上古汉语分词和词性标注语料库的构建

柯永红
北京师范大学民俗典籍文字研究中心, 中国文字整理与规范研究中心　北京市新街口外大街19号100875
yh5555@126.com

摘要

针对国内尚无开放的大规模上古汉语分词及词性标注语料库可用的问题，提出以人工为主+ 机器辅助的标注模式，构建一个包括46部文献的上古汉语分词及词性标记语料库。描述了语料选择、文本分词、词性标注和质量控制等建库过程，分析了该语料库词长、词频、词用等分布，评估了标注质量。已经完成标注的语料库包括323余万字、217万余词。与EvaHan2022基测集和盲测集的分词、词性标注一致度分别为93.70%、89.49%和92.83%、89.86%。该语料库可用于古汉语研究、辞书编撰、语言教学、人工智能等多个领域。

关键词： 上古汉语 ； 分词 ； 词性标注 ； 语料库

Construction of Ancient Chinese Word Segmentation and Part-Of-Speech Corpus

Yonghong Ke
Research Center for Folklore, Classics and Chinese Characters,
Research Center for Collation and Standardization of Chinese Characters
No.19, Xinjiekouwai St, Haidian District, Beijing, 100875, P.R.China
yh5555@126.com

Abstract

In order to solve the problem that there is no open and large-scaleannotated ancient Chinese word segmentation and part-of-speech corpus available in China, a manual and machine assisted tagging model was proposed to construct an Ancient Chinese word segmentation and part-of-speech corpus including 46 documents. The construction process of corpus selection, text segmentation, part-of-speech tagging and quality control are described. The distribution of word length, word frequency, word usage and tagging quality of the corpus are analyzed. The corpus that has been annotated includes more than 3.23 million characters and 2.17 million words. The agreement of word segmentation and part-of-speech tagging with EvaHan2022 base set and blind test set is 93.70%, 89.49% and 92.83%, 89.86%, respectively. The modified corpus can be used in many fields such as ancient Chinese research, dictionary compilation, language teaching, artificial intelligence and so on.

Keywords: Ancient Chinese , word segmentation , part-of-speech , corpus

1 引言

本文为国家社科基金课题“上古汉语词标记语料库及应用系统构建研究”（项目编号：20BYY127）的阶段性成果。

国内尚无开放使用的大规模古籍分词和词性标记语料库，给古籍研究和保护带来极大不便，也限制了人工智能技术在古籍文献上的深层应用。比如，在古籍全文检索语料库输入词条“天然”、“引领”，会检出错误的句子：

例1. 有**天然**后有地，有地然后有别。《鬻子·汤政汤治天下理第七》
例2. 如有不嗜杀人者，则天下之民皆**引领**而望之矣。《孟子·梁惠王上》

例1中“天然”不是一个词。例2中“引领”表示“伸长脖子”，是词组而不是词，而这样的错误需要具备一定的古文阅读素养才能发现，极易对使用者产生误导。

本文的目标是建设一个覆盖上古时期主要文献的分词和词性标记语料库。古籍所对应的古代汉语的面貌与现代汉语有很大不同，其内部各阶段也有较大的分别 (王力, 1980)，本文讨论的是上古汉语分词和词标注语料库的建设。

2 研究现状

2.1 古汉语分词和词性标记语料库建设现状

我国机读语料库的建设始于1979年。在各界的共同努力下，古籍资源库和语料库的建设取得了丰硕的成果，极大方便了古汉语研究和应用。从整体上看，我国的古汉语语料库建设有三个方面的特点：（1）目前建成的古汉语语料库多是没有分词和词性标记的电子文本语料库，如中华古籍资源库(中国国家图书馆)、中华经典资源库(国家语言文字工作委员会)、中国基本古籍库(北京爱如生数字化技术研究中心)、瀚堂典藏古籍数据库(北京时代瀚堂科技有限公司)、书同文古籍数据库(北京书同文数字化技术有限公司)、国学宝典(北京国学时代文化传播股份有限公司)、籍合网(古联（北京）数字传媒科技有限公司)、国学大师(国学大师)、识典古籍(北京大学-字节跳动数字人文开放实验室) 北京大学CCL语料库(北京大学中国语言学研究中心)、鼎秀古籍全文检索平台(北京翰海博雅科技有限公司)、中国哲学书电子化计划(STURGEON D.)等，这些语料库解决的是古籍文字逐字索引问题。（2）古汉分词和词性标记语料库数量极少，如南京师范大学先秦汉语标注语料库(陈小荷等, 2013)、南京农业大学典籍平行语料库(黄水清, 王东波, 2021)等，尚未完全开放访问。（3）台湾“中研院”古代汉语标记语料库（以下简称ASC）和英国谢菲尔德大学历时汉语语料库（SCC）是世界范围内具备一定规模的、开放的分词和词性标记语料库。ASC的文献数量较多、标注质量高且提供了较为丰富的词检索功能。但是，ASC采用的术语和标签集与大陆存在差异，且有访问限制。SCC的所收语料较少、标签通用性较低且存在一些明显的断句、分词或标注错误。

整体而言，国内尚无完全开放的大规模古汉语分词和词性标注语料库。

2.2 古汉语自动分词和词性标注技术研究现状

古汉语自动分词主要有基于词典和统计的方法、基于序列标注的方法和基于深度神经网络的方法。基于深度神经网络的分词和词性标注方法相对于传统方法有明显的优势，俞敬松等 (2020)、Tang Xuemei等 (2022)在部分语料上分词的F值分别为95.32%、98.46%。第一届古代汉语分词和词性标注国际评测中，盲测集上封闭测试分词和词性标注的F值分别达到93.64%和87.77% (李斌, 袁义国, 芦靖雅等, 2023)。总体而言，预训练模型、外部知识引入以及多模型融合方法大大提升了古汉语自动分词的效果。

但是，当下古汉语的自动分词仍然不能满足严谨的科研场景的需求。汉语的大部分复音词是从短语演变而来的，在古汉语阶段，有的复音词还处于由短语向词的转变过程之中，这些复音词身上还带着短语的特征，这给词和短语的区分带来很大困难。北京大学数字人文研究中心“吾与点”是目前最新、最先进的古汉语分词系统之一。以“吾与点”2024年1月的版本为例，在其中输入三组例句，可以得到如下分词结果：

例1. 佑/问/长/：/有/**妻子**/乎/？/对/曰：/有/妻/未/有/子/也/。《后汉书·吴佑传》
例2. **妻子**/好合/，/如/鼓/瑟琴/。《诗经·小雅·常棣》
例3. 是故/君子/有/终身/之/忧/，/无/**一/朝**/之/患/也/。《孟子·离娄下》
例4. 强/而后/可/，/ **一/朝**/而/获/十/禽/。《孟子·滕文公下》
例5. 残贼/之/人/谓/之/**一/夫**/。《孟子·梁惠王下》
例6. 耕者/之/所/获/，/**一/夫**/百亩/。《孟子·万章下》

上述分词结果中例4、例6、例8存在错误的分词。例4中的“妻子”表示“配偶和子女”，是词组。例6中“一朝”表“一时”义，是词。例8中的“一夫”表示“众叛亲离之人、独夫”，是词。

我们今天所能见到的古汉语语料是封闭的、有限的、不完整的，一些结构在短语向词转变的过程中存在词和短语两种形态，或者作为词出现的频次很低（表现为未登录词或低频词），导致可用来机器学习的语料不足，这也是限制机器自动分词准确性的主要原因，而这样的问题在可预见的一段时间内仍将存在。古汉语中存在许多兼类、词类活用以及词的临时用法，给词性的标注带来了很多困难。而且，自动分词的错误会叠加词性标注之上。

目前，上古汉语缺乏相应的词性标记规范。《信息处理用现代汉语词类标记规范》(标准号GB/T 20532—2006) 及其修订工作(杨丽姣等, 2021)为现代汉语汉语语料库建设提供了重要依据。化振红(2021)的《建立中古汉语语料库分词规范的若干问题》讨论了中古汉语词的切分及分词规范问题，为上古汉语分词及词性标注提供了重要的参考。

综上所述，在现有技术条件下，古汉语分词和词性标注仍需要大量的专业人工标注和检验工作，这也是建设大规模、高质量古汉语分词和词性标记语料库的主要困难。

3 上古汉语分词和词性标记语料库的建设

我们的语料库总体遵循了一套系统化、规范化的建库流程，总体思路可以用下图表示：

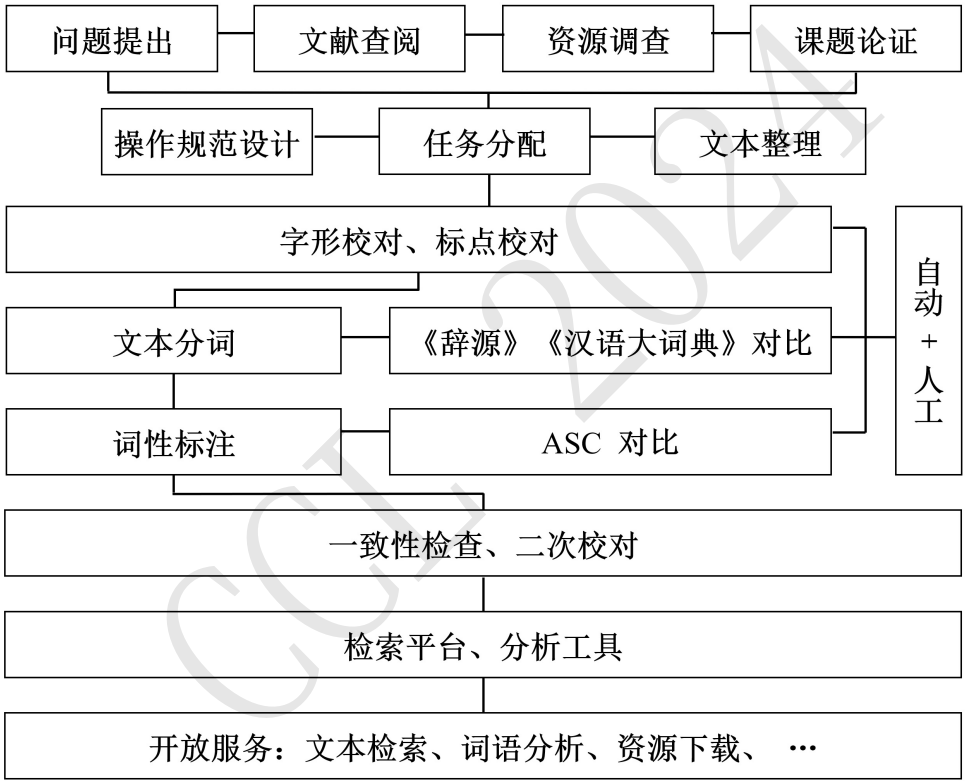

Figure 1: 建库过程

首先，我们通过广泛的文献查阅和资源调查完成课题论证。在确定了研究内容和建库目标后，设计了操作规范，进行任务分配和文本整理，为后续工作做好铺垫。在语料标注阶段，进行了字形和标点校对、文本分词、词性标注等一系列步骤，同时结合自动校对与人工检查的方式进行了比对，以确保数据的准确性和完整性。之后，通过一致性检查和二次校进一步提升语料标注质量。在完成语料标注的基础上，通过检索平台和分析工具，为用户提供包括文本检索、词语分析、资源下载等在内的开放服务。

如前文所述，在现有技术条件下，上古汉语分词和词性标注仍需要大量的专业人工标注和检验工作。为建成高质量的大规模古汉语分词和词性标注语料库，本文采用了自动工具为辅+人工为主的标注模式，主要基于两个方面的考虑：（1）在自动标注阶段，使用机器对专有名词、数词、量词、连词、代词等出现较多且自动标注准确率高的词进行预标注，再由人工

对剩余部分进行标注并对机器标注部分进行校对，可以减轻大规模语料标注中人工标注的工作量。（2）在人工标注阶段，充分借鉴已有研究成果，使用《辞源》《汉语大词典》以及ASC等权威辞书和高质量语料库协助人工标注和校对，进一步提升标注准确率。限于篇幅，我们重点介绍语料选取、文本分词、词性标注和质量控制几个方面的工作。

3.1 语料选取

上古汉语主要包括先秦和两汉时期。根据李零(2008)的统计，先秦两汉典籍现存115种，其中先秦60种、秦1种、汉54种。我们以文献均衡性和重要性为主要依据，选取了上古时期的46种文献。如表1所示：

文献类别		先秦	两汉
传世文献	六艺类（10种）	《左传》《公羊传》《谷梁传》《诗经》《尚书》《仪礼》《礼记》《大戴礼记》《周礼》《周易》	
	史书类（3种）	《国语》《战国策》	《史记》
	子书类（23种）	《论语》《孝经》《孟子》《荀子》《墨子》《老子》《庄子》《冠子》《慎子》《商君书》《韩非子》《邓析子》《吕氏春秋》《管子》《晏子》《孔子家语》《孔丛子》	《新语》《新书》《新序》《淮南子》《春秋繁露》《说苑》
	兵书类（5种）	《司马法》《六韬》《孙子》《吴子》《尉缭子》	
	方技类（2种）	《素问》《灵枢》	
	经传小学类（1种）		《韩诗外传》
出土文献	（2种）	马王堆帛书、睡虎地秦简	

Table 1: 语料库所包含的文献

上述文献已经包含了上古时期的主要文献，能够满足日常使用需求。随着研究工作的进展，我们还在增加新的文献，最终的目标是覆盖目前已有的全部文献。

3.2 文本分词

将句子切分成词串是建设分词和词性标记语料库核心工作。但是，古汉语特别是上古汉语的分词与现代汉语相比，情况要复杂得多。在古汉语中，有大量的复音词处于词和词组的中间状态。同样一个结构在现代汉语是词，在古汉语有时是词有时是词组。比如“大国”“大雨”和“大海”“大我”，“大国”和“大雨”意义为“大的国”“大的雨”，与之相应的还有“小国”“小雨”，应当是词组。“大”是“海”自身的属性和特点，没有“小海”之说，“大海”应当是词。“大我”尽管有对应的“小我”，但其中的“大”和“小”不和“我”直接发生关系，应当看作是词(王宁, 2008)。马真(1981)指出“确定先秦时期的合成词，根据的只是有限的书面材料，不能像区分现代汉语的词和词组那样可以采用‘替换法’‘扩展法’‘插入法’等，更不能简单地使用是否经常连用作为定词的标准”。具体操作而言，我们遵循几个主要的分词原则：

(1) 意义方面，根据语素组合后是否表达一个整体的意义来判定是否为复音词。构成词的两

个或多个语素组合后表达一个整体的意义，且不同于参构语素按照结构关系推导出的意义(卜师霞, 2018)。语素组合一般有产生新义、意义泛化或具体化、意义偏指或构成偏义等几种情况。

例1. **左右**曰："待先生，如此其忠且敬也。"《孟子·离娄下》
例2. 予死于**道路**乎？《论语·子罕》
例3. 古者**大事**，必乘其产。《左传·僖公十五年》
例4. 王曰："舍之！吾不忍其**觳觫**，若无罪而就死地。"《孟子·梁惠王上》
例5. 则易直子谅之心**油然**生矣。《礼记·乐记》

例1中"左右"作为一个整体构成了新的意义，例2中"道路"构成了一个更加概括的意义，例3中"大事"的意义具体化，特指战争。例4中"觳觫"表达一个不可拆分的、整体的意义，例5中"然"作为词缀不表义。因而，上述五个例子加粗部分都应当划分为词。

(2)语法方面，根据语素组合结构的稳固性判定是否为复音词。构成复音词的两个或多个语素组合应结合稳定，不能拆开或者随意扩展，否则是词组。或者组合结构的多个成分中，有一个一般不单用。

例6. 退食自公，**委蛇委蛇**。《诗经·羔羊》
例7. 昔我往矣，杨柳**依依**。《诗经·采薇》
例8. 子曰："父母之年，不可不知也。"《论语·里仁》
例9. 鲍叔牙为人，刚愎而上悍。刚则犯民以暴，愎则不得民心。《论语·里仁》
例10. 他人有心，予**忖度**之。《孟子·梁惠王上》

例6和例7中"委蛇""依依"不能拆开，否则没有意义，因而是词。例8中"父母"可以扩展为"父与母"，是词组。例9中"刚"和"愎"在后面的句子中以单用，因而前句的"刚愎"是词组。例10中"忖"表"揣度"义时，一般都不单用，因而"忖度"应作为一个分词单位。

(3)根据复音词与其构词语素的词性变化判定是否为复音词。毛远明(2000)指出词义的变化和词性的转变是同步的，词性转变了，应该是复音词。

例11. 凡祀，**启蛰**而郊。《左传·桓公五年》
例12. 孔张，君之昆孙，子孔之后也，**执政**之嗣也。《左传·昭公十六年》

例11中"启"和"蛰"都是动词，"启蛰"变为一个名词作状语，词性发生了转变，因而可以判定为复音词。例12中"执政"本为动宾结构，组合后变为名词性成份，可以判定为复音词。

(4) 根据上古汉语单音词占据绝对优势的特点，对于尚未完全固定的语素组合，坚持从严的标准，能分则分。

例13. 王何必曰利？亦有**仁义**而已矣。《孟子·梁惠王上》
例14. **仁义**忠信，乐善不倦。《孟子·告子上》

尽管"仁"和"义"作为组合出现的频次很高，但"仁"和"义"均可单独使用，且组合后意义也未发生改变，因而组合尚未完全固定，应切分为两个单位。

(5)同一语素组合有时是词，有时是短语，需要根据具体语境判定是否属于一个分词单位。

例15. 苟行王政，**四海**之内皆举首而望之。《孟子·滕文公下》
例16. 禹之治水，水之道也。是故禹以**四海**为壑。《孟子·告子下》

上面两个句子中，前一个"四海"表"天下"之义，是一个词，后面的"四海"表"四方之海"之义，是词组。

(6) 借助权威工具书和已有电子资源检验分词结果。我以《辞源》和《汉语大词典》作为复音词的检验参照、以ASC校对词性标注，开发了辅助工具在标注过程中为标注者提供提示性信息。上述工具书和电子资源是提示性的，不作为标准。

3.3 词性标注

王力(1990)指出"在划分词类的时候，不但要重视结构方面（形态方面），而且要重视意义方面。应该把结构和意义看成一个有机的整体。"划分词类时，我们遵循是功能和意义结合的标准。和现代汉语相比，古汉语的词类划分更加复杂，具体而言，需要重点关注兼类、词类活用和词的临时用法：

（1）古汉语中词的兼类非常普遍，需要在语境中依据意义和功能进行判断。以"使"为例：

例17. **使**勿使能殖，则善者信矣。《左传·隐公六年》
例18. 公怒，绝宋**使**。《左传·隐公九年》
例19. **使**能，国之利也。《左传·文公六年》

上面三个“使”分别应当分别标注为动词、名词和副词。

（2）古汉语中的词类活用也远多于现代汉语，常有名词作动词、形容词作动词、数词作动词等词类活用现象出现。比如“东”本来是名词，在：

例20. 秦师遂**东**。《左传·僖公三十年》

中“东”活用作为动词使用。在词性标注过程中，需要根据具体语境判断，这要求标注者有很好的古代文献阅读素养。

（3）此外，古汉语中还有一些词的临时用法，应当参照语境义划分词类。如：

例21. 天地车**轮**。《吕氏春秋·大乐》

高诱注：“轮，转。”据殷国光(2008)的统计，“轮”在春秋战国文献中出现的34例用法均为名词，义为“车轮”。殷国光认为此例中“轮”作“转”是一种临时意义，应以“常功能为标准”，将其划分为名词。我们认为殷国光关于词义的分析是正确的，但是在归类时，我们以“语境义为标准”，将其归入动词，这么做一方面符合该词的实际用法，同时也可以在语料库中保留更多的语言现象，有利于语言研究。

上古汉语缺乏相应的词性标记规范，在进行语料标注时，研究者通常会建立一套自用的标注集，但总体而言，关于词的大类划分的区别不大。杨伯峻、何乐士(2001)在《古汉语语法及其发展》中提出了一套全面、细致的词类划分方案，我们以两位先生的研究为主要参照，并考虑与《信息处理用现代汉语词类标记规范》(标准号GB/T 20532—2006, 标准号GB/T 20532—2006)尽量兼容，将词类划分十三个大类。词类和对应的标记如下表所示：

序号	标记	词类	词例
1	a	形容词	大
2	c	连词	而
3	d	副词	渐
4	e	叹词	呜呼
5	j	兼词	诸
6	m	数词	一
7	n	名词	东
8	o	拟声词	关关
9	p	介词	以
10	q	量词	杯
11	r	代词	吾
12	u	助词	者
13	v	动词	学

Table 2: 词类和标记

3.4 质量控制

由于自动分词还不能很好地区分上古汉语的词与非词，我们采用了人工为主+自动工具为辅的标注模式。根据李斌等(2012)对25部先秦文献的统计，名词、动词和人名是数量最多的词类。专有名词、数词、量词等数量多且自动标注准确率高，我们使用自动标注工具对这几类词进行预标注，然后由教师、研究生、本科生组成的标注小组标注剩余部分并对预标注内容进行检查。小组标注完成后再开展二次人工校对，校对工程中发现的难点、疑点问题由项目团队专家团队通过会议评审解决。过程可由下图表示：

标注人员的前期培训、分词与词性操作手册、古籍注疏参考资料都是必要的质量控制过程和材料。我们也开发了一些辅助工具来减少人工标注的错误，如标注标签自动校对和自动提示等。下图展示了利用外部资源自动校对的功能：

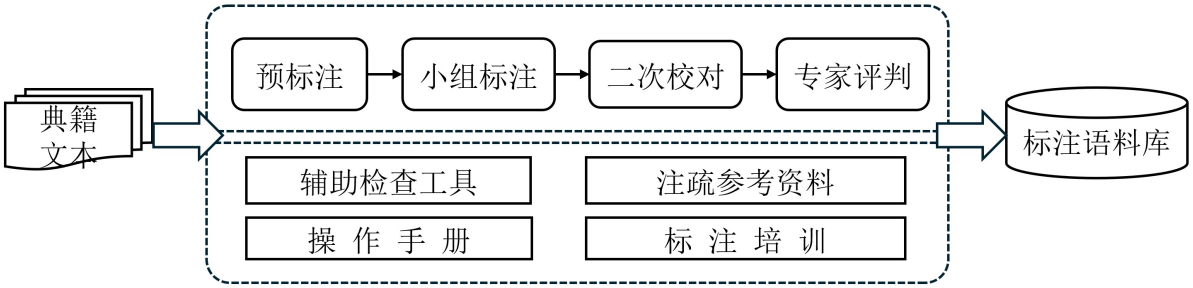

Figure 2: 建库过程

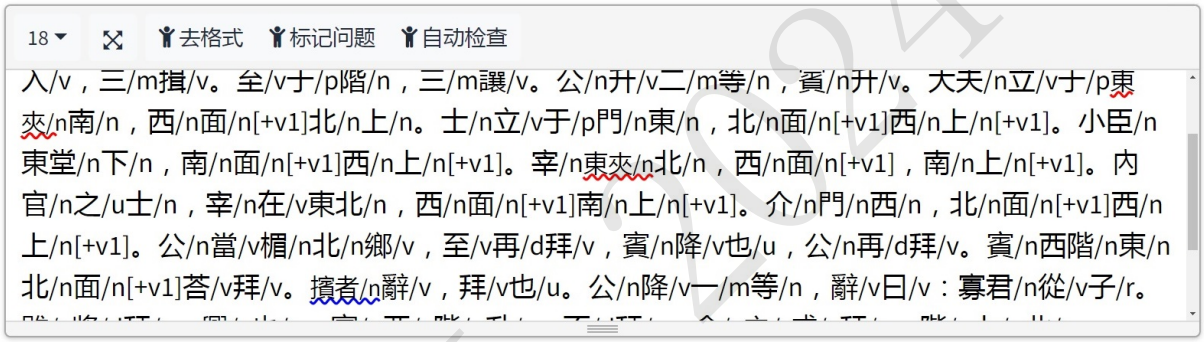

Figure 3: 辅助标注工具

上图中红色波浪线表示与ASC的分词不一样，蓝色波浪线表示《汉语大词典》中未收录此复音词，提示标注者注意检查。随着完成语料的增加，我们也将逐步完善辅助标注工具，让人工标注+ 智能辅助的方式取得更高的质量和效率。

4 上古汉语分词和词性标记语料库内容、质量及应用

4.1 语料库的内容

自2020年课题启动以来，我们已经完成46部典籍的标注工作，其中17部已经完成二次校对。根据统计，语料库的词长如下表所示：

词长	词数	百分比%	词形数	使用率
1	1907750	87.69	6925	275.49
2	237473	10.92	37506	6.33
3	22758	1.05	6265	3.63
4	5670	0.26	2419	2.34
5	1473	0.07	872	1.69
大于5	372	0.02	331	1.12

Table 3: 语料库的词长分布

王宁(1996)指出，汉语词汇的积累经历了原生阶段、派生阶段到以双音合成词为主的合成阶段。从上表的统计来看，呈现了先秦及两汉阶段汉语词汇的特点：从词的数量来说，先秦文言文中的单音词占据绝对优势，明显多于复音词；从词的使用频率来说，单音词的使用频率明显高于复音词。我们发现，三字词四字词多是专名（如周幽王、楚隐王、古公甫、强弩都尉、二十八宿、骇鸡之犀）、数字（如七十一、十分之二）以及一些特殊结构（如隐隐兮、焉、、暖暖姝姝），词长为5及以上的基本都是数字（如三千七十一、二百五十八、四百六十余）。

我们将语料库中的词频按照词性排序，可以得到下表：

序号	词性	数量
1	名词	733885
2	动词	627803
3	助词	186071
4	副词	160679
5	代词	145181
6	连词	106758
7	形容词	94693
8	介词	75649
9	数词	37919
10	量词	4778
11	兼词	1657
12	叹词	423
13	象声词	372

Table 4: 语料库的词性分布

可以看出，名词、动词、助词数量最多。上古汉语代词中副词、代词的占比高于现代汉语，副词在上古阶段仍处于虚化的过程中，代词一部分属于实词、一部分属于虚词。

4.2 语料库的标注质量

我们将该语料库（下文简称ACP）与第一届古代汉语分词和词性标注国际评测（EvaHan2022）发布的数据集（下文简称EVA）(李斌, 袁义国, 芦靖雅等, 2023)进行了对比。EVA语料主要包括《左传》《史记》和《资治通鉴》。《资治通鉴》成书于北宋，我们计算时将此部分排除，保留《左传》和《史记》。将ACP、ASC的标注数据与EVA进行了句子级别的对齐后，以EVA为基准，进行F1值的计算，结果如下表所示：

由于ASC、ACP及EVA的词性标记存在较大差异，我们只计算了ASC的分词指标，没有计算其词性标注指标。以ASC为基准，则ACP分词F1为95.59%，可见分词方面，二者的一致性很高。我们重点分析了ACP与EVA的标注差异。标注差异主要有三种情况。首先，二者对分词粒度的把握不同，如：

例1. ACP：及/c **宋人/n** 盟/v 于/p 宿/n
例2. EVA：及/c **宋/n 人/n** 盟/v 于/p 宿/ns

	数据集	分词F1	词性标注F1	备注
EvaHan2022最佳指标	基测集	96.16%	92.05%	封闭测试
	盲测集	93.64%	87.77%	封闭测试
ASC 指标	基测集	92.50%	/	/
	盲测集	92.90%	/	/
ACP 指标	基测集	93.70%	89.49%	/
	盲测集	92.83%	89.86%	/

Table 5: ACP、ASC与EVA 标注一致性对比

例3. ACP：**郑/n 公子忽/n** 为/v 质/n 于/p 周/n
例4. EVA：**郑公子忽/nr** 为/v 质/n 于/p 周/n

上例中，可以看到ACP和EVA对名词不同的处理方式。由于名词出现频次最高，这样的情况占据了二者差异的主要部分。在EVA中，我们也可以看到对“郑公子忽”这一单位不同的处理方式：

例5. EVA：**郑/ns 公子忽/nr** 在/v 王所/n

其次，ACP和EVA对一些结构有不同的判定，如：

例6. ACP：庄公/n **寤/a 生/v** ，/w 惊/v 姜氏/n ，/w 故/c 名/v 曰/v 寤生/n 。/w
例7. EVA：庄公/nr **寤生/v** ，/w 惊/sv 姜氏/nr ，/w 故/c 名/n 曰/v 寤生/nr 。/w 例句中出现了两次“寤生”。“庄公寤生”中“寤生犹言逆生，现代谓之足先出”(杨伯峻, 2001)，“寤生”应当标注为短语，ACP的标注更为恰当。第二个“寤生”作为人名出现，ACP和EVA都标注为名词。又如：

例8. ACP：去/v 顺/n 效/v 逆/n ，/w **所/u 以/p** 速/v 祸/n 也/u 。/w
例9. EVA：去/v 顺/n 效/v 逆/n ，/w **所以/c** 速/sv 祸/n 也/y 。/w

王力(1989)指出“所以”完全变为连词大约是从晋代开始，杨伯峻、何乐士(2001)认为“‘所以’作为一个连词，则大约是汉以后的事。”“所以”在上古时期是两个词，现代汉语的“所以”是由介宾结构逐步虚化而来的。因而，ACP将上例中的“所以”标注为两个词。

最后，ACP和EVA对词类及标记的划分并不一致，ACP的词类标记有13个，EVA的词类标记有21个，尽管我们已经将词标记进行了对齐，但这也在一定程度上影响了词性标注F1的计算。

上古汉语的词的划分历来是难点，实际标注过程中也有不同的操作标准，因而导致各家的标注结果会有所不同。对于难以判定的语素组合，ACP比EVA的分词粒度更小，ACP采取能分则分的策略。总体而言，我们认为，ACP的分词粒度符合上古汉语的实际情况，词性标注准确，能较好地满足科研和应用的需要。

4.3 语料库的应用

上古汉语分词和词性标记语料库不仅可以提供更加准确、灵活的词检索和统计，研究者更可以对定量模式进行定性的功能解释，对于汉语词汇的意义、结构和用法研究包括词汇面貌的整体描写、词义分析、词用调查、类型调查、形态特征分析等有重要的作用。

近年来“大数据”“深度学习”等前沿技术对于现代的汉语研究和应用起到了很大的推动，研究人员越来越渴望借助计算机对古籍开展全面、深层的研究。上古汉语分词和词性标记语料库可以作为一个基础的资源用作机器训练和学习，作为前沿技术服务于古汉语研究和应用的支撑性资源。

我们正在开发语料库检索平台和统计分析工具，未来将开放给语言研究、语言教学、辞书编撰、机器翻译等领域使用。

5 结语

语料库加工的本质是语言知识的重新整理、发现、形式化、规范化等工作，是最基本、最重要的基础研究之一。论文介绍了上古汉语词性标注语料库的构建过程的思路、难点和过程，并从词汇词频、词性和词长和分析了语料库的分布。论文的研究有助于填补学界缺少开放的大规模、深标注古汉语语料库的空白。论文的特色之处在于采用了人工标注为主+ 自动化工具辅助的模式，对词和词组的处理较为准确。下一步，将从三个方面继续开展工作：增加语料库的内容，最终实现对上古文献的全覆盖；继续对已经标注的语料开展人工和自动校对，提高标注的一致性和准确性；分批发布语料，共享给学术界和工业界使用。语料库加工的本质是语言知识的重新整理、发现、形式化、规范化等工作，是最基本、最重要的基础研究之一。论文介绍了上古汉语词性标注语料库的构建过程的思路、难点和过程，并从词汇词频、词性和词长和分析了语料库的分布。论文的研究有助于填补学界缺少开放的大规模、深标注古汉语语料库的空白。论文的特色之处在于采用了人工标注为主+ 自动化工具辅助的模式，对词和词组的处理较为准确。下一步，将从三个方面继续开展工作：增加语料库的内容，最终实现对上古文献的全覆盖；继续对已经标注的语料开展人工和自动校对，提高标注的一致性和准确性；分批发布语料，共享给学术界和工业界使用。

参考文献

GB/T 20532-2006, 信息处理用现代汉语词类标记规范[S]. 北京: 中国标准出版社, 2007.

STURGEON D. *Chinese Text Project*[EB/OL]. [2023-04-01]. https://ctext.org/

TANG X M, SU Q. 2022. *That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory*, volume 1. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7830–7840, Dublin, Ireland. Association for Computational Linguistics.

北京爱如生数字化技术研究中心. 中国基本古籍库[EB/OL]. [2023-04-01]. http://dh.ersjk.com/

北京大学中国语言学研究中心. *CCL*语料库检索系统[EB/OL]. [2023-04-01]. http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp

北京大学-字节跳动数字人文开放实验室. 识典古籍[EB/OL]. [2023-04-01]. https://www.shidianguji.com/

北京国学时代文化传播股份有限公司. 国学网[EB/OL]. [2023-04-01]. http://www.gxbd.com/

北京翰海博雅科技有限公司. 鼎秀古籍全文检索平台[EB/OL]. [2023-04-01]. http://103.242.200.9/ancientbook/portal/index/index.do

北京时代瀚堂科技有限公司. 瀚堂典藏古籍数据库[EB/OL]. [2023-04-01]. https://www.hytung.cn/

北京书同文数字化技术有限公司. 书同文古籍数据库[EB/OL]. [2023-04-01]. http://guji.unihan.com.cn/

卜师霞. 源于先秦的现代汉语复音词初探[M]. 北京: 中华书局, 2018.

陈小荷, 冯敏萱, 徐润华等. 先秦文献信息处理[M]. 北京: 世界图书出版公司, 2013.

古联（北京）数字传媒科技有限公司. 籍合网[EB/OL]. [2023-04-01]. https://www.ancientbooks.cn/

国家语言文字工作委员会. 中华经典资源库[EB/OL]. [2023-04-01]. https://zhjdzyk.eduyun.cn/

国学大师. 国学大师[EB/OL]. [2023-04-01]. http://www.guoxuedashi.net/

化振红. 建立中古汉语语料库分词规范的若干问题[J]. 语言研究集刊, 2021, (06): 151-167.

黄水清, 王东波. 国内语料库研究综述[J]. 信息资源管理学报, 2021, 11(03): 4-17+87. DOI: 10.13365/j.jirm.2021.03.004.

李斌, 冯敏萱, 陈小荷. 基于标注语料库的先秦汉语词汇统计[R]. 武汉: 第十三届汉语词汇语义学研讨会, 2012.

李斌, 袁义国, 芦靖雅等 . 第一届古代汉语分词和词性标注国际评测[J]. 中文信息学报, 2023, 37(03): 46-53+64.

李零. 简帛古书与学术源流（修订本）[M]. 北京: 三联书店, 2008.

马真. 先秦复音词初探[J]. 北京大学学报(哲学社会科学版), 1981, (5): 77.

毛远明. 左传词汇研究[M]. 重庆: 西南师范大学出版社, 2000.

王力. 汉语史稿[M]. 北京: 中华书局, 1980.

王力. 汉语语法史[M]. 北京: 商务印书馆, 1989.

王力. 王力文集*(*第十六卷*)*[M]. 山东教育出版社, 1990: 315.

王宁. 当代理论训诂学与汉语双音合成词构词研究[A]. 沈阳, 冯胜利主编. 当代语言学理论和汉语研究[C]. 北京: 商务印书馆, 2008.

王宁. 训诂学原理[M]. 北京: 中国国际广播出版社, 1996.

杨伯俊, 何乐士. 古汉语语法及其发展（修订本）[M]. 北京: 语文出版社, 2001.

杨伯峻. 春秋左传注（修订本）[M]. 北京: 中华书局, 2016.

杨丽姣, 肖航, 刘智颖. 《信息处理用现代汉语词类标记规范》修订研究[J]. 语言文字应用,2021,(03):111-120.

殷国光. 《吕氏春秋》词类研究[M]. 北京: 商务印书馆, 2008.

俞敬松, 魏一, 张永伟等 . 基于非参数贝叶斯模型和深度学习的古文分词研究[J]. 中文信息学报, 2020, 34(06): 1-8.

中国国家图书馆. 中华古籍资源库[EB/OL]. [2023-07-01]. http://read.nlc.cn/thematDataSearch/toGujiIndex