面向国际中文教育的知识图谱数据集构建方法与应用

王韩[1]，王千予[1]，兰韵诗[1]，王晔[1]，梁远远[2]，丁安琪[2]
1. 华东师范大学数据科学与工程学院，上海 200062；
2. 华东师范大学国际汉语文化学院，上海 200062

摘要

针对国际中文教育资源分散、标准体系并存、题目与知识点关联不显式、偏误与文化资源难以复用等问题，构建了面向国际中文教育的知识图谱数据集CFL-KG。通过整合等级标准、汉字、词汇、语法、题目、偏误、文化、成语和多媒体等资源，并经过数据清洗、字段规范、实体对齐、关系标注和质量校验，形成了可计算、可追溯、可持续维护的结构化数据。当前数据快照包含389，111个节点和1，017，421条关系，覆盖多类教学对象及其关联。该数据集可支持跨标准备课、测评资源追溯、偏误诊断、文化解释和智能问答等应用，为国际中文教育的数据建设和智能化服务提供基础支撑。

关键词

国际中文教育；知识图谱；数据集构建；多源融合；教学应用

中图分类号：TP391　　文献标志码：A

Construction method and application of a knowledge graph dataset for international Chinese language education

Wang Han[1], Wang Qianyu[1], Lan Yunshi[1], Wang Ye[1], Liang Yuanyuan[2], Ding Anqi[2]
1. School of Data Science and Engineering, East China Normal University, Shanghai 200062, China;
2. International College of Chinese Studies, East China Normal University, Shanghai 200062, China

Abstract

International Chinese language education relies on heterogeneous resources such as proficiency standards, characters, words, grammar points, assessment items, learner errors, cultural knowledge and multimedia materials. However, these resources are often fragmented, weakly aligned and difficult to reuse in teaching applications. CFL-KG, a knowledge graph dataset for international Chinese language education, was constructed. Through data cleaning, field normalization, entity alignment, relation annotation and quality checking, multi-source resources were organized into computable and traceable graph data. The current snapshot contains 389,111 nodes and 1,017,421 relations. The dataset supports cross-standard lesson preparation, assessment-resource tracing, learner-error diagnosis, cultural explanation and intelligent question answering, providing a reusable data foundation for intelligent Chinese language education.

Key words
international Chinese language education, knowledge graph, dataset construction, multi-source fusion, teaching application

0 引言

数据集质量正在成为人工智能应用效果的决定性因素之一。在教育领域，数据不仅是模型训练或系统检索的输入，更承载着课程标准、知识组织、测评解释和教学干预等专业语义。对于国际中文教育而言，学习对象包括汉字、词汇、语法、文化和交际知识，教学过程同时受到《国际中文教育中文水平等级标准》、HSK 2.0、HSK 3.0以及各类教材体系的约束，测评资源又与题型、评分标准和知识点覆盖密切相关。因此，该领域的数据集构建并不是简单地汇总文本、表格或题库，而是要将多源异构资源转化为可计算、可追溯、可更新、可复用的领域数据资产[1-5]。

现有国际中文教育资源具有明显的分散性和异构性。一方面，汉字、词汇、语法、成语、文化知识、等级标准、题目资源、偏误记录和语料证据分布在不同机构、教材、题库、平台和人工整理文档中；另一方面，不同资源在字段命名、粒度层级、等级体系和使用场景上存在差异。例如，同一词语可能同时出现在不同等级标准和词典资源中，其释义、例句、拼音或词性字段存在不一致；同一题目可能隐含考查多个知识点，但题库中往往只保存题干、选项和答案；学习者偏误记录包含偏误类型、偏误原因和高频群体等教学价值信息，却难以与具体字词语法自动关联[1-2,6-7]。

上述问题直接影响数据驱动的教学应用。教师在备课时需要回答“某一标准某一级别应覆盖哪些字词语法和文化点”；命题或测评分析时需要追溯“某道题究竟考查哪些知识点、覆盖是否均衡”；偏误干预时需要定位“某个语法点常见错误是什么、应选取哪些例句和练习进行干预”；智能问答或大模型应用还要求回答可解释、证据可追溯。如果缺少结构化关系和统一数据模式，这些任务只能依赖人工检索和经验判断，难以形成稳定的数据服务能力。

针对上述需求，本文构建面向国际中文教育的知识图谱数据集CFL-KG。与一般语言资源库相比，CFL-KG不是仅保存字词条目或文本片段，而是围绕国际中文教育的专业任务，将语言知识、等级标准、测评资源、偏误证据和文化内容统一建模为图结构数据。本文重点讨论该数据集的构建方法、质量控制和典型应用，而非单纯介绍系统功能。

1 数据集需求与相关研究

1.1 知识图谱数据集建设需求

国际中文教育有鲜明的领域性。首先，知识对象有多层次结构，既包括汉字的拼音、笔画、部首、结构和多媒体书写资源，也包括词汇的词性、释义、搭配、例句、近义词和反义词，还包括语法点的句式结构、用法规则、例句和等级标签。其次，资源使用受到多套标准共同约束，同一教学内容可能需要在HSK 2.0、HSK 3.0和《国际中文教育中文水平等级标准》之间进行映射。再次，教育场景具有任务导向性，数据不仅要可检索，还要能服务备课、命题、评测解释、偏误诊断和文化说明[1-2]。

因此，面向该领域的数据集应满足四类要求：一是覆盖要求，即覆盖核心语言知识、等级标准、题目、文化、偏误和语料证据；二是对齐要求，即将不同来源、不同标准和不同粒度的数据组织到统一模型中；三是关系要求，即不仅保存实体属性，还要显式表达知识点之间、知识点与题目之间、知识点与偏误之间、语言知识与文化资源之间的关系；四是应用要求，即数据集能够通过查询、统计、可视化和接口服务进入真实教学工作流。

1.2 相关工作

本文的构建思路借鉴了四类前驱工作。第一，知识图谱与本体工程研究强调，领域数据的可复用性依赖概念层、关系层和约束层的明确建模。Gruber提出本体是共享概念化的显式规范，后续OWL等语义网技术进一步区分了类、实例、对象属性、数据属性和约束规则[3,8-10]。因此，本文不把CFL-KG仅视为Neo4j中的节点边集合，而是先定义领域本体，再以图数据库实现存储和查询。

第二，多源数据融合研究关注异构来源中的实体解析、字段映射、冲突处理和来源可信度问题[11-12]。这对CFL-KG尤其重要，因为国际中文教育中的词汇、语法、题目、文化和偏误数据来自不同标准、教材、题库和人工标注资源，若没有统一的对齐规则，会导致同名实体重复、等级标签冲突和关系方向不一致。

第三，教育知识图谱研究表明，图结构可以组织课程知识、教学资源和评测对象，并支持检索、推荐和精准教学[4-5]。但是，现有研究多聚焦通用教育或单课程场景，对国际中文教育中的多标准并存、语言单位类型复杂、偏误证据和文化资源融合讨论不足。

第四，专家审核与应用反馈机制来源于人机协同标注、弱监督数据构建和教育测评证据化思想[13-15]。题目–知识点、偏误–知识点和文化关联等关系具有明显的领域解释性，不能完全依赖自动抽取。因此，本文采用“自动生成候选—专家复核—应用日志反馈—版本更新”的闭环方式，保证高价值关系的可解释性和可维护性。

近期《大数据》相关研究也关注知识图谱关键技术综述与教育知识图谱结构分析问题，为本文的数据模式设计和图结构质量评估提供了参考[16-17]。

1.3 构建流程

本文采用“领域本体约束+多源数据融合+专家审核+应用反馈”的方式构建CFL-KG。首先，依据国际中文教育知识体系和教学任务设计数据模式，将实体类型划分为语言知识、评估资源、偏误证据、文化拓展和标准等级等类别；其次，面向不同数据源设计字段映射和清洗规则，将表格、文档、词典、题库和多媒体资源转换为统一格式；再次，通过规则抽取、词表匹配、人工标注和模型辅助等方式构建关系；最后，结合Schema校验、冲突检测、分层抽样复核和应用日志反馈持续更新数据。总体流程如图1所示[3-5,14]。

2 数据来源与数据内容

2.1 多源异构数据来源

CFL-KG整合的数据来源可分为标准类资源、语言基础资源、测评资源、偏误资源、文化资源和语义关系资源。表1概括了各类数据的资源对象、核心字段以及

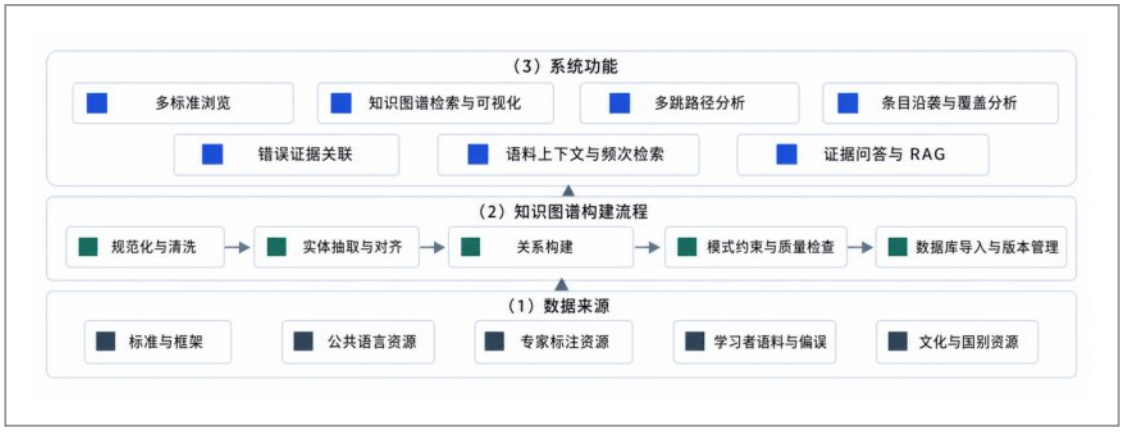


图1 CFL-KG数据集构建流程

在图谱中的对应对象。可以看出，该数据集的建设对象不仅包括字词语法等基础语言材料，还包括题目、偏误、文化、多媒体和关系型数据，因此需要统一的数据模式和质量控制流程。

表1 CFL-KG主要数据来源与内容

数据类别	资源对象	核心字段或内容	对应图谱对象
语言基础数据	汉字、词汇、语法、部首、音节/拼音	拼音、笔画、部首、释义、词性、短语、例句、发音、书写动画等	Character、Word、Grammar、Radical、Pinyin
等级标准数据	HSK 2.0、HSK 3.0、国际中文教育等级标准	等级、能力描述、音节数、词汇量、语法点数量、标准来源	HSKLevel、HSK30Level、InternationalLevel
测评资源数据	题型模板、题目实例、阅读材料	题干、选项、答案、解析、题型、难度、考查知识点	Question、QuestionTemplate及题目考查关系
偏误证据数据	发音偏误、字形偏误、词汇偏误、语法偏误	偏误类型、偏误内容、高频群体、偏误原因、纠正建议	Error及偏误关联关系
文化拓展数据	文化点、成语、典故、国情知识、多媒体	文化介绍、来源、例句、文化等级、典故、褒贬义、图片/视频路径	Cultural、Idiom、CulturalStage
语义关系数据	近义、反义、共现、组成、学习依赖、文化关联	关系端点、关系类型、来源、关系等级或置信标记	SYNONYM、ANTONYM、CO_OCCURS_WITH等

2.2 数据内容组织

从数据内容看，CFL-KG具有“属性数据+关系数据+证据数据”的复合结构。属性数据主要描述实体自身特征，如汉字的笔画和部首、词语的拼音和释义、语法点的类型和例句、文化点的介绍和来源。关系数据描述实体之间的教育语义联系，如词语与等级标准的来源关系、题目与语法点的考查关系、词语之间的近反义关系、

汉字之间的学习依赖关系、词语与文化点的文化关联关系。证据数据用于支持可解释应用，包括题目文本、例句、语料库上下文、偏误原因和多媒体路径等。

这种组织方式使数据集不再停留于“资源清单”层面，而是能够支持面向关系的查询和推理。例如，教师可以从某一等级出发检索应掌握的字词语法，再沿着关系找到相关题目、常见偏误和文化解释；研究者可以统计不同等级中题目覆盖的知识点分布，分析某类偏误与特定语法点或学习者群体之间的关系；智能系统可以在生成回答时引用图谱路径和语料证据，从而降低无依据生成的风险。

3 知识图谱数据模式与本体设计

3.1 领域本体模型

本文将数据模式明确区分为本体概念层和图谱实例层。前者规定国际中文教育对象的概念边界、上下位层级、属性范围和跨类语义关系；后者是在图数据库中以节点、边和属性形式存储的具体数据。借鉴本体工程和语义网中的术语体系[3,8-10]，本文将领域本体定义为式（1）：

$$O=(C,H_C,R,H_R,A,D,Cons) \quad (1)$$

式中，C 表示概念类集合，H_C 表示概念类之间的上下位层级关系，R 表示对象关系集合，H_R 表示关系层级或关系归并，A 表示数据属性集合，D 表示属性取值域，Cons 表示端点类型、方向、基数、必填字段等约束。知识图谱实例可表示为：

$$G=(V,E,P,\tau,\rho) \quad (2)$$

式中，V 表示实体节点集合，E 表示关系边集合，P 表示属性集合，τ:V→C 将每个节点映射到本体概念类，ρ:E→R 将每条边映射到本体关系类型。

在 CFL-KG 中，概念类 C 首先划分为五个上位概念：语言知识类、评估资源类、偏误证据类、文化拓展类和标准等级类。每个上位概念进一步划分为若干子类型；这些子类型在本体层面是概念子类，在图数据库落库时表现为节点标签或节点类型。需要区分的是，本体层的“子类型”回答“某类对象由哪些更具体对象构成”，而跨类语义关系回答“不同类型对象如何发生教育语义联系”。表 2 给出了 CFL-KG 的核心概念体系、子类型及其类型间关系。

3.2 分层数据模型及设计依据

CFL-KG 采用面向教学任务的分层数据模型。语言知识层解决“学什么”的问题，评估资源层解决“怎么考”的问题，偏误证据层解决“错在哪里、如何干预”的问题，文化拓展层解决“如何解释和拓展”的问题，标准等级层作为枢纽连接不同标准体系。分层依据主要来自三个方面：第一，国际中文教育等级标准和 HSK 规格本身以字、词、语法、能力描述和等级要求组织教学目标；第二，真实教学工作流通常围绕备课、测评、偏误反馈和文化说明展开；第三，国际中文教育常用资源（如教材、题库、学习者语料和文化材料）本身呈现出异构但相对稳定的信息结构，可自然映射为语言知识、评估资源、偏误证据和文化拓展等核心对象类型。由此形成的 Schema 既符合领域对象边界，也便于将教学任务转化为可查询的图路径。

3.3 关系类型与约束

关系类型是该数据集区别于一般资源库的关键。形式上，任一关系 r∈R 均被定

表2 CFL-KG本体核心概念体系

上位概念（本体层）	子类（概念层）	类型内部关系	与其他上位概念的关系
语言知识类	汉字、词汇、语法点、成语、部首、音节/拼音	词汇由汉字组成（COMPOSED_OF）；词语之间可形成近义、反义、共现关系（SYNONYM/ANTONYM、CO_OCCURS_WITH）；语法点之间可具有前驱依赖（DEPENDS_ON）	通过FROM_LEVEL对齐到标准等级；通过ASSESSES/EXAMPLE_OF被评估资源考查；通过ERROR_RELATED_TO被偏误证据指向；通过CULTURAL_RELEVANCE与文化拓展类形成文化关联
评估资源类	题型模板、题目实例、阅读材料、样卷	题型模板、样卷和阅读材料组织为评估资源属性；题目与考点的图谱关系由ASSESSES/EXAMPLE_OF表示	通过ASSESSES/EXAMPLE_OF考查语言知识；通过FROM_LEVEL归属或对齐到标准等级，为偏误干预提供练习材料
偏误证据类	发音偏误、字形偏误、词汇偏误、语法偏误、偏误原因、学习者群体	偏误实例归属偏误类型，并关联偏误原因、纠正建议和高频群体	通过ERROR_RELATED_TO指向具体语言知识，为教学干预提供证据
文化拓展类	文化点、成语、典故、国情知识、跨文化交际案例、多媒体文化材料	典故、案例和多媒体材料作为文化资源的扩展属性或证据材料记录；成语与文化知识的联系由CULTURAL_RELEVANCE表达	通过CULTURAL_RELEVANCE解释和扩展语言知识；通过FROM_LEVEL按标准等级或文化阶段组织
标准等级类	HSK 2.0等级、HSK 3.0等级、国际中文教育等级、文化等级/阶段	等级序列、能力描述和数量要求构成标准等级属性；不同标准之间通过共享等级端点和FROM_LEVEL关系进行比较	作为FROM_LEVEL的目标节点连接语言知识、评估资源和文化资源，支持跨标准检索

义为domain(r)→range(r)的带约束对象关系，其中domain(r)表示关系起点允许的概念类型，range(r)表示关系终点允许的概念类型，实例边可写作e=(v_i,r,v_j)。只有当τ(v_i)∈domain(r)且τ(v_j)∈range(r)时，该关系才满足Schema约束。为体现国际中文教育的数据特点，CFL-KG关系设计遵循三个原则：一是标准可定位，即核心语言对象、题目和文化资源能够追溯到一个或多个等级标准；二是测评可解释，即题目不是孤立文本，而应显式指向被考查的知识点；三是教学可干预，即偏误、文化说明和语料证据能够与具体语言知识形成可追溯路径。

从表3可以看出，CFL-KG的关系设计并非一般知识图谱通用关系的简单堆叠，而是围绕国际中文教育的数据特点展开。第一，多标准并存决定了FROM_LEVEL等标准对齐关系的必要性；第二，教学测评需要解释题目与知识点之间的对应关系，因此需要ASSESSES/EXAMPLE_OF等测评追溯关系；第三，二语学习过程中的偏误具有教学诊断价值，因此需要ERROR_RELATED_TO等偏误证据关系；第四，

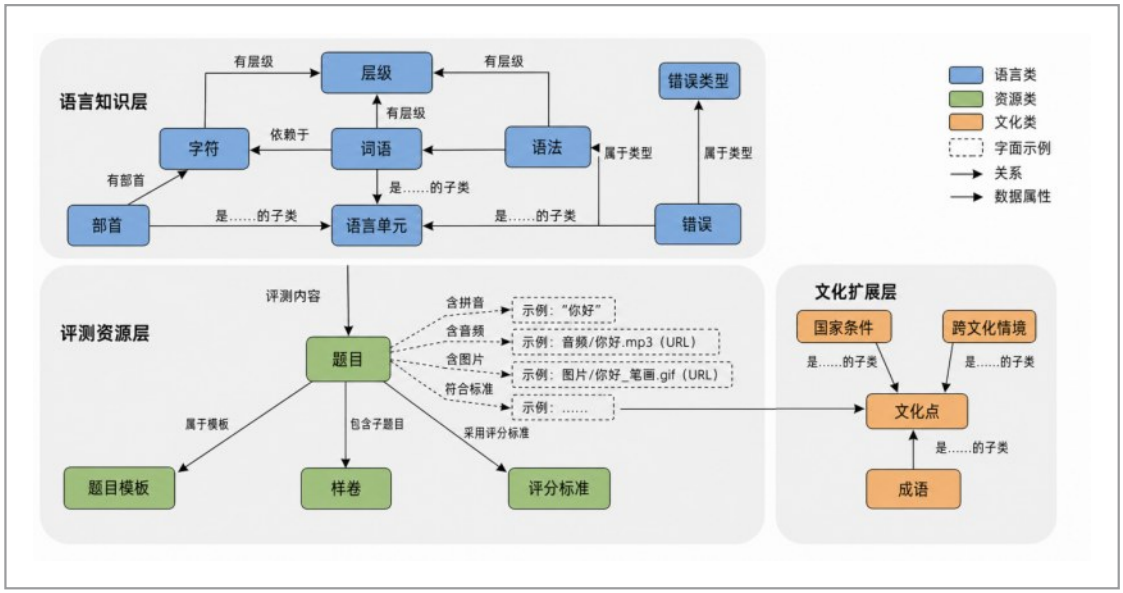


图2 CFL-KG核心数据模式

表3 主要关系类型、约束、示例及典型应用

关系类型	约束	示例	数据特点	典型应用
FROM_LEVEL	语言知识/评估资源/文化知识 -> 标准等级	“把字句” -> HSK 3.0二级；“春节” -> 文化初级	多标准并存，同一知识点需在不同标准和等级体系中定位	跨标准检索、等级比较与备课筛选
ASSESSES/EXAMPLE_OF	评估资源 -> 语言知识	题目“我把书看完了” -> “把字句”	题目常隐含考查目标，需要将题干、答案和知识点显式连接	测评追溯、题库覆盖分析与试题解释
ERROR_RELATED_TO	偏误证据 -> 语言知识	“把书看了完”偏误 -> 结果补语语序	二语学习者偏误具有诊断价值，需要绑定到具体字、词或语法	偏误诊断、纠错建议与干预材料选择
CULTURAL_RELEVANCE	语言知识/成语 -> 文化知识	“粽子” -> 端午节；“画蛇添足” -> 成语典故	国际中文教育强调语言知识与文化理解共同教学	文化解释、文化资源推荐与语用说明
DEPENDS_ON	语言知识 -> 语言知识	“把字句” -> 基本宾语结构	语言知识具有学习顺序和前驱依赖，不宜孤立呈现	学习路径组织与先备知识检查
SYNONYM/ANTONYM	词语 <-> 词语	“高兴” <-> “开心”；“容易” <-> “困难”	近义和反义词辨析是二语词汇教学中的高频需求	词汇教学、近义辨析与反义对比
CO_OCCURS_WITH	词语 <-> 词语	“承担” <-> “责任”	词语搭配和共现关系体现教学语料中的使用习惯	搭配发现与例句检索
COMPOSED_OF	词语/成语 -> 汉字	“汉语” -> “汉”“语”	汉字构成体现汉语形义结合特点	字词结构分析与书写教学

国际中文教育强调语言学习与文化理解的结合，因此需要CULTURAL_

RELEVANCE等文化拓展关系。上述关系共同将分散的字词语法、题目、偏误和文化材料组织为可查询、可追溯、可分析的教学证据链。

4 数据集构建方法

4.1 数据预处理与字段规范

数据预处理首先解决不同来源之间的字段不一致问题。对于结构化表格，按照预设字段映射表转换为统一属性；对于半结构化文档，先解析为段落、表格和列表，再依据实体类型抽取字段；对于多媒体资源，采用资源路径、文件类型、适用对象和来源说明进行登记；对于语料库证据，保留检索词、上下文窗口、频次和来源标记。

字段规范遵循“必需字段+可选字段+来源字段”的原则。必需字段包括实体名称、唯一标识、实体类型和来源；可选字段根据不同节点类型设置，如Character包含拼音、笔画、部首、发音和书写动画，Word包含释义、词性、短语、例句、近义词、反义词和标准化拼音，Grammar包含语法名称、类型、解释和例句，Question包含题干、选项、答案、解析、题型和难度。来源字段记录数据出处、整理者、更新时间和审核状态，用于后续追溯。

4.2 关系候选构建与审核

在完成实体记录的字段规范后，首先围绕这些记录生成关系候选并进行专家审核，再进入实体对齐、关系对齐与融合。这样安排的原因是：对齐与融合不是单纯的字符串合并，而需要依赖已确认或待确认的关系上下文。例如，“把字句”“‘把’字句”和“处置式”是否可以合并，不能只看名称相似度，还要考察它们是否被同类题目考查、是否对应相近等级、是否具有相同例句或解释；“端午”和“端午节”是否应融合，也需要结合其文化关联对象、例句和来源证据判断。

候选关系来自四类方法：第一，规则生成，如词语到汉字的组成关系、内容到等级的FROM_LEVEL关系；第二，词典或标准映射，如近义、反义和语法等级关系；第三，统计生成，如基于语料窗口频次得到的CO_OCCURS_WITH关系；第四，模型辅助与人工标注，如题目–知识点、偏误–知识点和文化关联关系。对于确定性较强的关系，可由规则自动生成并通过Schema约束校验；对于解释性较强的关系，采用“候选生成—专家审核—状态标记”的流程，审核状态包括通过、修改、待定和删除。

4.3 实体对齐、关系对齐与融合

在关系候选完成并经过专家审核后，进入实体对齐、属性融合和关系对齐阶段。对齐与融合要解决三个层面的问题。第一是实体层对齐，即识别不同来源中指向同一对象的记录，解决同名、异名、同形异义和跨标准重复等问题。第二是属性层融合，即处理同一实体在不同来源中释义、拼音、词性、等级、例句等字段冲突。第三是关系层对齐，即识别重复关系、方向冲突、端点不一致和跨标准关系归并问题，使同一教学事实在图谱中具有一致表达。

实体对齐采用精确匹配、拼音归一化、同义名称匹配、字段相似度和关系上下文校验相结合的方式。关系上下文是本节的关键逻辑：若两个候选实体共享相同或高

度相似的 FROM_LEVEL、ASSESSES、ERROR_RELATED_TO 或 CULTURAL_RELEVANCE 关系，则可作为融合或建立等价/近似关系的重要证据；若名称相似但关系上下文差异较大，则保留为不同实体并记录歧义标记。关系对齐则依据端点类型、关系方向、来源可信度和应用语义进行处理，确保合并后的关系既满足 Schema 约束，也符合教学解释。

4.4 入库与版本维护

完成候选关系审核与对齐融合后，实体和关系连同来源、时间、审核状态和版本号写入图数据库。入库前执行 Schema 检查，包括节点标签是否合法、关系端点类型是否符合定义、必需字段是否缺失、等级字段是否在枚举范围内、是否存在重复实体和孤立节点等。对于待定关系和冲突属性，系统保留审核状态并进入后续复核队列，避免未经确认的数据直接影响教学应用。

CFL-KG 采用 Neo4j 存储图结构数据，同时保留结构化中间表和资源文件索引。版本维护分为数据版本、Schema 版本和资源版本：数据版本记录实体和关系增删改，Schema 版本记录节点类型、字段和关系类型变化，资源版本记录多媒体和语料证据文件变化。应用反馈则通过查询日志、错误修订记录和教师反馈进入下一轮候选数据生成与审核。

5 数据规模与质量评估

5.1 数据规模统计

在当前数据快照中，CFL-KG 包含 389,111 个节点、1,017,421 条关系，覆盖汉字、词汇、语法、题目、文化、成语、部首、音节/拼音、等级和偏误等主要对象。表 4 给出了总体规模统计；图 3 以左右两个子图展示节点类型和关系类型的分布情况。结果显示，词汇节点占比较高，体现出国际中文教育资源中词汇教学和词汇关联的基础地位；关系类型中组成、题目考查和共现关系占比较高，说明数据集既覆盖语言结构，也覆盖测评和语料证据。

表4 CFL-KG 数据规模概况

指标	数值	说明
节点总数	389,111 个	当前图谱快照中的实体节点总量
关系总数	1,017,421 条	当前图谱快照中的边总量
节点类型数	10 类(统计口径)	覆盖语言、测评、文化、等级和偏误等核心对象；系统管理口径包含更多扩展标签
关系类型数	7 类(统计口径)	按核心关系类别合并统计，覆盖标准对齐、组成、近反义、考查、偏误和文化关联等
标准对齐比例	97.8%	至少关联一个标准等级的实体占比
核心语言单元	20,000+个汉字；250,000+个词汇；1,000+个语法点	用于语言知识层组织和检索
评估与文化资源	10,000+道题目；2,000+篇阅读材料；	用于测评追溯和文化拓展

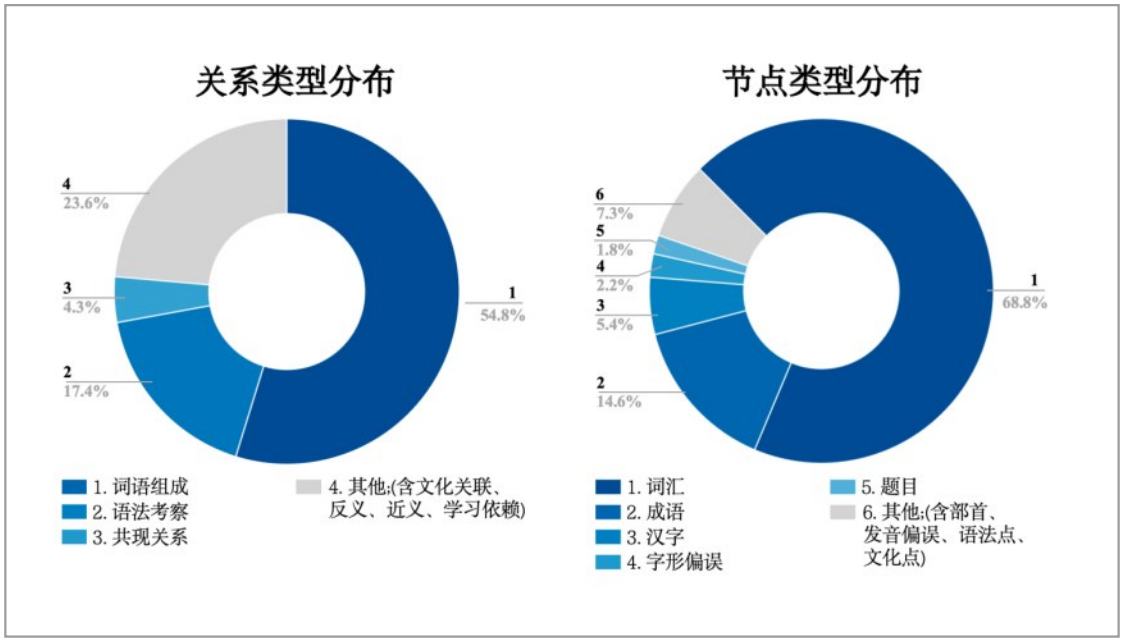


图3 CFL-KG节点类型分布与关系类型分布

5.2 覆盖性与结构质量

覆盖性与结构质量用于判断数据集是否足以支撑典型教学任务。表5显示，多源融合后重复实体率和属性冲突率处于可复核范围内，Schema违规边率为0.6%，说明大部分关系符合预定义模式。表6进一步给出图结构连通性和三跳可达性：最大连通分量占比为94.6%，知识点到题目、偏误和文化点的短路径可达性较高，表明数据集能够支撑以知识点为中心的关系追溯。但发音偏误的三跳可达比例相对较低，提示后续仍需扩充发音偏误证据。

表5 数据融合一致性与Schema校验结果

评价指标	结果	解释
重复实体率	2.8%	多源融合后检测到的疑似重复实体比例
属性冲突率	1.9%	同一实体在不同来源中存在关键属性冲突的比例
重复实体数	10,895	需要合并或人工复核的候选实体数
属性冲突实体数	7,393	需要确认权威来源或保留多值的实体数
Schema违规边率	0.6%	关系端点类型或方向不符合模式定义的边占比
Schema违规边数	6,105	自动校验后需复核的边数

表6 图结构连通性与三跳可达性

评价指标	结果	说明
弱连通分量数	37	反映图中整体连通块数量
最大连通分量占比	94.6%	大部分节点处于核心知识网络中
孤立节点比例	0.7%	度为0的节点比例较低
平均最短路径	6.8	最大连通分量内部的平均路径长度
图直径	24	最大连通分量中的最长最短路径
三跳可达题目比例	90.4%	知识点在3跳内可到达相关题目比例
三跳可达偏误比例	76.8%(文字偏误);43.9%(发音偏误)	反映偏误证据覆盖仍有提升空间
三跳可达文化点比例	65.2%	反映语言知识与文化拓展的关联程度

5.3 实例级正确性审核

实例级正确性采用分层抽样人工审核。抽样对象覆盖题目-知识点关系、偏误-知识点关系、等级对齐关系、近反义关系、文化关联关系和关键属性。表7显示，各类关键关系的准确率处于0.89-0.95之间，关键属性准确率为0.97，审核者一致性κ均高于0.80。上述结果说明，当前数据快照在核心教学关系和基础属性上具有较好的可靠性，但文化关联和偏误关联仍需要持续复核。

表7 实例级数据质量抽样审核结果

类别	审核项	样本数	准确率	一致性κ
关系	题目-知识点关系	220	0.92	0.84
关系	偏误-知识点关系	210	0.89	0.81
关系	等级对齐关系	200	0.95	0.90
关系	近义/反义关系	200	0.93	0.88
关系	文化关联关系	180	0.90	0.82
属性	拼音/部首	300	0.97	0.91

6 数据服务与典型应用

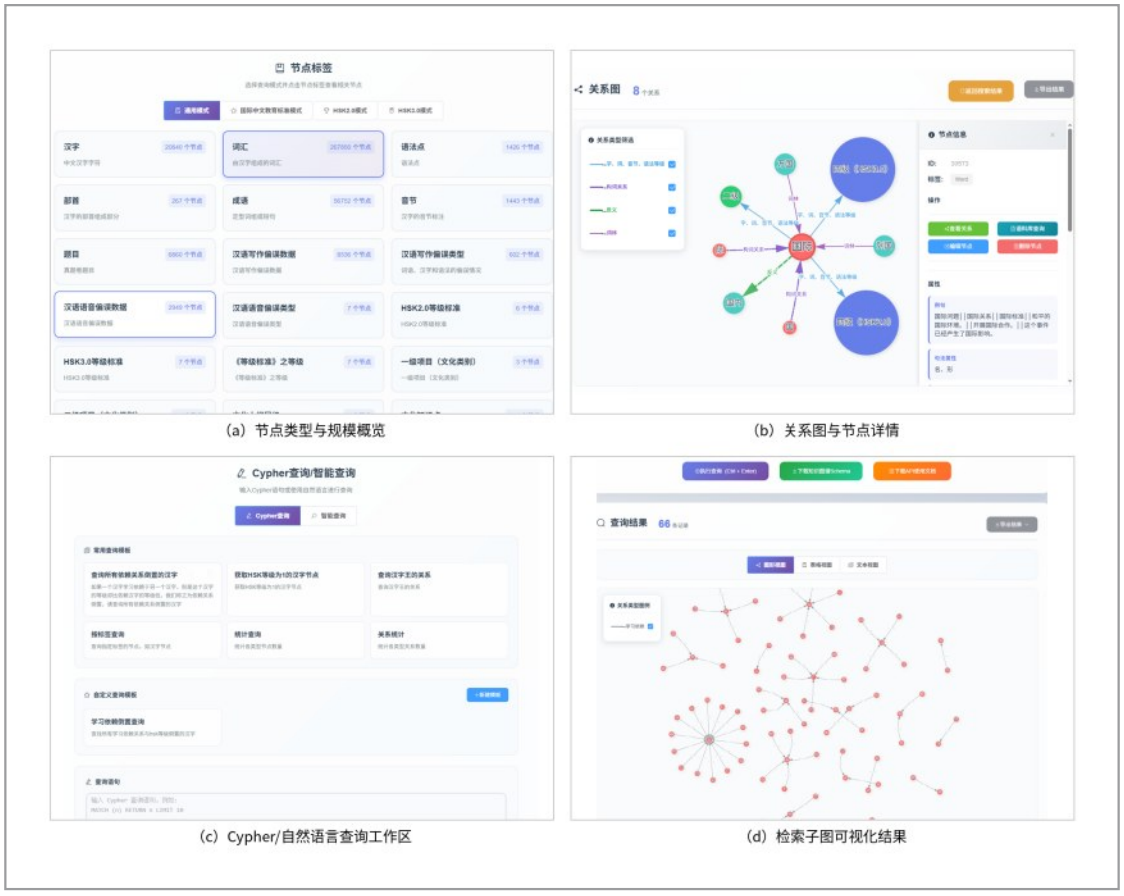
6.1 数据服务方式

为了使数据集进入实际教学和研究流程，CFL-KG提供图数据库查询、节点与关系检索、自然语言转Cypher、语料库检索、元数据统计、权限控制和API接口等服务。用户可以按标准、等级、节点类型和关键词筛选数据，也可以围绕某一实体进行一跳或多跳关系扩展。对于非技术用户，系统将常用查询封装为模板，降低图查询语言使用门槛。

如图4所示，数据集在实际服务中的典型应用入口，包括按标准模式浏览节点类型、查看关系图与节点详情、使用Cypher或自然语言查询，以及返回检索子图的可视化结果。该图用于说明CFL-KG已经能够以服务方式支撑数据检索、关系追溯和证据组织；本文的重点仍是数据集构建与应用验证，而不是单纯展示系统界面。

6.2 跨标准备课应用

真实备课场景中，教师通常不是简单查询某个词条，而是需要围绕“某一课程等级是否覆盖必要语言项目”进行决策。以初级综合课备课为例，教师先选择“HSK 3.0二级”作为入口，系统返回该等



(a) 节点类型与规模概览　(b) 关系图与节点详情

(c) Cypher/自然语言查询工作区　(d) 检索子图可视化结果

图4　数据集应用服务示例

级下的汉字、词汇和语法点；随后教师继续展开某一词语或语法点的FROM_LEVEL、DEPENDS_ON和CULTURAL_RELEVANCE关系，判断其是否需要前置复习或文化说明。若同一知识点在《国际中文教育中文水平等级标准》中等级不同，教师可以将其标记为“跨标准差异点”，在课前设计补充说明。该案例体现出CFL-KG不仅提供检索结果，还通过等级关系和前驱关系帮助教师进行课程边界判断。

6.3 测评资源追溯应用

在题库建设或试卷复核中，教师需要确认题目是否覆盖目标知识点，并避免某一知识点被过度或不足考查。以“把字句”相关题目复核为例，系统从Grammar节点出发，经ASSESSES/EXAMPLE_OF关系检索相关题目，再根据题型、难度和等级字段进行筛选。若某份样卷中多个题目均连接到同一语法点，而目标等级下其他语法点缺少题目，则说明试卷覆盖不均衡。与关键词检索相比，图谱追溯的优势在于：即使题干中没有直接出现“把字句”字样，只要题目被标注为考查该知识点，仍可被检索和统计。

6.4 偏误驱动教学干预应用

偏误驱动干预是该数据集的重要特色。以结果补语语序偏误为例，教师在批改中发现学习者频繁出现“看了完”“写了好”

等表达后，可在系统中检索相应偏误节点，沿 ERROR_RELATED_TO 关系定位到相关语法点，再查看偏误原因、高频群体、典型例句和相关练习题。该过程把错误现象、知识点解释和练习材料放在同一证据链中，有助于形成“错误定位—原因解释—例句对比—练习巩固”的教学干预路径[6-7]。

6.5 文化解释与知识增强问答应用

文化解释场景体现了语言知识与文化资源的融合价值。以“端午节”为例，系统可返回其文化说明、典故来源、相关词语“粽子”“龙舟”、相关成语或节日主题，并进一步连接到适合的等级和题目材料。在知识增强问答场景中，系统先检索图谱实体、关系路径和语料证据，再组织回答。与仅依赖文本相似度检索的方式相比，图谱数据能够提供更清晰的来源、路径和关系约束[18-19]。

6.6 应用效果初步验证

为了验证数据集在实际任务中的可用性，设置跨标准备课、测评追溯和偏误干预三类任务，比较人工检索、关键词检索和图谱数据服务三种方式。表 8 中的准确性和证据完整性均采用 0–1 评分，时间指标为任务中位完成时间。结果表明，图谱数据服务在准确性和证据完整性上均优于人工检索和关键词检索，并显著缩短任务完成时间，说明显式关系建模能够提升教学资源组织和证据追溯效率。

表 8　应用任务初步验证结果

方法	准确性(0–1)	证据完整性(0–1)	中位完成时间/s	特点
人工检索	0.83	0.76	210	依赖经验,跨资源整理耗时较长
关键词检索	0.86	0.70	145	查找较快,但难以保证关系链完整
图谱数据服务	0.92	0.88	95	能够返回关系路径和结构化证据

7 讨论

7.1 数据集的领域价值

CFL-KG 的主要价值不在于提供单一检索系统，而在于将国际中文教育领域长期分散的数据资源组织为可复用的数据底座。对于教师，该数据集能够降低跨标准备课和资源查找成本；对于测评研究者，它提供题目–知识点–等级之间的可追溯结构；对于偏误分析研究，它将学习者错误与语言知识和干预材料连接起来；对于智能教育应用，它提供比纯文本语料更稳定的结构化知识和证据链[4-5]。

7.2 数据治理与开放边界

教育知识图谱数据集的开放需要兼顾复用价值与安全边界。CFL-KG 中部分标准、词典、文化和公开资源可以通过元数据、样例数据或接口服务开放；题库、学习者偏误和多媒体材料可能涉及版权、隐

私或授权限制，适合采用脱敏、分级权限、摘要统计或受控API方式提供服务。后续可进一步建立数据卡片，明确数据来源、适用范围、质量指标、限制条件和引用方式。

7.3 局限性

当前数据集仍存在三方面不足。第一，高价值关系构建成本较高，尤其是题目–知识点关系、偏误–知识点关系和文化关联关系仍依赖专家审核。第二，跨标准对齐存在粒度差异，同一知识点在不同标准中的等级边界不完全一致，需要引入置信度和争议标记。第三，偏误证据和文化多媒体资源仍需持续扩充，特别是发音偏误、多模态材料和真实课堂反馈数据的覆盖有待提升。

8 总结

本文围绕国际中文教育数据资源建设与应用需求，构建了面向国际中文教育的知识图谱数据集CFL–KG。该数据集融合语言知识、等级标准、测评资源、偏误证据、文化内容和语料证据，设计了多标准对齐的分层数据模型，并形成从多源采集、字段规范、实体对齐、关系构建、专家审核到质量检测和服务发布的完整流程。基于当前数据快照的统计和抽样审核结果表明，CFL–KG在规模覆盖、结构连通、关系正确性和应用可用性方面能够支撑跨标准备课、测评追溯、偏误干预、文化解释和知识增强问答等场景。未来将进一步扩充偏误和多媒体数据，完善跨标准对齐的置信度标记，形成更稳定的数据版本发布与受控开放机制。

参考文献：

[1] Ministry of Education of the People's Republic of China, National Language Commission. Chinese Proficiency Grading Standards for International Chinese Language Education[S]. Beijing: Beijing Language and Culture University Press, 2021.

[2] Chinese Testing International. HSK test syllabus and related proficiency specifications[EB/OL]. ChineseTest, 2010–2023.

[3] Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs[J]. ACM Computing Surveys, 2021, 54(4): 1–37.

[4] Chen P, Lu Y, Zheng V W, Chen X, Yang B. KnowEdu: A system to construct knowledge graph for education[J]. IEEE Access, 2018, 6: 31553–31563.

[5] Qu K, Li K C, Wong B T M, Wu M M F, Liu M. A comprehensive survey of knowledge graphs in education[J]. Electronics, 2024, 13(13): 2537.

[6] Corder S P. The significance of learners' errors[J]. International Review of Applied Linguistics, 1967, 5(4): 161–170.

[7] Selinker L. Interlanguage[J]. International Review of Applied Linguistics, 1972, 10(3): 209–231.

[8] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199–220.

[9] Noy N F, McGuinness D L. Ontology development 101: A guide to creating your first ontology[R]. Stanford Knowledge Systems Laboratory, 2001.

[10] Hitzler P, Krotzsch M, Rudolph S. Foundations of Semantic Web Technologies [M]. Boca Raton: Chapman & Hall/CRC, 2009.

[11] Lenzerini M. Data integration: a theo-

retical perspective[C]//Proceedings of PODS. 2002: 233-246.

[12] Dong X L, Srivastava D. Big data integration[M]. Cham: Springer, 2015.

[13] Corbett A T, Anderson J R. Knowledge tracing: modeling the acquisition of procedural knowledge[J]. User Modeling and User-Adapted Interaction, 1995, 4 (4): 253-278.

[14] Ratner A, De Sa C, Wu S, Selsam D, Re C. Data programming: creating large training sets, quickly[C]//Advances in Neural Information Processing Systems. 2016.

[15] Musen M A. The Protege project: a look back and a look forward[J]. AI Matters, 2015, 1(4): 4-12.

[16] 李紫宣, 白龙, 任韦澄, 等. 代码大语言模型赋能的知识图谱关键技术综述[J]. 大数据, 2025, 11(02): 3, 19-28.
Li Z X, Bai L, Ren W C, et al. A survey on key technologies of knowledge graphs empowered by code large language models[J]. Big Data Research, 2025, 11(02): 3, 19-28.

[17] 李美子, 伍云芳, 卢淑怡, 等. 基于拓扑结构与相似度信息融合的教育知识图谱节点重要性评估模型[J/OL]. 大数据: 1-25[2026-06-01].
Li M Z, Wu Y F, Lu S Y, et al. A node importance evaluation model for educational knowledge graph based on fusion of topological structure and similarity information[J/OL]. Big Data Research: 1-25[2026-06-01].

[18] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems. 2020.

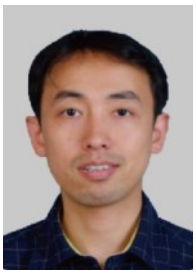
[19] Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models[EB/OL]. arXiv:2210.03629, 2022.

作者简介

王韩（2001–），男，硕士在读，华东师范大学，在读研究生，主要研究方向为教育知识图谱。

王千予（2002– ），女，硕士，华东师范大学，研究生，主要研究方向为自然语言处理，反事实数据生成。

兰韵诗，女，博士，华东师范大学，副教授，主要研究方向为知识图谱，智能问答以及其他与自然语言处理相关的任务。



王晔，男，博士，华东师范大学，专任研究员，主要研究方向为Web数据管理，海量数据挖掘，分布式系统。

梁远远，男，博士，华东师范大学国际汉语文化学院，主要研究方向为自然语言处理、大语言模型以及国际中文教育。

丁安琪，女，教授，华东师范大学国际汉语文化学院，主要研究方向为汉语作为第二语言学习者研究、国际汉语教师教育以及国际汉语教师教育。

收稿日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: