先秦词网构建及梵汉对比研究*

卢雪晖
南京师范大学文学院
1164800955@qq.com

徐会丹 陈思瑜
北京师范大学汉语文化学院中文信息处理研究所
{1935374738;1141522940}@qq.com

李斌✉
南京师范大学文学院
libin.njnu@gmail.com

摘 要

先秦汉语在汉语史研究上具有重要地位，然而以往的研究始终没有形成结构化的先秦词汇资源，难以满足古汉语信息处理和跨语言对比的研究需要。国际上以英文词网（WordNet）的义类架构为基础，已经建立了数十种语言的词网，已经成为多语言自然语言处理和跨语言对比的基础资源。本文综述了国内外各种词网的构建情况，特别是古代语言的词网和汉语词网，然后详细介绍了先秦词网的构建和校正过程，构建起了涵盖 43591 个词语、61227 个义项、17975 个义类的先秦汉语词网。本文还通过与古梵语词网的跨语言对比，尝试分析这两种古老语言在词汇上的共性和差异，初步验证先秦词网的有效性。

关键词：词网；古汉语词网；跨语言对比；古文信息处理

The Construction of Pre-Qin Ancient Chinese WordNet and Cross-Language Comparative Study between Ancient Sanskrit WordNet and Pre-Qin Ancient Chinese WordNet

Xuehui Lu
School of Chinese Language and Literature, Nanjing Normal University
1164800955@qq.com

Huidan Xu Siyu Chen
Institute of Chinese Information Processing, Beijing Normal University
{1935374738；1141522940}@qq.com

Bin Li✉
School of Chinese Language and Literature, Nanjing Normal University
libin.njnu@gmail.com

Abstract

Pre-Qin ancient Chinese plays an important role in the study of Chinese history. However, previous studies have not formed structured pre-Qin vocabulary resources, which is difficult to meet the needs of information processing and cross language comparison of ancient Chinese. Based on the semantic class structure of WordNet, dozens of language WordNets have been established in the world, which has become the basic resource for multilingual natural language processing and cross language comparison. This paper summarizes the construction of various WordNets at home and abroad, especially the ancient language and Chinese WordNets, and then introduces the construction and calibration process of the wordnet for Pre-Qin ancient Chinese (named PQAC Wordnet(PQAC-WN)) in detail, covering 43,591 words, 61,227 senses, and 17,975 synsets. Through the cross language comparison with the ancient Sanskrit WordNet, this paper attempts to analyze the similarities and differences of the two ancient languages in vocabulary, and preliminarily verify the effectiveness of the PQAC-WN.

Keywords: WordNet, ancient Chinese WordNet, cross-language contrast, ancient Chinese information processing.

* 基金项目：国家社科基金项目（18BYY127）；国家社科基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”（15ZDB127）；江苏高校哲学社会科学优秀创新团队建设项目（2017STD006）；江苏省社会科学基金项目（20JYB004）；国家语委科研项目（YB135-61）。

1 引言

先秦汉语是从公元前 2146 年发展至公元前 221 年的古代汉语，属于上古汉语，是汉语的源头，在汉语史研究中具有重要地位。从词汇发展的角度来看，在先秦时代，汉语的名词、动词、形容词等逐渐形成系统，其中有很多也是作为基本词或词素一直沿用到现代（王力，1980；郭锡良，2000）。对先秦汉语词汇的深入研究，有助于加深对汉语词汇变化规律和汉语词汇现状的认识，也可以为先秦汉语研究以及以先秦文献为基础的其它学术研究提供有效的支持。但是目前的已有研究多是对专书或专类的个案研究，很难形成完整的体系，没有形成结构化的先秦词汇资源。

借助词网（WordNet）这一国际通用的词汇语义框架，建构先秦词网，可以直接对先秦词汇系统的全貌进行研究，为先秦词汇语义研究提供新的材料和新的方式。目前，徐会丹等人（2019）已初步构建了涵盖 39623 个词语、57355 个义项和 16431 个义类的先秦词网，并通过对已构建的先秦词网的计量分析，系统展现了先秦词汇语义系统的整体概貌，为词汇语义研究提供了新的材料和新的方式。同时，作为结构化的知识资源，先秦词网也可以为资源稀缺的先秦汉语提供机器可用的知识资源，弥补先秦汉语信息处理资源匮乏的不足，满足先秦汉语信息处理的需要（张颖杰，2017）。

先秦词网目前虽然已经取得了很大的成果，但构建的准确率和覆盖率仍有待提高。一是先秦汉语词义映射到词网时存在不精确映射和错误映射，导致现已构建的先秦词网中存在部分错误，准确率有待进一步提高；二是由于目前的映射方法存在一定局限性，在某个先秦汉语词义在可以映射到词网的多个节点的情况下，却只会被映射到词网的一个结点上，导致映射不完全，使得整体的映射覆盖率也还不是很高。因此本文将进一步在现已构建的先秦词网的基础上进行校正，继续完善先秦词网的构建体系，构建一个准确率和覆盖率更高的先秦词网。

另外，相比于语法和语音，不同语言间在词汇语义上的对比能更为有效地体现语言的共性和差异。词网作为一个国际化的词汇语义框架，为进行跨语言对比提供了可能。和先秦同时期发展的是印度和希腊—罗马等古国，都是人类文明以及语言学的发源地。因此，本文将从跨语言对比的角度出发，将先秦词网和古梵语词网进行对比，探讨先秦汉语和梵语这两种古老语言的词汇语义系统的异同，从而进一步证明先秦词网这一词汇数据库的应用价值。

全文的结构如下：第 2 节梳理了相关工作。第 3 节简述了先秦词网的构建和校对过程，进一步完善了先秦词网的构建体系。第 4 节初步统计了先秦词网和古梵语词网的对比研究结果。第 5 节是结论和未来工作。

2 相关工作

2.1 先秦词汇语义研究

先秦汉语是古汉语研究的重要课题。先秦汉语的资源有出土文献和传世文献两大系列，但出土文献较为零散，传世文献为主要研究对象，而传世文献的数量并不多。比如中国台湾“中央研究院”开发的“古汉语标记语料库”，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，对重要的先秦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。南京师范大学建立的先秦汉语标注语料库收录了先秦传世的主要文献，有 25 种，如《左传》、《吕氏春秋》等，共 180 多万字（李斌，2010）。总体来说，先秦汉语的资源还较为匮乏。

近年，在先秦汉语词汇研究领域取得了丰富的研究成果。根据搜集的文献资料，主要分为三个方面：（1）先秦专书词汇研究，如《左传》、《诗经》、《孟子》等先秦文献都有了对应的专书词典；（2）先秦专类词汇研究，既包括根据词汇领域分出的语义类别词，也包括根据词频、词义、词形等标准分类得到的常用词、同义词、复音词等；（3）先秦专书专类词汇研究，这类研究是上面两类的结合，指的是针对先秦的某部文献进行的某个类别词汇的研究。对上述的先秦词汇研究进行总结，我们发现，先秦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，重视文献诠释，而轻视理论资源的建设。虽然成果十分丰富，但缺少能够系统展现先秦词汇语义全貌的总结性研究。

第二十届中国计算语言学大会论文集，第108页-第120页，呼和浩特，中国，2021年8月13日至15日。

针对先秦汉语这种资源稀缺语言，结构化的知识资源对于其信息处理是十分重要的，但以往的知识资源往往未经处理很难被直接应用（张颖杰，2017）。以《尔雅》为例。《尔雅》是中国最早的词典，首创按词义分类编排的体例，囊括释诂、释言、释训、释亲、释官、释器等共计 19 个门类，其中，《释诂》、《释言》、《释训》3 篇释一般词语，其余 16 篇释特殊词语，且前 3 篇各条大都以同义词类聚，训释详尽具体，后 16 篇从语义层面在各种意义范畴中集中辨释了大量特殊文化词语，共收录 4300 多个词语。《尔雅》以独特的编排体例和简明的训释方法，为后世字典、词典的编纂树立了典范。但《尔雅》一书在创建伊始，其主要的对象是人而非计算机，义类系统较为粗略，没有很好的结构化，无法直接为自然语言处理任务所用。但是，这种按照词汇的概念进行分类的方法仍对历代学者研究先秦的词汇语义具有极大帮助。

词网(WordNet)是一个结构化的语义资源，可以提供基本的词汇概念及其关系信息，是一个十分基础且重要的语义资源（袁毓林，2008）。借助词网这一国际通用的词汇语义框架，建构先秦词网，可以为先秦的词汇语义研究提供新的材料和新的方式。

2.2 词网研究

2.2.1 词网发展概述

自 20 世纪 80 年代以来，语义分析一直是自然语言处理中的研究热点和难题，而语义词典是语言知识用于语义分析的重要基础。词网（Princeton WordNet， PWN）是一个大型的英语词汇数据库1，它来源于 1978 年普林斯顿大学 Miller 教授所主持的一个知识工程的项目（Miller, 1995）。词网不同于传统词典的是，它是根据词义而不是词形来组织词汇信息。在词网中，名词、动词、形容词和副词被组织成同义词集合（Set of Synonyms，Synset），并在各个同义词集之间建立一种指针，以此来表示各种语义关系，包括同义关系、反义关系、上下位关系、整体部分关系、蕴涵（推演）关系等等。目前，WordNet 忽略英语中较小的虚词集， 共涵盖了名词、动词、形容词和副词的 111223 个概念，有近 95600 种不同的词型，组成了 70100 个同义词集合。

词网作为结构化的语义资源，在展现词汇语义系统概貌、挖掘词汇语义特点等方面具有重要作用，也被成功地用于词义消歧、语言自动处理、双语及多国语机器翻译、检索系统等一系列语言工程，已经逐渐成为国际通用的框架结构。90 年代后，各国竞相参照 PWN 开发本国语言的词网语义词典。例如荷兰、西班牙、意大利、英国、法国、德国、捷克、爱沙尼亚等国家都参与了构建 EuroNet 系统（Vossen et al.，1998）。2002 年，Stamou 等人开发了代表巴尔干语言的基本概念及其之间的语义关系的一个多语言资源 BalkaNet（Stamou et al.，2002），包括中欧和东欧的希腊语、土耳其语、罗马尼亚语、保加利亚语、捷克语和塞尔维亚语。Vossen 在荷兰成立国际组织“全球词网联盟”（Global WordNet Association， GWA）[2]，为讨论、共享和连接世界上所有语言的词网提供了一个平台。截至 2020 年，该平台已有 77 种词网项目，涵盖了世界上 200 多种语言。

近年来，国内也出现了越来越多中文词网的相关成果，它们大多集中于现代汉语词网的构建。目前，已经开发了 7 个中文词网。分别是北京大学计算语言学研究所 2001 年构建的中文概念词典（Chinese Concept Dictionary, CCD）（于江生 等，2002）、东北大学 2003 年构建的中文词网（Northeastern University Wordnet, NEW）（Zhang Xi et al.，2003）、台湾大学和中央研究院 2004 年构建的汉语双语本体论词网（Sinica Bilingual Ontological Wordnet, BOW）（Huang C R，2003）、东南大学 2008 年构建的中文词网（Southeast University Wordnet, SEW）（Renjie Xu et al.，2008）、台湾大学和中央研究院 2010 年构建的中文词网（Chinese WordNet, CWN）（Chu Ren Huang et al.，2010）、南洋理工大学 2013 年构建的中文开放式词网（Chinese Open Wordnet， COW）（Shan Wang et al.，2014）、曲阜师范大学 2018 年构建的多融合中文词网（Multi-Fusion Chinese Wordnet，MCW）（Mingchen Li et al.，2020）。这些中文词网对中文信息处理的发展起到了良好的促进作用。

2.3.2 古代语言词网发展概述

[1] https://wordnet.princeton.edu/
[2] http://globalwordnet.org/

古代语言作为现代语言的来源和基础，具有很高的语言学、文学、文献和史料价值。如今，世界上已经兴起了针对资源稀缺的古代语言的信息处理技术的探索（张颖杰，2017）。在 GWA 上，古拉丁语、古希腊语、古梵语以及希伯来语等都被建构注册。本文研究先秦词网的构建、完善和统计分析，所以在此对古代语言词网的构建做一简要介绍。

目前，词网的构建方式主要有两种，Vossen（1999）将这两种方法分别称为扩展法和合并法。所谓扩展法，是指将 PWN 中的同义词集合等价翻译为目标语言，再根据目标语言的特点，继承并修改 PWN 中的语义关系，从而构建目标语言的词网。所谓合并法，指的是目标语言的词网首先独立于 PWN 构建，定义自己的同义词集合和关系，再将目标语言的词网与 PWN 的概念和关系对齐。先前的工作表明，与合并法相比，扩展法更通用，更容易实现映射，且避免了大量的词典编纂工作；但是，它的构建过程深受 PWN 框架影响，最终将较少保持目标语言的特有属性。近年出现的一些用古代语言构建的词网，如拉丁语、古希腊语、古梵语以及希伯来语等古代语言大多是采用扩展法构建的。

希伯来语（Ordan and Wintner，2007）[3]作为第一种设计了词网的闪族语言，基于多语词网(Multilingual 词网，MWN)中的原始希伯来语词网，采用扩展法进行了构建。目前包含 5261 个同义词集合，7735 个词语。作为跨语言词汇数据库对齐的测试案例，为构建词网时如何处理语言之间的差异提供了有效借鉴。

而古拉丁语词网（Latin WordNet, LWN）[4]在一定程度上与希伯来语的构建历程相似。它最初由 Stefano Minozzi 在 2004 年至 2008 年间创建，同样是多语词网（MWN）的组成部分，包含 9378 个古拉丁词语和 8973 个同义词集合。埃克塞特大学（University of Exeter）在此基础上进行扩展，添加了约 30000 个词语，覆盖从古代到晚期（甚至更远）的拉丁语，目前已包含 35603 个古拉丁词语和 46702 个同义词集合。

古梵语词网（Sanskrit WordNet，SWN）[5]在 William Michael Short 的指导下，由意大利帕维亚大学（University of Pavia）和英国埃克塞特大学（University of Exeter）共同创建。如今它包含 33604 个古梵语词和 32151 个同义词集合。

古希腊语词网（Ancient Greek WordNet，AGWN）[6]（Bizzoni et al，2014）则是自动构建古代语言词汇语义资源的新范式。它从“希腊—英语”词典中提取同义词集合，根据“由同一个英语单词或短语翻译的希腊语词语很可能是同义词或至少在语义上是密切相关的”这一前提，将希腊语同义词集合与 PWN 中含有对应英语单词的集合相关联，从而构建成一个完整的古希腊语词网。目前，它涵盖了 66319 个古希腊词语和 45367 个同义词集合。

相比于以上语言，古汉语本身作为孤立语，几乎没有形态上的变化，并以书面语言的形式被记录在了历史文献中，为构建提供了便利。但由于古汉语与现代汉语存在着巨大的差异，现代汉语中现有的词汇资源，如 CCD 等不能直接用于古汉语词网的研究。针对以上问题，国内学者借鉴了国外古希腊语、古梵语等古代语言构建词网的经验，构思建造了古汉语的词网。

张颖杰等（2014）提出了一种自动构建中古汉语词网（middle ancient Chinese WordNet, MidacWordNet）的通用的方法。他们基于扩展法，使用解释性历时词典《汉语大词典》作为原始资源，从中提取中古汉语词汇及其释义，CCD 被用作映射到 PWN 上的中介。最终得到的中古汉语词网中包括 31345 个中古汉语词义，涵盖了 GCD 中中古汉语全部词义的 35%，映射准确率达到了 80%，显示了中古汉语和英语词汇方面的巨大差异。2017 年，张颖杰（2017）等又提出了一种自动构建先秦词网(Pre-Qin ancient Chinese WordNet, PQAC-WN)的方法。同样采取扩展法，以《汉语大词典》作为原始资源，以 CCD 和《同义词词林》的合并后得到的 CCD-Extend 作为映射中介，将先秦词义映射到 PWN 同义词集合上，并在映射过程中，应用了基于图的词义消歧方法。最后构建的

[3] http://cl.haifa.ac.il/projects/mwn/
[4] https://latinwordnet.exeter.ac.uk/
[5] https://sanskritwordnet.unipv.it/
[6] https://greekwordnet.chs.harvard.edu/

PQAC-WN 包括 43255 个词义，覆盖 GCD 中先秦汉语全部词义的 79.99%，准确率达到 85%以上。

徐会丹等（2019）同样采取扩展法，对先秦时代的词汇进行了人工映射，初步构建了先秦词网(Pre-Qin ancient Chinese WordNet, PQAC-WN)，具体构建方法是首先借助基于《汉语大词典》建立的汉语词义演变时代数据库，以书证朝代为“先秦”进行筛选，形成“先秦汉语—现代汉语”的古今释义词表；然后借助 CCD 中的“英语—现代汉语”的双语资源，以现代汉语为中介，人工标注获得“先秦汉语—英语”的映射结果，最终构建起了涵盖 39623 个词语、57355 个义项、16431 个义类的先秦词网。

纵观上述构建成果，各个古代语言的词网在构建和扩充的过程中不断扩大词语数量和 PWN 的映射规模，其中古拉丁语、古梵语、古希腊语和先秦词网的规模较大。各词网间的具体对比情况如表 1 所示：

词网	词语	义项	义类	构建方法
希伯来语词网	7735	-	5261	扩展法
古拉丁语词网	35603	-	46702	扩展法
古梵语词网	33604	-	32151	扩展法
古希腊语词网	66319	-	45367	扩展法
先秦词网(张颖杰等（2017）)	-	43255	-	扩展法
先秦词网(徐会丹等（2019）)	39623	57355	16431	扩展法

表 1 各词网间的具体对比

在前人的基础上，下一节将介绍本文采取的先秦词网的构建及校对方法。

3 先秦词网的构建及校对

3.1 构建方法

构建一种新的语言的词网主要有两种方法：合并法和扩展法。前者需要大量的先秦词典编纂工作，后者可能忽略先秦词汇的文化特色，两种方法各有利弊。借鉴国内外已有的研究成果，本文仍沿用了徐会丹等（2019）的构建方法，采用扩展法，以 PWN 作为构建先秦词网的起点。该构建思路是将先秦词汇的义项映射到 PWN 的义类，继承 PWN 的概念节点、语义关系和层次结构；然后针对先秦词汇语义的特点进行节点调整，保留特有的语言属性。下面简单介绍一下构建具体过程。

3.1.1 词表来源

我们以《汉语大词典》作为原始资源，从中获取目标构建的词网需要涵盖的所有的先秦词汇语义信息，这项工作的实现，以前人对《汉语大词典》中所有义项年代信息的标注结果为基础。刘雪扬（2015）以词义演变为切入点，人工标注了《汉语大词典》中 45 万多个义项的年代信息，形成汉语词义演变时代数据库。本文就是基于这一现有数据库，以书证朝代为“先秦”进行筛选，获得了先秦词库。最终，我们得到的先秦词库包含词汇 45498 个，义项 63230 个。该词库包括只在先秦时代产生的先秦时代特有词汇，也包括在到当代的进程中在某一时代消亡，或一直沿用至今的词汇。

3.1.2 映射中介

CCD 是现成的英汉双语语义资源，可以直接投入使用。它以 1997 年发布的词网 1.6 版本为基础，整体上可以看作是词网的中文版本。CCD 是构建先秦词网、实现先秦汉语到英语映射的一个很好的中介选择。

本文将从先秦词库出发，将先秦词语的每一个义项都映射到 CCD 对应的现汉同义词集合（即义类）中，多义词会被映射到多个集合中。通过映射，我们获得了先秦同义词集合与 CCD 的中文同义词集合（CSynset）的对应关系，而 CCD 作为 PWN 的汉化版本，提供中文同义词集合（CSynset）

和英文同义词集合（Synset）的对应关系，由此可以抽取“先秦同义词集合—英文同义词集合”的对应关系。

3.1.3 映射方法

我们在将先秦词网的同义词集与 PWN 的同义词集联系起来的映射过程中主要遵循了以下三种映射方法：

（1）**直接映射**，为一个先秦义项提供 PWN 中精确匹配的义类。例如：

词形	义项	CCDID	SYNSET	CSYNSET
矜	端庄	01476068s	decent	沉稳 稳重 端庄

表 2 先秦词“矜”义项的映射结果

（2）**上位映射**，把先秦汉语中无法直接精准映射到 PWN 的节点上的义项映射到 PWN 中的一个上位义类。上位映射并不是为词网的框架的层级结构新增节点，在词网中建立起某一确定的义类或同义词集，而是为了更精确的体现先秦词汇的概念系统，与直接映射有所区分而采取的一种映射方法。例如：

词形	义项	CCDID	SYNSET	CSYNSET
來	省亲，特指已嫁女子回娘家省亲	01368235v	go_home head_home	回家 返家

表 3 先秦词“來”义项的映射结果

（3）**新增节点**，为一些先秦特有义类增加新的中间义类节点，并根据上下位关系将其添附于词网 框架中的相应父节点上。采取“父节点+新增子节点”的两级编号方式进行标注。例如新增节点“礼器”的编号为父节点“器皿”的编号加上子节点序号，即“03574512n-01”：

词形	义项	新增节点	CCDID	SYNSET	CSYNSET
榃	古代礼器。为长方形的木承盘	03574512n-01	03574512n	vessel	罐 器皿 容器 盛器

表 4 先秦词“榃”义项的映射结果

3.2 校正方法

我们随机抽取了徐会丹等（2019）构建的先秦词网中 57355 个义项中的 500 个义项，发现其语义映射的精确度大约为 90.4%左右，仍存在着 9.6%左右的错误映射和暂无映射的情况。统计结果如表 5 所示：

类型	个数	占比	共计
精确映射和上位映射	442	88.4%	90.4%
新增节点映射	10	2%	
错误映射	33	6.6%	9.6%
暂无映射	15	3%	
总计	500	100%	100%

表 5 徐会丹等（2019）构建的先秦词网的精确度抽样统计

在这一背景之下，本文将在现已构建的先秦词网的基础上继续校正和改进，构建一个准确率和覆盖率更高的先秦词网。

3.2.1 补充和纠错

为了提高先秦词网的准确率和覆盖率，我们首先进行了补充和纠错，改正了发现的暂无映射和错误映射的义项。目前，已补充了 5823 个义项的映射，所有的义项都通过精确映射、上位映射或新增节点映射与 PWN 进行了映射。同时，纠正了 2683 个义项的映射错误，提高了映射的准确率。补充和纠错的具体例子如表 6 所示。

项目		补充	纠错
词语		譁譁	謗言
义项		舒徐迟缓貌	怨恨、指责的话
例句		《荀子・乐论》:“盡筋骨之力以要鐘鼓俯會之節，而靡有悖逆者，衆積意譁譁乎!”	《左传・成公十八年》:“舉不失職，官不易方，爵不踰德，師不陵正，旅不偪師，民無謗言，所以復霸也。”
原映射结果	CCDID	-	00560484v
	SYNSET	-	reprehend
	CSYNSET	-	指责 谴责 责备 非难
	DEFINITION	-	express strong disapproval of
	CDEFINITION	-	表达强烈的不满
现映射结果	CCDID	02289454s	05033972n
	SYNSET	relaxed	rebuke reproof reproval reprehension reprimand
	CSYNSET	和缓 弛缓 放松 缓和 舒缓 迟缓	指责 斥责 训斥 谴责 责备
	DEFINITION	made less tense or rigid	an expression of criticism and censure
	CDEFINITION	不紧张的或不严格的	一种批评或谴责的表示

表 6 补充和纠错的具体例子

虽然通过人工进行语义标注映射的精确度相比其他方法有更高的精确度，但由于是由人来进行是否概念对应的判断，难免会带有主观性，从而带来错误，难以做到真正的客观。我们也将继续在今后进行多次差错校对，来提高标注的准确性，降低主观性。

接下来，我们进一步改进了构建时的映射方法，主要包括将映射关系从一对一映射改为多对多映射，以及对新增节点的映射方法做了进一步的细化和补充。

3.2.2 改进映射关系——从一对一映射到多对多映射

在前人的构建规范中，先秦义类完全通过先秦词的义项与英语义类的一对一映射得到。这导致虽然最终一个英语义类可能包含多个先秦词义项，但一个先秦义项只会被包含在一个英语义类中，从而整体的映射覆盖率并不是很高，CCD 覆盖了 PWN 中的全部义类共 117659 个义类，而目前构建的先秦词网只覆盖了 PWN 中的 16413 个义类，很多应该覆盖的义类没有覆盖到。

本文对这种构建方法做出了改进，将映射方式改为多对多映射，即一个先秦义项可能会被包含在 PWN 的多个义类中，也就是多个同义词集合（Synset）中，而多个不同的先秦义项也有可能被映射到 PWN 的同一个义类中。如表 7、表 8 中先秦词“倉”和“蒼”的义项都是“青色”，都映射到了 CCD 对应的“青色”义类，进而映射到了英语同义词集合“blue bluish blueish”和“blue blueness”这

两个义类上。其中，词形、义项、例句等都是先秦词库中的字段，CCDID、SYNSET、CSYNSET、DEFINITION、CDEFINITION 等都是 CCD 中的字段。CCDID 字段和 PWN 中的 ID 码（Synset_id）是完全一致的。

词形	义项	例句
倉	通“蒼”。青色	《仪礼・聘礼》：“繅三采六等朱白倉。”参见“倉龍”、“倉玉”。
蒼	青色(包括蓝色和绿色)。	《诗・秦风・黄鸟》：“彼蒼者天，殲我良人!”唐 韩愈《条山苍》诗：“條山蒼，河水黄。”陈其通《万水千山》第八幕：“狂风暴雨只能吹掉苍松的一些枝叶。”

表 7 先秦词“倉”和“蒼”的义项和例句

CCDID	SYNSET	CSYNSET	DEFINITION	CDEFINITION
00370667s	Blue,bluish, blueish	青色	tinged with blue or purple from cold or contusion	因为寒冷或者外伤而带有淡淡的蓝色或者紫色
03883954n	Blue, blueness	蓝色,青色,天蓝色	the color of the clear sky in the daytime	白天晴朗的天空的颜色

表 8 先秦词“倉”和“蒼”义项与 PWN 的映射结果

通过这种映射方法，我们就可以得到“先秦同义词集合—英文同义词集合”的多对多的对应关系。把映射关系从一对一映射改为多到多映射之后，目前的先秦词网涵盖了 43591 个词语、61227 个义项、17881 个义类，对 PWN 的整体映射覆盖率得到了提高。

3.2.3 新增义类的改进

目前的先秦词网中包含 44 个新增义类节点，如：“省亲”、“礼冠”、“谥号”、“卦”、“揖”、“穴位”等，包含新增的 794 个词语，体现了先秦词汇的特点。本文在原来的基础上，又新增了 76 个的大的语义类，如“庄稼单位”、“庙”、“钟鼎”、“玉”、“服饰”、“农历”等，包含新增的 1050 个词语，进一步体现了先秦词汇的特点。目前的先秦词网共包括 120 个新增义类，包含 1844 个词语。

对新增义类进行补充的同时，我们还做了更细致的处理，进一步考虑了新增义类节点下的层级结构，实现了从新增概念节点到新增概念子树的转化。下面以义类节点 04624523n“medical_science 医学”下的新增义类节点 04624523n-01“中医”的处理为例来进行说明。

新增义类	义类节点 ID	义项示例	义项数量
身体器官	04624523n-01-01	水藏,中医指肾脏。	42
穴位	04624523n-01-02	尺澤,针灸穴位名。位于肘横纹上肱二头肌腱桡侧。	44
治疗方法	04624523n-01-03	湯熨,中医的一种治疗方法。用热水熨帖患处以散寒止痛。	15
医学基本理论	04624523n-01-04	神藏,中医学谓神气蕴藏于内腑。	52

表 9 新增义类举例

在接下来的数据校对和修正过程中，仍需要对新增节点进行补充以及更细致的处理。通过不断的校正和改进，我们构建起一个准确率和覆盖率都更高的先秦汉语词网。具体结果如表 10 所示。

词网	词语	义项	义类
徐会丹等（2019）构建的先秦词网	39623	57355	16431
补充和纠错后的先秦词网	43591	61227	16672
多对多映射后的先秦词网	43591	61227	17899
改进新增义类后的先秦词网	43591	61227	17975

表 10 先秦词网的构建和校对数据

4 梵汉跨语言对比

作为语言学的研究课题，仅仅完成先秦词网的构建工作是不够的，还需要进行计量分析。徐会丹等（2019）就通过对已构建的先秦词网进行了统计分析，并与具有相同结构的 PWN 进行了对比分析。其构建的先秦词网涵盖 39623 个词语、57355 个义项和 16431 个义类。平均 1 个语义类含有 2.4 个词、3.5 个义项；平均 1 个先秦词含有 1.4 个义项。57355 个先秦义项能够映射到 CCD 的有 56525 个，占总数的 98.6%。而无法在 CCD 中找到对应概念、需要增加新节点的义项数量为 830 个，占比约 1.4%。这些统计结果说明先秦词网为探究先秦特色语义提供新的资源和方式。

本文试通过先秦词网和古梵语词网的初步跨语言对比分析，进一步证明本词汇数据库的应用价值，初步观察古老语言在词汇上的共性和差异。先秦词网和古梵语词网都是以 PWN1.6 版本为基础进行构建的，继承了 PWN 的概念节点、语义关系和层次结构。这使得进行二者间的对比是具有可行性的。

4.1 映射情况对比

经过一定的校对和补充，目前已构建的先秦词网涵盖 43591 个词语和 17975 个义类。古梵语词网涵盖 33604 个词语和 32151 个义类，先秦词网和古梵语词网具体的映射统计结果如表 11 所示：

类别	词语个数 （先秦词网/古梵语词网）	义类个数 （先秦词网/古梵语词网）
精确映射和上位映射	41747/17090	17855/9750
新增节点映射	1844/20736	120/22401
总数	43591/33604	17975/32151

表 11 各词网的映射结果

进一步统计包含 PWN 在内的三个词网的各词性的同义词集数量即义类数量，统计结果如表

12 所示。

词网	名词	动词	形容词	副词	共计
先秦汉语词网	8491（120 独有）	4674	3554	1295	17975
古梵语词网	27815(22401 独有)	1764	2278	294	32151
PWN	82115	13767	18156	3621	117659

表 12 各词网的义类数量

这些宏观数据可以让我们大体把握到先秦汉语和古梵语词汇语义系统的整体概貌和差别所在。两种语言的词汇数量和义类丰富度都很高，其中古梵语的义类丰富度更高一些。与古梵语相比，先秦汉语中的义类总量较少，其中名词义类较少，而动词义类较多。其原因可能是：古梵语的构建虽然是映射到 PWN 上的，但同时也独自增加了很多新的义类而没有和 PWN 映射，所以古梵语中的义类数量会相对较多。而先秦义类完全通过先秦词的义项与英语义类的映射得到，还没有添加很多新的义类，且这个过程中存在许多上位映射的情况，也就是说，许多先秦义项只是对应到了比较典型的义类上，没有进行更加细致的划分，所以会导致义类数量相对较少。这也是我们在后期校对先秦词网的过程中希望改进的地方。

4.2 共有义类

在词网中，词语根据同义关系聚合成同义词集合，代表不同的语义类。这些语义类被分为各基本类别中，每一类包含若干同义词集合。下面列举了 11 种名词和 5 种动词的起始概念及其对应的义类数量，其中，|先秦占比-古梵语占比|表示某个起始概念在先秦汉语义类中的占比减去该概念在古梵语义类中的占比的绝对值，绝对值越大，表示该概念在先秦汉语和古梵语中的分布差异越大。

名词起始概念	先秦义类		古梵语义类		*|先秦占比-古梵语占比|*
	个数	占比	个数	占比	
动作	1188	0.14	667	0.12	0.02
物质	340	0.04	273	0.05	0.01
认知	509	0.06	299	0.06	0.01
动物	325	0.04	156	0.03	0.01
状态	594	0.07	427	0.08	0.01
数量	172	0.02	174	0.03	0.01
植物	181	0.02	149	0.03	0.00
时间	74	0.01	59	0.01	0.00
处所	254	0.03	140	0.03	0.00
情感	123	0.02	109	0.02	0.00
身体	215	0.03	172	0.03	0.00
其他	4516	0.51	2789	0.52	0.00
总计	8491	-	5414	-	-

表 13 先秦汉语和古梵语名词义类的分布对比

动词起始概念	先秦义类		古梵语义类		|先秦占比-古梵语占比|
	个数	占比	个数	占比	
变化	794	0.17	330	0.19	0.02
接触	469	0.10	152	0.09	0.01
社会交互	426	0.09	175	0.10	0.01
行动	384	0.08	124	0.07	0.01
运动	276	0.06	103	0.06	0.00
其他	2325	0.50	1213	0.49	0.00
总计	4674	-	1764	-	-

表 14 先秦汉语和古梵语动词义类的分布对比

根据上述统计数据，我们可以发现：

（1）无论是名词义类还是动词义类，在总量上先秦汉语都略高于古梵语。但先秦汉语和古梵语的名词、动词义类在大多数概念类别中的分布比例基本一致。这从语义分布的角度进一步表明两种语言的概念系统大致重合，概念系统具有普遍性。

（2）先秦汉语和古梵语在分布上存在一定差异的名词类别有：动作、物质、认知、动物、状态、数量等，存在一定差异的动词类别有：变化、接触、社会交互、行动等。这一定程度地体现出了概念系统的特异性。

4.3 新增节点的异同

先秦汉语和古梵语词汇语义系统在词网的整体概貌之下，无法映射而新增的词更能反映这两种古老语言在词汇上的共性和差异。经过对新增节点进行补充以及更细致的处理，先秦词网目前共新增了 120 个大的节点，包含 1844 个词汇。古梵语新增了 22401 个小节点，都是名词节点，大都是一些专有名词，且没有对新增的义类进行进一步的语义类的划分。因此进行新增节点的对比时，我们通过梵语词网新增义类的英语释义，尝试将其划入某一类语义类，然后再与先秦词网进行对比。

下面以“祭祀”类义类和“神话人物”类义类进行简单分析。

	新增义类示例	词型示例
先秦汉语	祭祀名，古代因征战出师而祭天	禷
	祭祀用的酒肉	胙
	古代祭祀时献酒三次，即初献爵、亚献爵、终献爵，合称“三獻”	三獻
古梵语	a sacrificial rite connected with liquids（与液体有关的祭祀仪式）	रसकर्मन्
	a religious rite preparatory to a sacrifice or any solemn observance（为祭祀或任何庄严仪式做准备的宗教仪式）	स्वस्तिवाचन
	name of a particular sacrificial text（特定祭文的名称）	यजुस्

表 15 新增“祭祀”类义类示例

表 15 中这些“祭祀”类下的义类和词汇可以很直接地说明先秦汉语词汇和古梵语词汇都存在着很多祭祀类的词汇。这些词汇在 PWN 找不到对应的词汇，因此都新增了很多这类节点。祭祀时往往有一定的规范，包括祭祀的活动、仪式、祭品、祭文、祭器等，古人希望通过这些来表示崇敬并求保佑，反映了人类初期对自然界、祖先或宗教的崇拜。

	新增义类示例	词型示例
先秦汉语	传说中的天神名	東皇太一
	传说中的水神名	玄冥
	传说中的山神名	磈氏
古梵语	name of a Buddha （佛教的创始人佛陀）	देवराज
	name of a Bodhisattva（菩萨）	सर्वसत्त्वप्रियदर्शन
	name of a Brāhman（婆罗门）	शम्पाक

表 16 新增“神话人物”类义类示例

表 16 中这些“神话人物”类下的义类和词汇可以很直接地说明先秦汉语词汇和古梵语词汇虽然都有大量的神话人物类词汇，但在具体人物和产生背景上仍存在着很大不同。当时的先秦人民主要是想象存在着某一类掌管具体事物的神仙，从而掌管天地的运行，大多都是中国人民在当时不能解释很多现象的情况下想象出来的产物，比如“玄冥”为传说中的水神。而古印度是以宗教为中心的神话之邦，婆罗门教（印度教的前身）、佛教和耆那教先后兴起，让人们在不能理解世界上的各种未知事物时能够寻找安慰，在古梵语词网中，新增了很多这类节点，大多都是印度社会里被奉为神的宗教人物，例如佛教的创始人佛陀 Buddha（देवराज）。虽然崇拜的神有所不同，但这都说明了古代人民对自然的敬畏和崇拜。

综上，可以看出，通过对先秦词网和古梵语词网的初步的跨语言对比，我们可以大致探究这两种古老语言的词汇语义系统的异同。如果能够对两者进行更详细的对比分析的话，或者与更多建构了词网的语言进行跨语言对比的话，将会得到更多的词汇语义映射情况和词汇语义关系对比等方面的信息。

5 结语

先秦词网能够展示先秦词汇语义系统的整体面貌，满足古汉语信息处理和跨语言对比的研究需要，为探究先秦特色语义提供新的资源和方式。本文借助国际通用的词网框架，以《汉语大词典》中的先秦部分和 CCD 为基础资源，进一步在现已构建的先秦词网的基础上进行了校正，完善了先秦词网的构建体系，构建一个准确率和覆盖率更高的先秦词网，涵盖 43591 个词语、61227 个义项、 17975 个义类。另外，本文还进一步从跨语言对比的角度出发，将先秦词网和古梵语词网进行对比，探讨了这两种古老语言的词汇语义系统的异同，进一步证明了先秦词网这一词汇数据库的应用价值。

参考文献

[1] Tufis D, Cristea D, Stamou S. BalkaNet: Aims, methods, results and perspectives. a general overview[J]. Romanian Journal of Information science and technology, 2004, 7(1-2): 9-43.

[2] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[3] Vossen P. Euro WordNet: general document[J]. 2002.

[4] Zhang Y, Li B, Dai X, et al. PQAC-WN: constructing a WordNet for Pre-Qin ancient Chinese[J]. Language Resources and Evaluation, 2017, 51(2): 525-545.

[5] Li M, Zhou Z, Wang Y. Multi-Fusion Chinese WordNet (MCW): Compound of Machine Learning and Manual Correction[J]. arXiv preprint arXiv:2002.01761, 2020.

[6] Bizzoni Y, Boschetti F, Diakoff H, et al. The Making of Ancient Greek WordNet [C]//LREC. 2014, 2014: 1140-1147.

[7] Zhang Y, Li B, Wang X, et al. Mapping word senses of middle ancient Chinese to WordNet [C]//2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE, 2014, 1: 446-450.

[8] ZHANG L, LI J, HU M, et al. Implementation of Chinese WordNet [J]. Journal of Northeastern University (Natural Science) vol, 2003, 24.

[9] Huang C R. Sinica BOW: integrating bilingual WordNet and SUMO ontology[C]//International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003. IEEE, 2003: 825-826.
[10] Xu R, Gao Z, Pan Y, et al. An integrated approach for automatic construction of bilingual Chinese-English WordNet [C]//Asian Semantic Web Conference. Springer, Berlin, Heidelberg, 2008: 302-314.
[11] Huang C R, Hsieh S K, Hong J F, et al. Chinese WordNet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing[J]. Journal of Chinese Information Processing, 2010, 24(2): 14-23.
[12] Wang S, Bond F. Building the chinese open WordNet (cow): Starting from core synsets[C]//Proceedings of the 11th Workshop on Asian Language Resources. 2013: 10-18.
[13] Xu H, Chen S, Cai J, Cao L, Wan C, Li B. 2020. The Construction and Statistical Analysis of Pre-Qin Ancient Chinese WordNet [J]. International Journal of Knowledge and Language Processing, 2021, 11(3): 48-61.
[14] Minozzi S. 2017. Latin WordNet, una rete diconoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval. In Paolo Mastandrea, editor, Strumenti digitali e collaborativi per le Scienze dell'Antichita`, number 14 in Antichistica, pages 123–134.
[15] 王力. 汉语史稿.中册[M]. 科学出版社, 1958.
[16] 杨合鸣.诗经词典[M].湖北：辞书出版社,2012.
[17] 张颖杰.资源稀缺语言的词汇语义资源自动构建方法研究[D].南京大学,2017.
[18] 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, 2:39–45.
[19] 张俐,李晶皎,胡明涵,姚天顺.中文词网的研究及实现[J].东北大学学报,2003(04):327-329.
[20] 于江生,俞士汶.中文概念词典的结构[J].中文信息学报,2002(04):12-20+44.
[21] 姚天顺,张俐,高竹.词网综述[J].语言文字应用,2001(01):27-32.
[22] 吴思颖,吴扬扬.基于中文词网的中英文词语相似度计算[J].郑州大学学报(理学版),2010,42(02):66-69.
[23] 袁毓林.面向信息检索系统的语义资源规划[J].语言科学,2008(01):1-11.
[24] 郭锡良.先秦汉语名词、动词、形容词的发展[J].中国语文,2000(03):195-204+286.