也、曰、不、以、而、其、于、於、為”，相对频率3.55%~0.8%。而1998年上半年《人民日报》语料的词频统计结果，前10个最高频汉字词是：“的、在、了、和、是、一、有、不、为、对”，相对频率4.9%~0.38%；仅发现一个词“不”是重合的。由于从使用频率上说古今汉语都是单字词占绝对优势，因此我们还对先秦文献与1998年上半年《人民日报》在汉字频率上做了一个比较。前100个最高频汉字，重合的仅有29个（已经消除简体和繁体的差别）<sup>5</sup>。

- 从概念表达的角度来观察，先秦汉语跟现代汉语有哪些错综复杂的词汇对应关系？例如，表示run的意思，先秦用“走”，现在用“跑”；表示walk的意思，先秦用“行”，现在用“走”。如何用计算机自动发现这些对应关系？
- 从概念表达的角度来观察，从先秦汉语到现代汉语有哪些词汇替换现象？例如“此”和“彼”替换为“这”和“那”。如何用计算机自动发现这些词汇替换关系？
- 从概念表达的角度来观察，先秦汉语沿用至今的词有哪些词义变化？例如“布”原指麻布；“說”原意为说服、喜悦。从信息处理的眼光来看，我们当然要关注意义微殊，但更关注意义迥异。立足于词而不是立足于字，我们认为古今字和通假字其实就是古文献中的意义迥异的同形词。因此，我们打算对先秦文献中的古今字、通假字做一个全面的标注。如何标注？现代语言中的词义消歧方法能移植过去吗？

说某书甚至某个时期的文献中有多少种古今字、通假字比较容易，但说这些

<sup>4</sup> 南京师范大学文学院的董志翘教授正致力于中古汉语（特别是佛典文献）词汇的统计和研究。

<sup>5</sup> 这29个汉字是：不成出大公國民能人日上生事是我下行一以用有于與月在者中主自。

字在某个概念上的频率分别是多少，就非得经过全面标注才行。全面标注可能会带来一些新的认识。例如，据初步考察，先秦文献中的“說”大都表示说服或喜悦，几乎没有现在“说话”这个意思，现在“说”这个意思，先秦文献中都用“道”或“曰”来表达。训诂学家们不关心这种问题，他们主要关心某字在古文献中有哪些普通人所不了解的意思。

除了词汇计算之外，我们还想在先秦文献的信息处理中做一些基于内容的计算。我们准备根据先秦文献建造历史事件、人名、地名三个知识库，对事件、人物和地点进行相关性度量，解决人名、地名的“同名异指”和“异名同指”问题，进行事件、人名和地名等历史知识的交叉检索。

左传、公羊传和谷梁传都是对春秋时期的历史做详略不一或者侧重点不同的阐述，是研究那个时期的历史事件的很好的材料<sup>6</sup>。现代的信息抽取（或事件抽取）技术也许能派上用场。

先秦文献中人名、地名的标注不是主要问题。一则先秦传世文献数量有限，二则出版的文献中一般都对专名加了特殊标记。不过，目前可利用的电子文本通常没有专名标记。在先秦文献的信息处理中，研究人名、地名的自动标注技术，着力点恐怕主要在于解决地名沿革、人名的同名异指和异名同指问题。换言之，我们要挖掘人名和地名的知识背景，而不仅仅是给文本中的词加上人名、地名标记而已。

人名库方面，据我们对《左传》已经标注的12万词次统计，人名出现了7627次，约占4.6%。先秦时期，姓、氏、字、号的使用都有特殊的文化因素，但为了解古代文化，首先需要解决人名的同名异指和异名同指问题，否则容易造成查全率和查准率的下降。

同名异指问题，例如“夫人姜氏”既可以指“定姜”、“归姜”、“生姜”、“哀姜”、“文姜”，还可以指“晋穆侯夫人姜氏”。

异名同指问题，例如郑伯克段于鄆故事中的“段”，其他称谓有“大叔段”、“大叔”、“共叔”、“共叔段”、“京城大叔”或者“郑共叔”。

地名库方面，我们收集了谭其骧先生主编的《中国历史地图集》先秦部分地名2807个。另外又收集了《左传》文本中的地名1078个，其中699个不见于历史地图集。这表明，先秦文献中大部分较小的地名不太可能出现在历史地图上。地名也有异名同地、异地同名的问题需要解决。

春秋战国是中国历史上第一个文化繁荣的时期。我们想通过历史事件、人物、地点的交叉检索来帮助大众更形象直观地了解那个时期的历史和文化现象，例如

---

<sup>6</sup> 三传还是研究词义的好材料，因为《春秋》微言大义，三传则致力于发掘词的语用意义。

历史事件的追踪、同一历史事件在不同文献中的表述、人物的历史参与和地理轨迹，用基于内容的计算技术更好地为大众服务。这样的构想有意义吗？是可行的吗？有没有更好的构想和切实可行的实现方法？请专家们不吝赐教。

参考文献：

常娥，侯汉清，曹玲 2007：古籍自动校勘的研究和实现. 中文信息学报，第2期.

陈琦潘 2003：武汉图书馆馆藏古籍善本数据库的建设与知识型数据库的实现. 图书馆论丛，第4期.

陈阳. 中文古籍数字化的成果与存在问题[J]出版科学. 2003. (04) .

李铎，王毅 2005：关于古代文献信息化工程与古典文学之间互动关系的对话. 文学遗产，第1期.

李国新. 中国古籍资源数字化的进展与任务[J]大学图书馆学报. 2002. (01) .

尉迟治平 2004：汉语信息处理与计算机辅助汉语史研究. 语言研究, 第3期.

伍宗文 2001：先秦复音词研究. 巴蜀书社.

于亭 2000：计算机与古籍整理研究手段现代化. 古汉语研究, 第3期.

张普 1989：中国古籍语料库的建立与标准化. 中文信息, 第2期.

张普、石定果 1989：计算机与古籍整理研究. 语文研究, 第4期.

郑永晓 2005：古籍数字化与古典文学研究的未来. 文学遗产, 第5期.