场景化国际中文教学资源知识图谱的构建

杨浩[1]，辛晶[2]，朱珊仪[3]，饶高琦[4]，荀恩东[1]

(1.北京语言大学 语言资源高精尖创新中心 北京 海淀 100083；
2.北京外国语大学 中国语言文学学院 北京 海淀 100081；
3.山东师范大学 国际教育学院 山东 济南 250014；
4.北京语言大学 国际中文教育研究院 北京 海底 100083)

摘 要：近些年，为支持国际中文教学，学界构建了大量的知识库，但大多都是针对某一具体的资源对象，比如搭配库、例句库等，其孤立性问题较为突出。在万物智能的时代背景下，国际中文教学也面临着数智化转型的问题，其对语言教学资源提出了更高要求，必须构建细粒度的、各个资源对象相互关联的知识图谱。在教学场景下，又特别注重“因材施教”，因此在构建教学用知识图谱时必须考虑知识的来源和用处，即场景化。本研究利用 BCC 结构检索工具关联各个资源实体， 充分考虑知识的来源以及适用的场景，构建了场景化的国际中文教学知识图谱，并初步进行了国际中文智慧教学的工程实践。

关键词：教学资源；知识图谱；中文教学；场景化

中图分类号：TP3　　**文献标识码**：A　　**文章编号**：1671-6841(2024)00-0000-00

DOI：10.13705/j.issn.1671-6841.2024***

Construction of a Scenarioized Knowledge Graph of International Chinese Teaching Resources

Hao Yang[1]，Jing Xin[2]，Shanyi Zhu[3]，Gaoqi Rao[4]，Endong Xun[1]

(1.Beijing Advanced Innovation Center for language Resources，Beijing 100083，China；
2.The School of Chinese Language and Literature，Beijing Foreign Studies University，Beijing 100081，China；
3.The School of International Education，Shandong Normal University，Jinan 250014，China；
4.Research Institution of International Chinese Language Education，Beijing Language and Culture University，Beijing 100083，China)

Abstract: In recent years, to support international Chinese language teaching, the academic community has developed numerous knowledge bases. However, these are often targeted at specific resources, such as collocation databases or example sentence collections, resulting in significant issues of isolation. In the era of ubiquitous intelligence, international Chinese language teaching faces the challenge of transitioning to a digital and intelligent paradigm, which imposes higher demands on language teaching resources. It is essential to construct a fine-grained knowledge graph that interconnects various resource entities. In educational contexts, there is a particular emphasis on "teaching according to the student's aptitude," which necessitates that the construction of knowledge graphs for teaching purposes take into account the provenance and application of knowledge, thereby

收稿日期：2024-XX-XX
基金项目：中文意合图的表征与生成方法研究（62076038）；北京语言大学国际中文智慧教育工程阶段性成果
第一作者：杨浩(1999-)，男，硕士研究生，主要从事知识图谱、AI 可解释性研究，E-mail:howyoung80@163.com
通讯作者：荀恩东(1967-)，男，教授，主要从事国际中文智慧教学、自然语言处理等研究，E-mail: edxun@blcu.edu.cn

ensuring Scenarioized relevance. This study employs the BCC structural retrieval tool to link various resource entities, carefully considering the origin of knowledge and its applicable contexts. Consequently, a contextualized knowledge graph for international Chinese language teaching was constructed, and preliminary engineering practices were conducted to explore its application in intelligent international Chinese language education.
Keywords: teaching resources；knowledge graph；Chinese language teaching；scenarioized

1 引言

随着计算机技术与人工智能技术的发展，智能化实践开始在各领域进行实践。2008 年，IBM 在《智慧地球：下一代领导议程》提出“智慧地球”的概念，随即引发了全球性的“智慧革命”,具体到各个垂直领域，又衍生出“智慧城市”、“智慧医疗”、“智慧交通”、“智慧教育”等概念。智慧教育是一个系统性的工程，智慧教学是智慧教育的核心，承担着实现智慧教育的重要任务[1]。但关于智慧教学的定义，目前学界还没有统一的说法。具体到国际中文教学领域，荀恩东[2]认为国际中文智慧教学就是“资源＋技术”的赋能,用“精标互联”的教学资源和语言智能技术赋能国际中文教学。其中“精标互联”就是要建立丰富的、细粒度的、各个资源实体强关联的国际中文教学知识库。

以往为了更好地服务国际中文教学，学界建立了大量的教学资源知识库，比如例句库、搭配库等。这些知识库虽然也有丰富的属性信息，但归根结底还是围绕着着某一资源为主体对象，其他资源信息作为其属性而存在。比如例句库，是以例句为主体对象，例句中富含的语法、词汇信息是作为其属性而存在的，这不符合“精标互联”所表达的互联性。互联性应该是资源对象在知识库中处于同等地位，资源对象间通过其存在的客观关系而相互联接。知识图谱是一种结构化的语义网络，能够以图的形式描述客观世界中的事物以及事物间存在的关系[3]。以知识图谱的结构形式来组织国际中文教学资源能够与“精标互联”的思想完美契合。

通常来讲，知识图谱分为通用领域知识图谱和垂直领域知识图谱，前者追求知识的广度，后者追求知识的深度[4]。目前，学界和工业界构建的各种知识图谱，无论是通用领域还是垂着领域，都只关注知识的存在性，即只要是知识是确实存的，就录入知识图谱，对于知识的场景适用性研究欠缺。而国际中文教学是一个讲究“因材施教”的教学场景，其对知识图谱的知识适用性要求极高，在构建知识图谱时，必须考虑知识从哪里来以及能够运用到哪里去，即场景化[5]。因此，本研究要建立的图谱不是一种简单的、一味追求资源数量的图谱，而是一种适用性高的场景化国际中文教学资源知识图谱。

在生成式大语言模型大获成功的背景下，有研究认为大语言模型将逐步取代知识图谱成为知识表示和获取的方法[6]。本研究认为，在国际中文教学领域，大语言模型不会代替知识库而成为知识的来源。首先，在知识的表示和存储上来讲，知识图谱将知识以图的形式进行存储，是一种结构化的显性知识库，具有可解释性[7]。大语言模型是通过学习大量的语料，将学到的知识以参数的形式存在于神经网络模型的权重中，是一种隐性的知识存储方式，不具备可解释性。再者，大语言模型还存在着事实性不准确问题，这在对知识准确度要求较高的国际中文教学场景不太适用，知识图谱能够很好的弥补这一缺憾。最后，通过构建事实准确的知识图谱，为语言模型的提供高质量的数据，提升模型的准确性也有很重要的意义[8]。

2 相关工作

国际中文教育进入新发展时期，面临向智能化转型的问题。智慧教育的核心是智慧教学，智慧教学的内涵就包括通过智能技术更好地建设数字化教学资源，推进教学资源的供给侧改革。通过资源和技术赋能，以实现国际中文教学提质增效变得愈发重要[2]。所以，要完成国际中文教育的智能化转型，语言资源知识库建设和语言智能技术创新是必要条件。

在语言资源知识库建设方面，学术界已经开展了大量的研究。邢丹[10]构建了介词结构搭

配库，王诚文等[11]从大规模语料中抽取介动搭配，助力语言教学研究。邵田等[12]从大规模结构树库中抽取两个动词连用的情况,为语言学本体研究提供了分类参考。王贵荣等[13]从语言本体的角度出发，总结了动宾搭配的知识体系，从 BCC 语料库中抽取动宾搭配知识，形成动宾搭配知识库。王雨[14]以《国际中文教育中文水平等级标准》为难易度控制标准，构建了等级可查、难度可控、应用方便的国际中文教育词语搭配知识库。刘志超[15]利用计算机技术从大规模真实语料中抽取搭配来构建具有语义关系标注的搭配库，为自然语言处理提供重要的知识资源。王璐[16]构建了三层搭配语义知识库并尝试了知识库的一些应用。胡韧奋[17]提出了汉语搭配的四种性质和九个类型，在二语教材语料库和互联网语料库构建了两个不同领域的大规模搭配知识库。钱小飞[18]建成了由 31003 种搭配构成的二元实词搭配知识库，并构造了词语搭配的可视化网络，直观地展示了词汇集群和频率关系。以上是搭配资源库建设的研究现状，在例句库建设方面，胡韧奋[19]构建了一个规模约 12 万句的面向对外汉语教学的话题语料库。以上搭配库和例句库都是从真实语料中标注、抽取、清洗而来。在生成式 AI 蓬勃发展的时代，朱奕瑾[20]基于 ChatGPT 利用思维链推导的方式建构了共识价值标准例句库，为基于生成式大语言模型的资源建设提供了示范应用。

在语言智能技术的创新上，荀恩东[21]研制了基于大数据的“北京语言大学语料库中心”(BLCU Corpus Center,简称 BCC)语料库，BCC 是服务语言本体研究和语言应用研究的在线大数据系统。BCC 除了线上服务外，还提供了个性化语料库构建工具包。用户可以使用私有的语料，将语料进行加工后，定制个性化语料库。BCC 支持对按照多层次结构标注体系[22]进行了组块结构分析的语料进行字符、词、短语、属性和结构信息为一体的复杂查询[24]。

上文提到的各种语言资源知识库都是具体地针对某一资源对象，目前关于研究如何将各资源实体进行关系连接，构建领域知识图谱还没有被探讨过。无论是服务于语言教学的知识库，还是服务语言学本体研究的知识库大都是从大规模语料库中获取，这对于教学场景来讲存在着数据噪声大、知识适用性低等问题。陆泉[23]提出了场景化知识图谱概念及其构建方法。场景化知识图谱是描述知识场景属性的知识图谱，关注知识的获得路径和适用场景。基于此本研究从教材、HSK 考试真题、练习题等可控数据源搜集语料，对知识图谱中的知识获取路径和适用范围进行明确定义，以期构建一个数据噪音低、适用性强的国际中文教学知识图谱。

3 语料采集与语料库建设

3.1 数据获取

为了尽可能地减少数据噪声，提高数据质量。本文选取的语料一律来自可控数据源，具体包括对外汉语教材、汉语水平考试(简称 HSK)真题、教材课后习题、国际中文智慧教学平台融课件[2]、国际中文学习词典以及各大中文学习网站[1]。

3.1.1 例句数据

对于汉语教师而言，真实语料、实用而科学的例句是进行课前备课的重要内容。语料库及其索引技术可以提供大量真实语料和统计数据，是实现教师例句设计过程现代化的有力工具[26]。但从各大开放语料库检索系统或者互联网上获取的例句存在超纲词、句式复杂、句群数量巨大等问题[27]。在实际教学中对于例句的要求是严格的，本文从数据源头上进行考虑，在各大主流对外汉语教材中获取例句约 10 万句、HSK 考试真题获取例句约 2 万句、国际中文智慧教学平台融课件获取例句约 2.5 万句，国际中文学习词典获取例句约 1.4 万句。上文提到的“精标互联”思想，其中“精标”主要体现在对资源对象属性信息的丰富。本文将获取的例句进行分词、词性标注后进行例句标注。在标注之前已经对例句文本进行了预处理，得到了例句的分词信息、词汇等级分布，并依据文本中出现的最高等级词汇预标了句子的等级。所以标注过程主要进行三步操作：①**判断句子是否适合作为例句**。②**为句子标注话题信息**。③**复核句子等级**。在确定了标注任务后，根据工作的内容开发了标注平台，平台同时兼具标注和质量检验功能，标注界面如图 1 所示。

[1] http://www.cltguides.com/main.action

图 1 例句标注操作图

Figure.1 Example sentences annotation operating

在判断句子是否适合作为例句时，把句子通顺、句义明晰、目标词语语义典型、目标词含义表达充分、无语境依赖等特点的句子判断为适合，把句子过长或过短、不成句、句义不明、上下文依赖强、句内焦点过多、有歧义、带有修辞、太过专业、有古文表达等特点判断为不适合。以上标准只是参考准则，标注员在实际标注过程中可在参考该准则的同时，不拘泥于准则灵活变通。基于以上判断标准，将会过滤掉一大批不符合要求的句子。在被过滤掉的句子中，有一些稍加修改便可作为例句，因此在判断的同时，也要考虑该句子是否有可以被修改成例句的可能。例如，“***乐善好施、扶危济困、助人为乐的美好品德，在今天的中国更是深入人心。***”存在句子长度过长、词汇太过高级等问题，将其修改为“***助人为乐是一种美好的品德。***”便可作为例句。

在判断句子的话题时，坚持“确定的标注、不确定的不标注”原则。本文按照表 1 的二级话题分类法确定句子的话题。

表 1 句子话题分类表

Table.1 Detailed Table of Sentence Topic Classification

一级话题	**二级话题**
日常生活	家庭生活；校园生活；职场生活；饮食；交通出行；交往；天气；购物；体育运动；医疗健康；个人经济活动；环保；新闻报道
教育	教学活动；学习状况
职业工作	就业离职；日常办公；公关活动；客户洽谈
文化	国情民生；风俗传统；地理名胜；语言文字；建筑；历史；军事；法律
文学艺术	各类故事；各类艺术；作品介绍；名人名家介绍；概念艺术
体验感悟	生活随感；人生感悟；社会现象评论
自然	动物；植物；自然现象；环境气候；人与自然
科技	网络与信息技术；科学常识；科技发明；专业学科知识
经济	经济现象；经济管理；商业贸易
各类信息	个人信息；事件信息；物品信息；时间信息

该标注体系由 10 个一级话题和 50 余项二级话题组成，基本覆盖了国际中文教学常见话题。

在判断句子等级时，主要根据句子的元信息(来源、句子长度、词汇等级分布)，将句子分为 1-7 七个等级，无法确定等级的标为 0。通过来源信息，可以大致确定难度范围，比如句子出现在教材、HSK 真题的位置，再辅以句子长度信息、词汇的等级信息最终确定例句等级，大部分的例句应该是和它词汇的难度匹配的。需要注意的是，有的词汇难度等级高，但是比较常用，学习得早，可以适当降级。例如，“***抱歉***”一词是六级词汇，当它出现在“***实在抱歉，我把这件事忘了***”一句时。参考它的出处信息《汉语教程(第三版)》——L21 理发，以及词汇等级分布：***实在/2 抱歉/6，我/1 把/3 这/1 件/2 事/1 忘/1 了/1,3***，最终把例句等级确定为 3。例句数据资源局部图如图 2 所示。

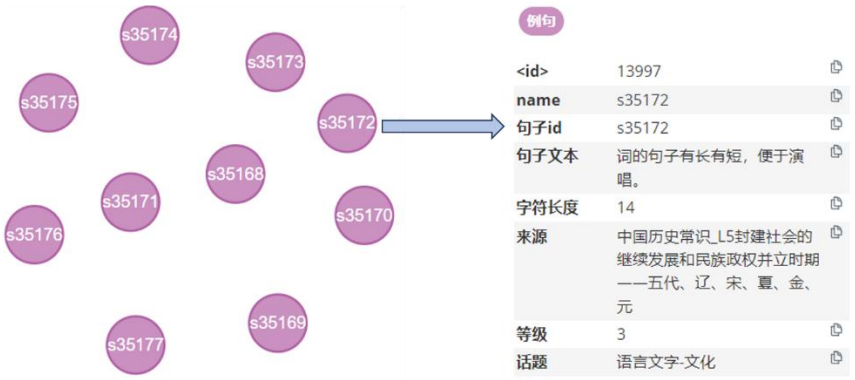

图 2 例句资源局部图

Figure.2 Partial map of example sentence resources

3.1.2 题目数据

题目中蕴含着关于知识点(主要是字、词)的自然标注信息、又拥有适用等级标签。但是真题数量有限，其囊括的知识点数量存在着很大的数据稀缺问题。为了使题目数据更加丰富，本文搜集了 2197 道课后习题和 20116 套国际中文智慧教学平台的融课件题目[25]。但这些题目都是无标签数据，对于适用的学习阶段没有明确标明。为了对无标签的试题进行定级，本文对 157 套 HSK 真题（一级 34 套、二级 25 套、三级 23 套、四级 26 套、五级 25 套、六级 24 套）进行等级词汇概率分布统计，在计算词汇等级分布概率时参考的是《新汉语水平(HSK)词汇大纲》的等级信息，计算结果如表 2 所示。

表 2 各等级试题词汇概率分布

Table.2 Vocabulary Probability Distribution for Different Levels of Test Questions

词汇 / 试题	HSK 1 级	HSK 2 级	HSK 3 级	HSK 4 级	HSK 5 级	HSK 6 级	超纲
一级试题	0.7301	0.0042	0.0008	0.0000	0.0000	0.0000	0.2649
二级试题	0.4209	0.2858	0.0099	0.0008	0.0008	0.0000	0.2817
三级试题	0.2917	0.1599	0.2650	0.0047	0.0008	0.0006	0.2773
四级试题	0.2160	0.1062	0.1352	0.2785	0.0070	0.0005	0.2567
五级试题	0.1603	0.0907	0.0941	0.0963	0.1214	0.0214	0.4158
六级试题	0.1264	0.0733	0.0805	0.0930	0.0891	0.0553	0.4825

纵向来看，对于各个等级的考试试题，其对应等级的词汇出现频率总是最高的。总体的

分布概率如此，那么每套试卷的词汇等级概率分布也应该与对应等级的词汇概率分布相近。相对熵(又称 KL 散度)可以描述两个概率分布之间的距离，相对熵越小，概率分布越相近。利用 KL 散度的算法思想来对试题进行定级，试题等级 L 评定过程可定义如下：

$$L = \arg\min_{i} \sum T(w) \log \frac{T(w)}{H_i(w)} \quad (1)$$

(1)式中，H_i表示i级的 HSK 试题等级词汇概率分布，T表示无标签试题的等级词汇分布，w为词汇等级变量。将搜集的题目数据进行等级评定后，各等级的题目数量占比如图 3 所示。

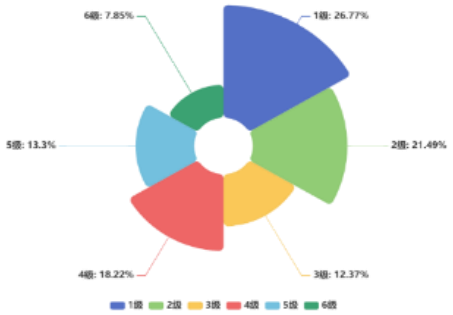

图 3 各等级题目占比图

Figure.3 Proportion of figures at different levels

3.1.3 其他数据采集

对于构建知识图谱的其他数据资源，例如汉字、词汇、语法点等信息主要来源于《国际中文教育中文水平等级标准》(简称《等级标准》)和国际中文教学指南网站。《等级标准》规定了二语学习者达到每一级中文水平应掌握的音节、汉字、词汇、语法的内容和数量。本文以《等级标准》提供的汉字表、词汇表、语法表作为汉字、词汇、语法对象的基本实例。《等级标准》只提供了关于汉字、词汇、语法的少量等信息，远不能满足教学需求。为此，需要借助其他来源的数据来丰富实体的属性信息。国际中文教学指南是中外语言交流合作中心建设的国际汉语教材研究和实用综合平台，网站提供了关于语言要素的丰富信息。利用网络爬虫，采集关于词汇的义项、中英文解释等信息，语法的分类、讲解、例句等信息。

-

3.2 构建工具介绍

在数据资源搜集完成之后，汉字、词汇、语法、例句等各资源对象都是孤立存在的。下一步就是要对各资源对象进行关联。为了方便资源对象的关联，本文选用 BCC 检索工具，对例句和题目数据构建索引。BCC 语料库可支持的语料类型包括：生语料、分词词性标注语料、句法结构树语料。本文将语料按照多层次结构标注体系[22]进行句法结构标注。例如句子“***金玉良缘是指林黛玉与薛宝钗，钗黛才是真爱!***”经过句法结构标注后，其句法结构树如图 4 所示。

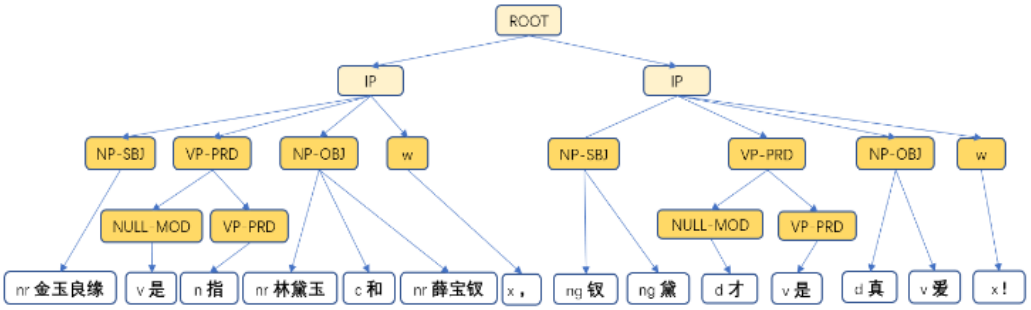

图 4 句法结构树示例图

Figure.4 Example of a syntactic structure tree

经过标注后的数据具有词性标记、短语功能标记、组块结构标记,利用 BCC 检索式[2]能对其进行一系列的复杂查询，包括字符级、词级、短语级、句法属性、句法结构关系等。例如，将语法点“***动词+得+形容词性词语***”编写成检索式“***v 得 a***”，可以检索到“***她笑得真开心***”、“***今天要过得开心哟***”、“***你球踢得真好***”等结果。

[2] 具体可参考 http://bcc.blcu.edu.cn/help

4　图谱构建

经过上述的工作，现已整理了约 2 万套带有等级标签的习题，借助 BCC 检索工具建立了来源于教材、HSK 真题、练习题的国际中文教学例句库约 15 万句，搜集了来源于《等级标准》的分等级词汇表、语法点表并且利用网络数据采集技术丰富了词汇表、语法点表的属性信息。在数据准备阶段完成后，后续的工作就是要对上述的数据进行充分的挖掘，然后对得到的各资源实体(主要是词汇、语法、例句)进行关联计算，建立国际中文教学资源知识图谱。图谱构建的大致流程如图 5 所示。

知识图谱中的知识包括实体的属性知识和实体间的关联知识两大类。知识图谱的知识获取也是围绕这两类数据开展。前文已经构建了例句数据集、题库的语料库索引。下文先从各语言要素实体的属性信息和内部关联进行阐述，然后再说明各不同资源实体的关联方法。

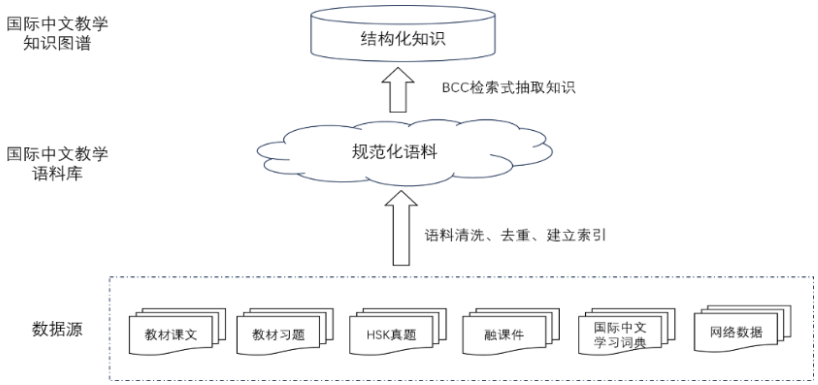

图 5　知识图谱构建流程图

Figure.5 Process for building an international Chinese language teaching knowledge graph

4.1 语言要素知识获取

4.1.1　字

字的全部实体为《等级标准》提供的 3000 个汉字。《等级标准》只提供了汉字的等级标签。为了对丰富汉字的属性信息，给真实教学场景提供有价值的数据，本文借助网络爬虫技术采集了汉字的拼音、笔顺、笔画数、部首、部件、中英解释、书写动画等信息，统计了汉字在国际中文教学语料库中的频次。除了汉字的属性信息，本文对汉字间的关系主要做了以下关联：同音、同笔画数、同部首、同结构、形近。知识图谱汉字的局部图(以“***爱***”字为例)如图 6 所示。

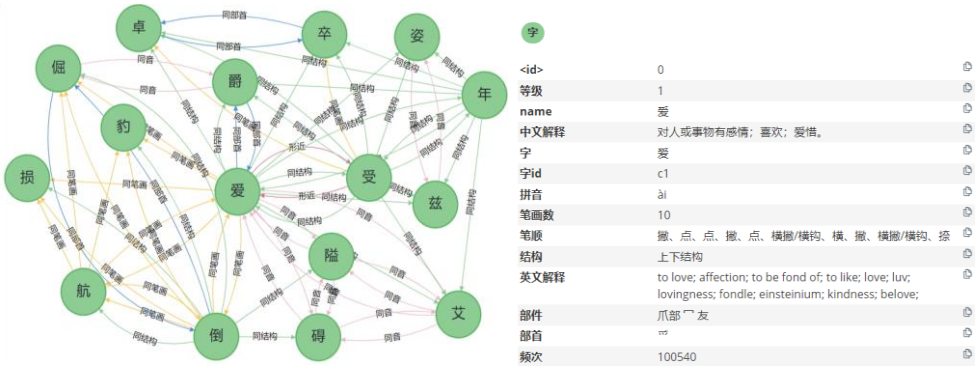

图 6　知识图谱汉字局部图(以“爱”字为例)

Figure.6 Partial knowledge graph of Chinese characters (taking the character ‘爱’ as an example)

4.1.2 词汇

词汇的全部实体为《等级标准》提供的11092个词汇。《等级标准》提供了词汇的拼音和等级属性信息。通过网络数据采集，爬取了词汇的中英文解释。在国际中文教学语料库中统计其出现的频次。词汇学习在语言学习中占据重要地位，词汇的学习要注重词汇间的关联，特别是词汇的搭配、共现。鉴于此，本文建立的知识图谱中词汇间的关系有：搭配、共现、同义、反义等。其中同义词和反义词来自网络数据采集，搭配、共现关系来源于语料库抽取。

关于词汇搭配库的构建，具体参照《服务国际中文教育的词语搭配知识库建设》[14]，该工作与本文都是为建设国际中文智慧教学平台而服务，属于同一个工作框架，本文不再对搭配库的建设进行赘述。

关于共现词汇，共现词汇的抽取是依据《等级标准》提供的词表，在建立的国际中文教学语料库中用脚本程序统计词汇的共现。主要进行以下几步操作：①采用 Jieba 分词工具对语料进行分词处理。②去除掉语料中的停用词。③统计词汇的共现频率。④去除掉与搭配库重叠的部分。⑤取共现频率排名前 10 位的共现词汇。抽取结果显示，在大规模语料库中进行词汇的共现关系抽取会与词汇的搭配数据有较高程度的重叠。抽取共现关系的初衷是找出跟词汇具有混淆、同属、同话题等关系的词汇。习题语料可以很好地满足这一要求，因为题目中经常出现一个词汇的近义、辨析项、同属等关系的词。因此，在抽取词汇的共现关系时，本文将语料库的范围缩小至习题语料。通过这一方法，能够较高质量地得到具有共现关系的词汇，图 7 以词汇“*茶*”、“*大学生*”为例，展示抽取结果。

通过以上工作形成的词汇知识图谱局部图如图 8 所示(“*幸运*”一词的知识图谱局部图)。

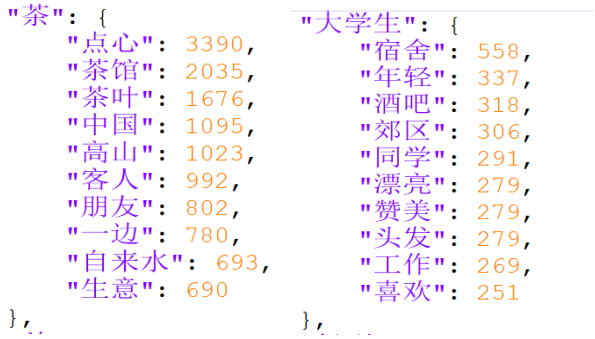

图 7 词汇共现示例图(以“茶”和“大学生”为例)

Figure.7 example of vocabulary co-occurrence (with ‘茶’ and ‘大学生’ as examples)

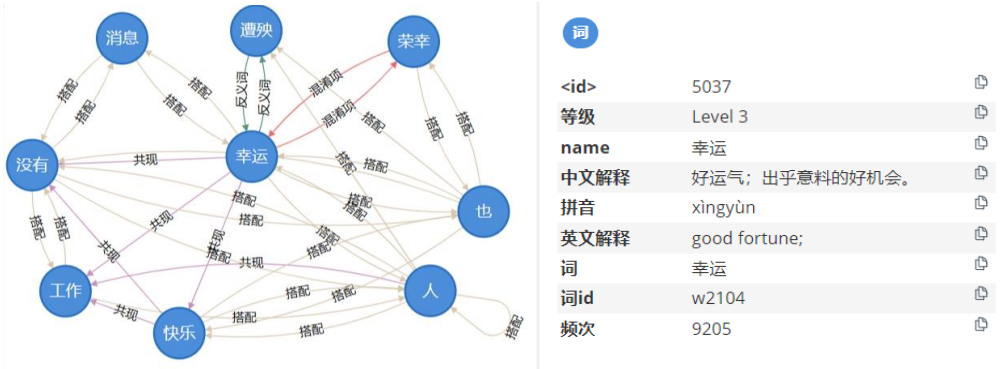

图 8 知识图谱词汇局部图(以“幸运”字为例)

Figure.8 Partial vocabulary knowledge graph (taking the term ‘幸运’ as an example)

4.1.3 语法

《等级标准》一共给出了 572 个语法点的等级、解释和用例。语法的学习，重点是在例句中体会语法的用法。因此本文关于语法的知识图谱重点在语法点和例句的关联，将在后文介绍。

4.2 场景属性信息

知识图谱的场景信息是用来描述知识从何而来，能用到哪里去的属性信息。前者描述知识来源、后者描述知识适用范围。本文构建的国际中文教学资源知识图谱，主要对例句和题目进行场景属性的标明。其中，例句的来源主要为对外汉语教材、国际中文学习词典、HSK 真题、融课件，深入内容体现为例句的话题属性。题目来源为 HSK 真题和融课件，深入内容体现为题目的题型、知识点(主要是词汇)的共现。适用范围属性为例句等级、题目所属套题的等级。

以上是对资源知识图谱场景信息的描述，其实对于教学场景而言，单单考虑资源的场景属性意义不大。实现智慧化的教学，必须结合学习者的行为数据，例如学习者年龄、国别、汉语水平、偏误信息等。只有将资源知识图谱和学习者基本信息和学习行为结合起来才能实现真正的场景化。本文探讨的是资源知识图谱的构建，对于学习者端知识图谱的建设本文不做探讨。

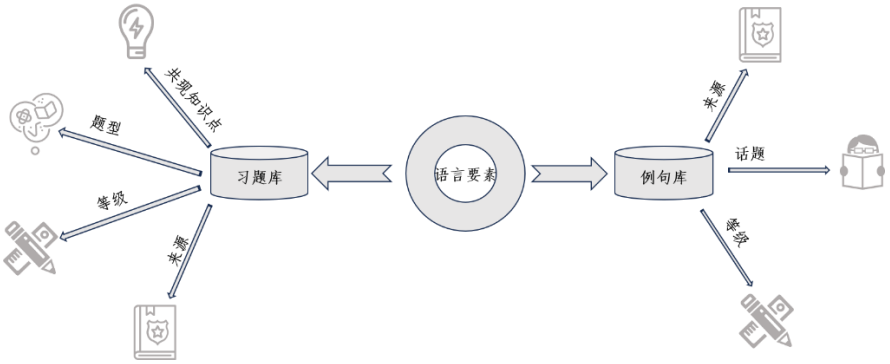

图 9 国际中文教学知识图谱场景信息属性图

Figure.9 Scene information attribute diagram of an international Chinese teaching knowledge graph

4.3 资源实体间的关联

综上所述，本文的教学资源实体共有以下几种：汉字、词汇、语法、例句、题目。前文已经对各资源实体的属性知识获取和各资源实体内部的关联关系进行了介绍，所以下面的工作就是对各资源实体进行关联计算。

4.3.1 题目的自然标注信息

首先关于题目，题目中已经存在天然的标注信息，比如选择题的空缺项、连线题的待连项、连词成句的词表等，这其实已经初步建立了字、词汇和题目之间的关联。

4.3.2 字与词汇和题目的关联

汉字与词汇、题目是简单的被包含与包含的关系。汉字和词汇可以通过简单的字符匹配建立联系。汉字和题目建立联系时，只与单独考察汉字的题目建立关联，例如根据拼音写汉字、汉字知识卡片等题目。

4.3.3 词汇与例句和题目的关联

词汇与例句的关联是一个非常复杂的工程，这涉及到词的不同读音、不同词性导致的多义项，即使是同一词性，也有不同的解释等因素。在进行词与例句的关联时，不能简单地通过字符串包含操作来确定词与例句的关系，这样太过粗糙。《国际中文学习词典》比较细粒度地给出了关于词的不同词性以及解释，本文参照其编写 BCC 检索式模板，使用脚本程序自动完成 BCC 检索式的编写，最后通过 BCC 检索式的主动检索建立词汇与例句的关联。图 10 以“***根本***”一词为例，给出其形成 BCC 检索式的过程。

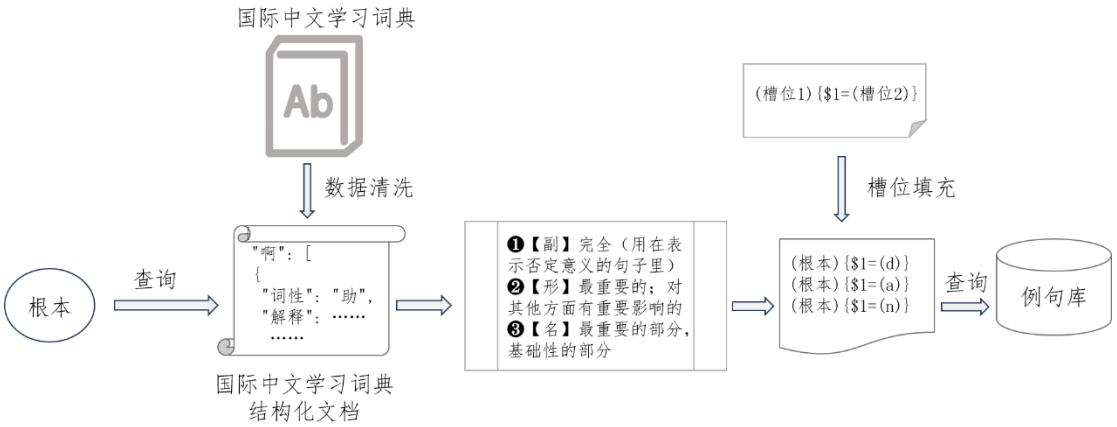

图 10 词汇检索例句流程图

Figure.10 Vocabulary retrieval example sentences flowchart

4.3.4 语法与例句和题目的关联

前文已经用 BCC 提供的个性化语料库建设工具对题目和例句数据集建立了索引形成了可供结构检索的例句库和习题库，支持使用 BCC 检索式对其进行检索。本文通过将各语法点映射成 BCC 检索式的方式，建立语法点与词汇和题目的关联。以检索关联的方式建立语法点知识图谱，工作的重点在于把语法点编写成检索式。《等级标准》附录将语法按照语素、词类、短语、固定格式、句子成分、句子的类型、动作的态、特殊表达法、强调的方法、提问的方法、口语格式、句群分为 12 大类语法项目，具体为 572 个语法点。本文进一步将 572 个语法点进一步拆分为 1149 个语法单元，按照每个语法单元进行 BCC 检索式的编写。语法单元在数据库中存储的逻辑结构如表 3 所示。

表 3 语法信息表

Table.3 Grammar information table

语法项目	语法点	语法单元	示例	等级	BCC 检索式
词类	名词（方位名词）	上	桌子上	1	n (f){$1=[上]}
词类	名词（方位名词）	下	树下	1	n (f){$1=[下]}
……					
句子成分	主语	名词作主语	衣服很好看	1	w n v w、w n v n w ……
句子成分	主语	代词作主语	他在看电视	1	n r w、w r v n w ……
……					
口语格式		该……了	十一点了，该睡觉了。	2	该*了 w
……					

通过以上步骤，本文完成了现有教学资源知识图谱的构建，国际中文教学资源知识图谱的局部图如图 11 所示。

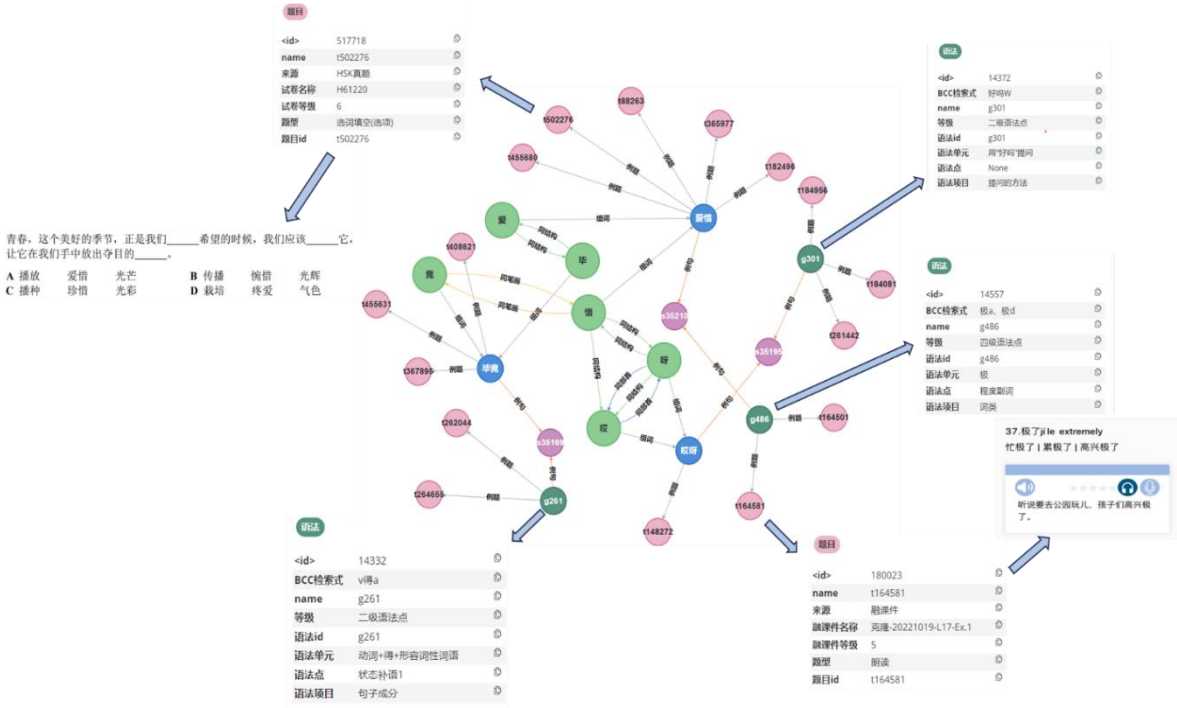

图 11 场景化国际中文教学资源知识图谱局部图

Figure.11 contextualized partial knowledge graph of international Chinese teaching resources.

5 图谱应用

本文将构建的知识图谱存储于开源图数据库 neo4j 中，用户可通过 Cypher 查询语言对知识图谱进行检索，该检索支持属性查找、关系查找等一系列复杂查询。目前该知识图谱已服务于北京语言大学国际中文智慧教学平台[3]，支持平台的基于语言要素的自动出题(如图 12)和题目检索。

一般来讲，自动出题可分为两大类，一是基于生成式模型的完全自动出题，二是基于内容检索的半自动化式自动出题。前者虽然题目类型丰富多样，但在题目内容上不能够达到完全可控，其在多媒体前端展示上由于其样式多变而落地困难。而基于内容检索的半自动化式自动出题，通过事先固定题目的样式模板，把重点放在题目的内容而非形式上，不但确保了题目质量而且较容易进行工程落地。国际中文智慧教学平台自动出题功能，通过事先固定题目的形式，以构建的国际中文教学知识图谱为内容支撑，能够高质量地完成基于语言要素的自动化出题。除此之外，知识图谱还可为教师课前备课、教学资源制作提供数据支持，通过和学习者行为数据的融合，实现个性化学习资源的推荐。

图 12 自动出题结果图

Figure.12 automated question generation result diagram

[3] https://classtest.blcu.edu.cn/

6 总结展望

本文以《国际中文教育中文水平等级标准》为引领，构建了含有汉字、词汇、例句、题目实体的国际中文教学资源知识图谱。该知识图谱不同于以往孤立的各教学资源库，真正实现了各不同类型资源的相互联通。图谱的数据来源于教材、题目、标准大纲等可控数据源，大大降低了数据噪声。国际中文教学资源知识图谱的构建一开始就是为服务国际中文智慧教学工程为出发点的，也有别于其它停留在理论设想层面的知识图谱。此外，在生成式大语言模型发展势头迅猛的背景下，各领域也开始构建面向垂直领域的大语言模型，国际中文教学资源知识图谱可以为构建国际中文教学大模型提供数据。但也应该意识到，国际中文教学知识图谱的构建是一个极其复杂、极其细致的工程。特别是词汇网络的构建，学界已经对此进行了大量的探讨，但落地实践鲜有，本文构建的知识图谱目前关于词汇只有搭配、共现、同义、反义等关系，要实现词汇的智能化教学、个性化学习，必须要构建细粒度的词汇先后序关系，因此应该继续丰富知识图谱的词汇网络关系。国际中文教学资源知识图谱的构建过程也有别于其他垂直领域，由于其服务于世界各国母语为非汉语的学生，需要确保知识图谱中的数据在民族习惯、价值观等方面不能有任何偏向，语言表达上也不能有语法上、词汇上的错误，因此需要大量的人工校对，耗时耗力。这也启发我们如何从工程角度出发，建立语言资源从数据采集到标注加工再到核准校对的工程化流程，借助人工智能算法进行先期过滤，以减少人力物力的付出。

参考文献：

[1] 王明丹,吴金航.智慧教学视域下的教师教学素养：意涵、结构与发展[J].教育科学论坛,2022(16):62-65.

[2] 荀恩东. 融课件：国际中文教育资源与技术的集成创新[J].语言教学与研究,2023(05):9-12.

[3] 王昊奋, 漆桂林, 陈华钧. 知识图谱: 方法, 实践与应用[M]. BEIJING BOOK CO. INC., 2019.

[4] 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述[J]. 计算机系统应用, 2019, 28(6): 1-12.

[5] 陆泉,陈静宇,陈帅朴等.场景化知识图谱及构建方法[J/OL]. 情报科学:1-19[2023-10-19].http://kns.cnki.net/kcms/detail/22.1264.G2.20230915.1111.011.html

[6] 车万翔,窦志成,冯岩松等.大模型时代的自然语言处理:挑战、机遇与发展[J].中国科学:信息科学,2023,53(09):1645-1687.

[7] Pan S, Luo L, Wang Y, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap[J]. arXiv preprint arXiv:2306.08302, 2023.

[8] Yang L, Chen H, Li Z, et al. ChatGPT is not Enough: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling[J]. arXiv preprint arXiv:2306.11489, 2023.

[9] 魏晖,吴应辉,苏向丽等.“国际中文教育集成创新”大家谈[J].语言教学与研究,2023(05):1-12.

[10] 邢丹,饶高琦,荀恩东等.基于大规模语料库的介词结构搭配库构建[J].中文信息学报,2020,34(11):1-8.

[11] 王诚文,饶高琦,荀恩东.基于结构检索的汉语介动搭配知识库构建[J].中文信息学报,2023,37(07):23-31.

[12] 邵田,翟世权,饶高琦等.基于结构树库的状位动词语义分类及搭配库构建[J].中文信息学报,2023,37(06):44-51+66.

[13] 王贵荣,饶高琦,荀恩东.基于大规模语料库的现代汉语动宾搭配知识库构建[J].中文信息学报,2021,35(01):34-42+53.

[14] 王雨,肖叶,荀恩东等.服务国际中文教育的词语搭配知识库建设[J].语言文字应用,2022(02):26-37.DOI:10.16499/j.cnki.1003-5397.2022.02.009.

[15] 刘志超. 汉语动宾搭配库构建技术研究[D].沈阳航空航天大学,2012.

[16] 王璐. 基于多知识源的二元搭配语义知识库的构建及应用[D].北京信息科技大学,2015.

[17] 胡韧奋,肖航.面向二语教学的汉语搭配知识库构建及其应用研究[J].语言文字应用,2019(01):135-144.DOI:10.16499/j.cnki.1003-5397.2019.01.017.

[18] 钱小飞.面向汉语国际教育的实词搭配知识库建设[J].语言文字应用,2020(04):132-142.DOI:10.16499/j.cnki.1003-5397.2020.04.020.

[19] 胡韧奋,朱琦,杨丽姣.对外汉语教学领域话题语料库的研究与构建[J].中文信息学报,2015,29(06):62-68.

[20] 朱奕瑾,饶高琦.基于 ChatGPT 的生成式共同价值标准例句库建设[J].云南师范大学学报(对外

汉语教学与研究版),2023,21(03):71-80.DOI:10.16802/j.cnki.ynsddw.2023.03.016.
[21] 荀恩东,饶高琦,肖晓悦等.大数据背景下 BCC 语料库的研制[J].语料库语言学,2016,3(01):93-109+118.
[22] 卢 露，矫红岩，李 梦，荀恩东 . 基于篇章的汉语句法结构树库构建 [J/OL]. 自动化学报，2020，（2）.
[23] 陆泉,陈静宇,陈帅朴等.场景化知识图谱及构建方法[J/OL].情报科学:1-19[2023-10-20].http://kns.cnki.net/kcms/detail/22.1264.G2.20230915.1111.011.html.
[24] 荀恩东.自然语言结构计算：BCC 语料库[M].人民邮电出版社,2023.
[25] 北京语言大学“国际中文智慧教学系统 2.0 版”发布[J].语言教学与研究,2023(02):113.
[26] 蔡建永. 基于语料库索引的对外汉语教学课前例句设计[C]//中文教学现代化学会.数字化对外汉语教学实践与反思.清华大学出版社,2010:431-436.
[27] 单天罡.基于语料库的对外汉语词汇例句收集研究[J].现代语文(语言研究版),2013(09):104-106.
[28] 宋柔,李斌,王宝鑫,杨子清,伍大勇,李辰,荀恩东,苏祺.“语言智能”多人谈[J].语言战略研究,2023,8(04):53-56.