基于意合图语义理论的结构标注体系与资源建设*

郭梦溪[1]，李梦[1]，荀恩东[2†]，饶高琦[3]，于钟洋[1]
[1]北京语言大学 信息科学学院
[2]北京语言大学 语言资源高精尖创新中心
[3]北京语言大学 国际中文教育研究院
guo_mengxi@foxmail.com

摘要

意合图是一种以事件为中心的多层次语义表示方法，由事件结构与实体结构构成，通过多层次语义体系设计，实现对事件的多层次分析。本文细化并制定了意合图标注规范，采用分层分级的标注策略，在自主研发的在线标注系统中对新闻语料和国际中文教育阅读语料进行了意合图1.0标注工作。通过本次标注，验证了意合图体系的合理性和可标注性，并构建了意合图语义资源库。

关键词： 意合图；资源建设；语义表示

System and Resource Construction Based on the Semantic Theory of Chinese-Parataxis-Graph

Mengxi Guo[1], Meng Li[1], Endong Xun[2*], Gaoqi Rao[3], Zhongyang Yu[1]
[1]School of Information Science, Beijing Language and Culture University
[2]Beijing Advanced innovation Center for language Resources,
Beijing Language and Culture University
[3]Research Institute of International Chinese Language Education,
Beijing Language and Culture University
guo_mengxi@foxmail.com

Abstract

The Chinese Parataxis Graph (CPG) is a multi-level semantic representation method centered around events, consisting of event structures and entity structures. By designing a multi-level semantic system, it enables detailed analysis of events. This paper refines and establishes the annotation standards for the CPG, employing a hierarchical and graded annotation strategy. The CPG 1.0 annotation was carried out on news corpora and international Chinese education reading materials using a self-developed online annotation system. This annotation process validated the rationality and annotability of the CPG system and resulted in the construction of a semantic resource database for the CPG.

Keywords: Chinese Parataxis Graph , Resource development , Semantic representation

基金项目：国家自然科学基金“中文意合图的表征与生成方法研究（62076038）；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）（23YCX136）
† 通讯作者

1 引言

在自然语言处理领域，语义分析作为一项重要任务，其核心目标在于从自然语言表达中分析其含义与意图，为计算机理解和处理自然语言提供基础支持。语义表示是将语义分析的结果形式化地进行隐式或显式地表征，隐式表征一般是基于参数的方法进行表征，显式表征一般是基于符号的方法进行表征。目前国内外存在多种符号化的显式语义表示理论，在表示形式和理念上各有不同。基于一种语义表示理论进行语料标注，是对理论科学性的有效验证，也是实现自动分析的数据基础。因此大多语义表示理论提出的同时也构建了相应的语义资源。

意合图是荀恩东近年来提出的一种以事件为中心的多层次语义表示方法，采用单根有向图的形式承载事件、实体、属性及其相互关系，对事件结构与实体结构进行有效表示，实现对事件的多层次分析。本研究基于意合图理论，制定了意合图标注规范，开展了意合图的标注工作，以此对意合图的合理性和可标注性进行验证，并在标注过程中不断打磨与完善意合图理论体系。所构建的语义标注资源可为自然语言处理与语言学研究提供数据支持。

2 相关研究

在资源构建方面，意合图在不同发展阶段，积累了具有不同特点的基础资源，为后期基于意合图的语义标注与分析奠定了基础。荀恩东在2019年首届事理图谱研讨会上公开提出意合图理论，随后经历了句法结构树库 (卢露等, 2022)与汉语块依存树库的理论 (钱青青等, 2022)与资源建设 (钱青青等, 2022)，为意合图的逻辑结构知识、论元结构知识和篇章知识打好基础 (钱青青等, 2022)。王诚文 (2021)初次制定了意合图的论元体系，并构建了非主客体论元格标，提供了大规模高质量的动词论元知识。荀恩东 (2023)正式提出了的早期意合图理论架构，并从宏观角度提出了意合图语义分析的资源需求。王贵荣 (2023)面向意合图开展了近一万句论元结构与事件间关系的标注工作，并对事件词与论元的标注情况进行了统计分析，总结了标注不一致的原因，为意合图体系的进一步完善与标注提供支撑。意合图的事件内结构中，除论元结构外，还有情态结构。在自然语言处理领域，相较于论元，对情态信息的研究相对较少，主要集中于情态识别与基于情态的情感分析。邵田 (2023)以大数据为基础构建情态义的状语组块与补语组块语义搭配库，并对其分类体系及语义分布情况进行深入研究，该工作为意合图情态体系的确定与标注提供了理论支撑与数据支持。

意合图经过了逐步探索，在大量实例标注中不断修正认识，完善其体系。在面向意合图构建的论元知识 (王诚文, 2021)与意合图论元标注结果分析 (王贵荣, 2023)的基础上，我们进一步完善了意合图的论元结构；在情态研究 (邵田, 2023)的基础上，我们从计算性与应用性的角度出发，确定了面向意合图的情态体系；考虑到时空信息对事件的重要性，我们对事件的时间信息作进一步细化，构建了意合图的时空结构体系；并且进一步明确了意合图关系事件的定义与作用。此外，除事件结构外，在实体结构方面，我们对实体属性作出通用层面的分类，补充了意合图的实体结构体系。通过上述工作，意合图理论与其语义体系得到进一步完善，目前已初步构建出意合图完整的通用性语义体系。因此，意合图的标注工作从局部标注推进至了整体标注。本文主要从意合图的语义体系内容与当前标注工作的推进两个方面展开。一些最新的相关概念在意合图理论论文 (郭梦溪等, 2024)中已有说明，该文将酌情略写。

3 意合图语义标注体系

在计算可行的前提下，基于在通用性下易于扩展的理念，意合图围绕事件表示，将语义分析转化为事件结构与实体结构的分析，设计了多层次通用语义体系，其层次结构如图1所示。本节主要从事件结构与实体结构两部分对意合图语义体系作说明。

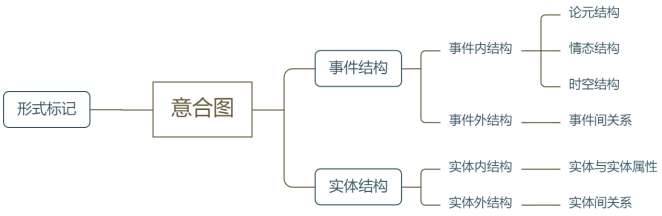

Figure 1: 意合图语义标注体系

3.1 事件结构

意合图的事件包括一般事件与关系事件。一般事件是对现实世界或可能世界中事物的动作行为或状态描述。除一般事件外，意合图还将单元的外部关系视作事件，称之为关系事件。 (郭梦溪等, 2024)事件结构分为事件内结构与事件外结构两部分，其中事件词的判断对事件结构的标注，以及标注员对意合图事件的理解尤为关键，因此本节从事件词的判断、事件内结构、事件外结构三个方面展开。

3.1.1 事件词

意合图将事件词作为事件的核心表达，事件词关联起事件内的各个成分，在事件间建立关系时，以事件词代表整个事件。根据事件词是否在句中有对应的语言单元，可以将事件词分为显性事件词和隐性事件词。

显性事件词能够在句子中找到相对应的语言单元，常为谓词，且不受句法位置限制。如“我哭肿了眼”，谓语中心“哭”与补语“肿”均为事件词；“即将远航的列车”与“列车即将远航”中的“远航”均为事件词。

隐性事件词在句子中不能找到对应的语言单元，是人为定义出的概念词语，可分为省略型与关系型两种，省略型事件词是事件词在句中缺省而进行的补全，而关系型事件词是意合图针对关系事件[0]定义出的事件词，是抽象出的关系词语。如下图所示，“目的关系”是代表目的关系事件的事件词，“Or”是代表选择关系事件的事件词。

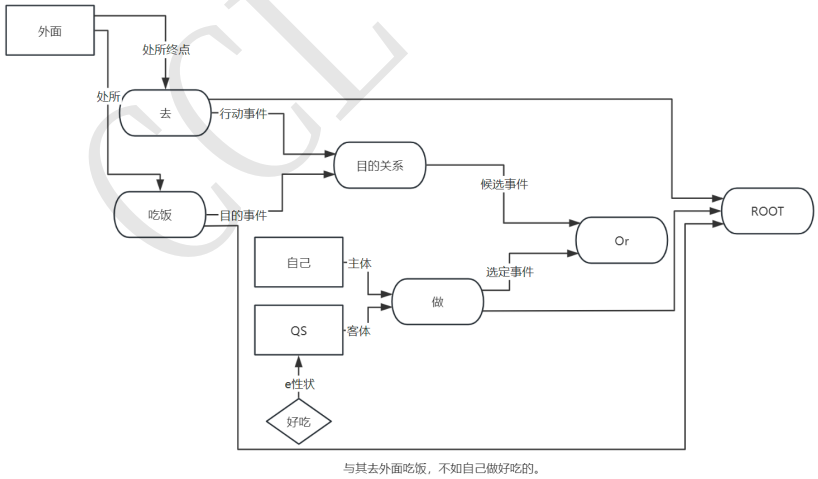

Figure 2: “与其去外面吃饭，不如自己做好吃的”的意合图表示

意合图选择事件词作为事件的核心表达，应选择能够表征事件核心概念的单元，一些不表征事件概念的虚化义谓词不应充当事件词，如形式动词、泛义动词等。其次，意合图的事件词可对应非连续的语言单元，如离合词，在形式化表征语义时，我们将其视为一个非连续性概念单元。对于语义凝固的超词形式的构式等，应视为一个概念单元。

[0]鉴于篇幅问题，本文不再详细解释关系事件，可参考《意合图：中文多层次语义表示方法》

3.1.2 事件内结构

事件内结构由论元结构、情态结构和时空结构构成，具体如下：

（一）论元结构

论元是事件的构成成分，意合图的论元分为一般论元与关系论元。一般论元是意合图一般事件的构成成分，与学界通常所说的论元角色内涵基本一致，常由实体充当，但非核心事件也可充当论元，此时非核心事件作为整体充当其他事件的构成成分。本次标注的一般论是面向事件词全集定义，不基于特定事件词针对性制定。根据与事件场景的关系，将一般论元角色分为核心论元与边缘论元，核心论元与事件场景紧密相关，边缘论元对事件场景起补充说明作用 (王诚文, 2021)。本研究的论元结构参考学界相关研究，在标注中不断打磨，最终形成包含3个核心论元和13个边缘论元的意合图一般论元体系。其中核心论元采用粗分类，将其聚类为主体、客体与邻体三类，具体情况如表1所示 (王贵荣, 2023)。边缘论元分为13类，具体情况如表2所示。

核心论元	类型	描述	示例
主体	施事	自主性动作、行为的发出者	小明打了他一下。
	当事	性质、状态或变化性事件的主体	宝塔很高。
	感事	非自主的感知性事件的主体	我太累了。
	领事	表示领属关系的领有者或整分关系中的整体	我有一本书。
	系事	类属关系或比喻关系的主体	我是中国人。
客体	受事	自发动作行为涉及的已存在事物	小明打了他一下。
	客事	非自发动作行为涉及的事件	洪水冲毁了房屋。
	成事	由施事的动作、行为造成的结果	他写了一本书。
	属事	表示领属关系的所属者，整分关系中的构成部分	我有一本书。
	类事	类属关系或比喻关系的客体	我是中国人。
	内容	动作、心理涉及但未发生变化的事件或获知信息	他知道这个消息。
邻体	与事	事件中有利害关系的间接客体	他送我一束花。
	同事	事件中所伴随的间接客体	你跟我去。
	基准	事件中进行比较和测量所参照的客体	我比他高。
	对象	事件中动作行为所针对的对象	我给她推荐了份工作。

Table 1: 核心论元

边缘论元	描述	示例
工具	动作、行为所凭借的器具	他用毛笔写字。
材料	事件过程中所凭借并消耗的物品	我拿面粉做馒头。
方式	事件中主体所采用的方式、方法	通过座谈会征询意见。
依据	事件所遵照的根据、条件等	依照宪法规定。
原因	引起事件的缘由	他以谋杀罪入狱。
目的	事件所要达到的目标	为了钱，他不择手段。
范围	事件所作用的领域、范围或覆盖面	领导就这个问题发表谈话。
数量	与事件相关的数量、幅度等量值	我比他高十公分。
数量源点	与事件相关的数量、幅度等的起始量	从50%上调至60%。
数量终点	与事件相关的数量、幅度等的终止量	从50%上调至60%。
状态	表示事件发生时主体所处或历经的状态或外部环境情况	他努力学习。 在逆境中成长。
状态源点	表示事件发生过程中主体的起始状态或外部环境的初始情况	从浑浑噩噩到努力学习。 从失败走向成功。
状态终点	表示事件发生过程中主体的终止状态或外部环境的最终情况	从浑浑噩噩到努力学习。 从失败走向成功。

Table 2: 边缘论元

除上述一般论元外，意合图的论元还包括关系论元，关系论元是关系事件的构成成分。意合图将事物间的关系也视为事件，包括实体关系事件与事件关系事件，并对每类关系事件定义了相对应的关系论元。关系论元在事件外结构与实体外结构中进行对应说明。

（二）情态结构

情态是说话人对事件的主观态度，以及事件的客观性抽象特点，是事件的属性信息，依附于事件。在与论元的区分上，情态作为事件的属性信息，强调事件的性质、特点等，而论元是事件的构成成分。从表达情态的词汇载体看，常为副词、助词、形容词等；从句法位置上看，常为状补语；从语义指向上看，大多指向事件词。在意合图中，属性与属性可先建立联系，形成嵌套属性，嵌套属性再依附于事件词，成为事件的属性。如句子“她只会说英语”中，情态信息“只”与“会”发生关系，作为事情词“说”的嵌套属性。意合图目前所定义的情态信息分为8类，分别为：程度、判断、语气、范围、情状、方式、能愿、结果，具体情况见表3所示。

情态	描述	示例
程度	事件的动作或状态达到的状况水平	这有点好看。我喜欢极了
判断	言者对事件内容的判断结果	你的答案不对。
语气	言者对事件的语气表达	河水难道倒流吗?
范围	事件的涉及范围	我只想睡觉。
情状	事件的情形	我悄悄地走了出去。
方式	事件的抽象方式特点	共同建设美好家园。
能愿	言者的主观意愿或事件的客观可能性等	我想去北京。
结果	动作或状态的结果，语义指向事件词	我们商量好了。

Table 3: 情态

（三）时空结构

时空结构包括与事件相关的时间信息与空间信息。意合图的时间信息目前包括6大类，16小类。具体情况如表4所示；意合图的空间信息目前包括4类，具体情况如表5所示。

时间信息	类型	描述	示例
时点	时间	事件发生的时间点或时期	他毕业于2023年。
	时间源点	事件开始的时间点或时期	他从早上八点学到晚上十点。
	时间终点	事件终止的时间点或时期	他从早上八点学到晚上十点。
时态	实现	表示事件的发生、出现	他去了办公室。
	进行	表示动作的进行或状态的持续	我在收拾行李。
	经历	表示曾经发生过某动作，存在某状态	我去过北京。
	将行	表示动作行为变化将要发生，或情况状态将要出现	我要去北京了。
	起始	表示动作或状态的开始，或进入一个新状态	她笑了起来。
	继续	表示已经开始的动作，或状态继续进行或存在	我听不下去了。
时段	时段	动作或状态持续的时间，或实现后所经历的时间等整段时间	我学了一个小时。
时频	时频	动作发生的频次或频率	他经常迟到。 他去过两次。

下页继续

Table 4: 时间信息

Table 4 – 续表

时间信息	类型	描述	示例
时序	表序	表示次序或重复的时间修饰语	我回家前，你把屋子打扫了。 我先休息一会儿。
	提前	动作发生的时间比预期的时间提前	它能够帮你提早发现敌人的踪影。
	准时	动作发生的时间按照预期时间准时发生	我按时完成了。
	延迟	动作发生的时间比预期的时间延迟	眼下催花技术则更侧重于遮光，让花儿延缓盛开。
时制	时制	与时间相关的其他时间修饰语。（不能归属上述语义类别的与时间相关的修饰语）	早上忽然下起了雨。

空间信息	描述	示例
处所	事件发生所在或途经的空间位置	他坐在椅子上一动不动。 他沿着这条路一直走。
处所源点	事件发生的起始空间位置	你打哪儿来? 我走出教室。
处所终点	事件发生的终止空间位置	我走进教室。 我回到中国。
趋向	事物随动作而移动的空间位置变化方向	他往东走了。 他跳下车。

Table 5: 空间信息

3.1.3 事件外结构

事件外结构即事件与事件间的关系，构成了上文所提到的事件间关系事件。意合图目前定义了8类事件间关系事件，每种事件间关系事件都由其对应的关系论元构成，具体情况如表6所示。

事件间关系	描述	关系论元	示例
时序关系	以时间为坐标轴，表述事件的前后顺序	先行事件 后继事件 伴随事件	他吃完饭<先行事件>看电视去了<后继事件>。
递进关系	具有递进关系的两个事件，其中一个事件作为基本事件，另一事件与基本事件相比，在意义上更进一层	基本事件 递进事件	他不但夺得了金牌<基本事件>，还打破了学校的纪录<递进事件>。
转折关系	具有转折关系的两个事件，其中一个事件提出某种事实或情况，另一事件转而述说与前面分句相反或相对的意思	让步事件 转折事件	小明虽然通过了考试<让步事件>，但是他一点也不骄傲<转折事件>。

下页继续

Table 6: 关系事件

Table 6 – 续表

事件间关系	描述	关系论元	示例
因果关系	具有因果关系的两个事件，其中一个事件表示原因，另一事件表示因此原因而导致的结果	原因事件 结果事件	因为太阳离我们太远了<原因事件>，所以看上去只有盘子那么大<结果事件>。
条件关系	具有条件关系的两个事件，其中一个事件提出影响事件进展的条件，另一事件说明在这种条件下所产生的推论	条件事件 推论事件	如果今天工作不努力<条件事件>，明天就得努力找工作<推论事件>。
目的关系	具有目的关系的两个事件，其中一个事件表示目的，另一个事件表示为达此目的采取的行动	目的事件 行动事件	为了搞好设计<目的事件>，技术人员吊在悬崖上进行工作<行动事件>。
并列关系	具**有选择**关**系的事件之间地位平等**	并列事件	这个西瓜又大<并列事件>又圆<并列事件>。
选择关系	具有选择关系的事件不能并存	候选事件 选定事件	你想逛商场<候选事件>还是看电影<候选事件>

3.2 实体结构

实体结构分为实体内结构与实体外结构。

3.2.1 实体内结构

实体内结构由实体与依附于实体的属性构成。意合图目前将通用层面的实体属性分为10类，具体情况如表7所示。

属性	描述	示例
数量	实体的数量、重量、尺寸大小、距离远近等物理属性的度量	我买了三斤苹果。 现有运输公司32家。 很多学生在操场。
时间	实体的存在时间或存在时间的一部分	这是唐代的一种风尚。 我相信未来的中国一定更好。
处所	实体的所在位置，以及实体的来源	公园里的花很漂亮。 给我带几瓶绍兴的黄酒。
参照	时间、空间方位、数量的参照对象	在两年前的会议上，中国力促亚太自贸区进程启动。 我推荐超市旁边的那家店。
功用	实体的用途	记者随运动员来到了比赛场地。
职能	实体的职责、职业、产业功能等	她是语文老师。 现有运输公司32家。
性状	实体内在的性质、质料等，以及实体外在的形态特征等	买个塑料凳子。 吸引更多优秀的毕业生。 会议室摆着一张圆形桌子。

继续下页

Table 7: 实体属性

Table 7 – 续表

属性	描述	示例
关涉	实体所关涉的内容	关于七夕的传说。 女性文学具有鲜明的女性意识。
指别	具有指别功能的成分	我们住在同一个小区。 其他人都不去。
表象	当单独看像领属关系，但放入该语境下并非领属关系的，往往是语义上有事件类缺省的特殊情况	他的诸葛亮很经典。 张三的英语很好。 用诗行架起心灵的桥梁。

3.2.2 实体外结构

实体外结构即实体与实体间的关系，构成了上文所提到的实体间关系事件。在一些情况下，事件可以作为整体与实体发生关系，此时该事件与实体间的关系，相当于实体与实体间的关系。意合图目前定义了5大类实体间关系事件，对每类关系作常见情况列举，细分类为开放类。每种实体间关系事件都由其对应的关系论元构成，具体情况如表8所示。

实体间关系	描述	关系论元	示例
同指关系	同一实体由不同词语指称时，不同词语间具有同指关系	Entity Event	李老师<Entity>生病了，他<Entity>女儿很担心。 经济上账目不清<Event>的问题<Entity>
领属关系	一个实体领有另一实体	现领有者 原领有者 从属者	我<原领有者>把书<从属者>送给了他<现领有者>。
整分关系	一个实体是另一个实体的组成部分	整体 部分	我<整体>哭肿了眼睛<部分>。
并列关系	两个或多个实体地位平等	并列实体	她买了香蕉<并列实体>和葡萄<并列实体>。
选择关系	两个或多个实体间不共存	候选实体	你喜欢吃香蕉<候选实体>还是葡萄<候选实体>

Table 8: 实体间关系

3.3 形式标记

为了向后期语义分析提供更多的形式化标记，我们对部分句法标记也进行了标注，主要有以下几类：

（一）介词标记

边缘论元常由介词牵引出，因此介词作为格标记，对论元角色的认定起着提示作用，当介引成分为事件时，也对意合图的事件关系认定起重要提示作用，因此介词作为最重要的形式标记之一进行标注。

（二）连接词标记

各种事件关系的连接词对事件关系的认定也同样起重要提示作用，连接词多为连词，也可以为副词等其他具有关联作用的词语。

（三）无实义方位词标记

对于不表示空间位置、时间、数量等的方位词，即无实义的方位词，视为形式标记，进行标注，如“在这个问题上，一定要一刀切”。

（四）话题标记

话题标记是根据对话题的理解和处理而产生的一种语言现象。现代汉语中，句首的“至于”“说到”“再说”等介词或动词后的体词性成分是话题，这些介词或动词性结构经常被称为话题

标记。句中语气词之前的体词性成分也是话题，因此这些语气词也可以被认为是话题标记，作为形式标记进行标注。

（五）框式结构标记

标注规范中所允许标注的框式结构以其自身意义凝固为第一原则，如“连……也”“对……而言”；以其引导成分的语义类固定第二原则，如“为了X起见”中的引导成分“X”往往表目的。

3.4 特殊现象的处理

3.4.1 省略现象

为保证语义的完整性，意合图允许新增节点进行省略补全，我们定义了事件词省略标记（EW）与实体省略（QS）标记。但由于汉语省略情况复杂，在标注时我们仅补充在该句语义标注中必不可少的语义成分。

事件词补全标记一般用于名词谓语句，除名词谓语句外，只有在句子中某些单元因事件词缺省而导致无法标注进合适的事件结构中时，属于必要性补全情况，如“我一会去超市，你（）呢？”。

实体补全一般有以下几种情况：

（一）“的”字短语

标注中，考虑到语义的明确性，我们将“的”字短语视为中心语省略的定中结构标注，如果在句子中能够找到其所指，补全后与其建立实体间关系以明确指代对象。如句子“红的（）好看”中“的”字短语的标注三元组为：(红,e性状,QS)。

但如果“的”字短语中含有事件表达，则往往不需要进行补全即可标注出其事件结构，该情况实际常出现于“是……的”句的表达中。如句子“饭是中午做的”在意合图中与“中午做饭”的标注一致。

（二）数量短语作为实体属性，实体省略

数量短语后常因表达的经济性省略实体，过往的一些标注常将其视为数量论元与谓词中心进行标注，意合图将其视为实体属性，对缺省的实体进行补充。如句子“我有三个（）”的标注三元组为：[(三个,e数量,QS),(QS,客体,有),(我,主体,有),(有,核心事件词,ROOT)]。

（三）非共识性本体与属性值的搭配

所谓共识性搭配指以实体代实体的通用属性与属性值直接搭配，如“男生很多”“他很高”，以“男生”代“男生的数量”，“他”代“他的个子”，但这种通用属性省略的搭配，已成为一种共识性搭配，一般不会产生语义不明的问题，因此不需要补全。但当句中出现非共识性搭配时，则会造成语义不明或直接推理至共识性搭配的表义，造成语义解析错误，因此应进行省略补全。如句子“她的脸特别红，你（）一点也不红”。

在标注中，标注员需对成分共享与省略补全进行区分，我们从是否出现先行语，以及是否与先行语完全对应的角度对二者进行区别，做不同的标注规范。

当句中无先行语，且符合上述省略补全情况，即可进行省略补全，如“红的没了”。

当承前省略时，多数情况下省略成分与先行语的语义结构是完全相同的，即与先行语完全对应，这时属于成分共享，不需要增添省略标记，如“他用毛巾揩抹了手脸，（他）穿好衣服。”

当省略成分能在句中找到先行语，但与先行语并非完全是同一内容，即与先行语不完全对应，可能是共享部分先行语，或共享全部先行语，但是自身还存在新的属性。在该情况下，如在上述省略补全范围内，则需要增添省略标记，并与先行语建立合适的实体间关系。如句中“四只杯子两只（）碎了”，补全省略实体，并与“四只杯子”建立整分关系。

3.4.2 特殊概念单元

对于汉语中的一些特殊句式，大多仍可以用意合图基本语义体系作出准确表示，但仍有个别句式，我们认为其句式在语义上可拆解为更为清晰，且更有利于下游任务应用的特殊概念单元。

本次标注实践中，我们在比较句中定义了特殊概念单元——比较项目（Comp）。比较句的主要语义是表达参与比较的各方在某一方面比较后的结果 (王贵荣, 2023)，一般由“比较主体、比较对象、比较项目、比较结果”四个部分组成，其中“比较主体”为主体、“比较对象”为邻体、“比较项目”为特殊概念单元标记Comp、“比较结果”为比较句的核心事件词。“比较对

象”常由标记词引出，一般为介词，标注为“CompPN”，以此对比较句进行形式标记。上述四个部分在句中可有省略，且比较主体与对象往往共享比较项目。如下例所示：

我的个子比他高。—(我, 主体, 高)、(他, 邻体, 高)、(个子, Comp, 高)、(比, CompPN, 高)

我的个子比他的高。—(我, 主体, 高)、(他, 邻体, 高)、(个子, Comp, 高)、(比, CompPN, 高)

我比他的个子高。—(我, 主体, 高)、(他, 邻体, 高)、(个子, Comp, 高)、(比, CompPN, 高)

我比他高。—(我, 主体, 高)、(他, 邻体, 高)、(比, CompPN, 高)（比较项目省略，属于共识性搭配省略，因此也无需补全）

上述例子中表示同一事件语义，如果根据基本语义体系与标注规范对比较句进行标注，若不考虑比较项目的缺省，其语义表示则不够准确；若考虑缺省情况，则常要面临与先行语不完全对应的省略情况处理，其语义表示不够简洁明了，且同一事件在表示上的一致性难以保证。因此，意合图在语义体系的构建上，在保证其基本原则的前提下，保留其一定的灵活性。

4 意合图语义资源建设

4.1 语料采样及处理

在语料的采样上，考虑到初期标注语料的规范性与代表性，因此选择了较为规范，且能体现语言生活的国际中文教育阅读语料。其次，考虑到后期能够在同一语料下与其他语义表示方法的标注结果进行有效对比，因此选择了NLP中具有代表性的宾州中文树库CTB（Penn Chinese Treebank）8.0的部分语料。在领域性上，后期将会根据落地场景针对性地选择领域化语料，暂不属本阶段工作范围。在上述考虑下，本次标注最终抽取了1754条来自BCC（Beijing Language and Culture University Corpus Center）(荀恩东等, 2016)国际中文教育语料库的阅读语料，3246条来自宾州中文树库CTB（Penn Chinese Treebank）8.0的新闻语料，共计5000条。

在语料的处理上，在自动抽取后又进行了人工筛选，确保其语法正确、语义自足、语体规范，随后进行分词处理。在分词上，为解决自动分词错误问题，我们在标注平台中设置了分词修正功能（见4.2.2）。

4.2 意合图标注实践

4.2.1 标注任务与规范

意合图语义层级分明，框架清晰，因此我们采用了分级分层逐步构建的原则，先行建立意合图框架，再逐一根据需求细化各部分内部结构。本次标注工作旨在建立意合图框架，标注内容包括细粒度的论元结构、粗粒度的情态结构、细粒度的事件间关系事件、粗粒度的实体内结构，而实体间关系事件仅标注了同指关系事件，领属关系事件与整分关系事件同实体内结构做一致标注。

分层分级展开标注能够减轻标注员的标注负担，合理控制标注员记忆内容以符合科学的标注心理，以此提高标注效率与质量。但我们在制定规范时，均从细粒度出发，标注员在前期培训中，也按照细粒度语义标签进行标注，以此加深标注员对每个语义标签的理解，避免粗粒度标注任务使标注员对语义体系的理解不够深入。因此，标注员即使在该阶段进行粗粒度标注，但其查阅规范时可看到详细的细粒度语义情况。

标注规范是资源建设的关键环节，对标注的准确性和一致性至关重要。本次任务的标注规范基于意合图语义体系展开，并根据标注情况动态修改或增加细则，不断完善规范内容。截至2024年4月，意合图标注规范共有七章，除上述基本语义体系内容外，还包括复杂情况的标注示范、标注过程中的常见问题汇总与阶段性质检反馈文档，整体内容近四万字，采用在线文档的方式将更新内容实时同步给标注员。

4.2.2 标注工具与流程

为提升标注效率和质量，我们自主开发了在线标注系统（图3）。通过边缘计算的方式，将用于检查标注结果的构图计算环节移动到终端设备进行，以提高响应速度并减少服务器的压力，从而支持系统处理大规模数据。该在线标注系统既可导入分词后的生语料，也可以导入标注过的语料进行二次标注。为了减少非认知性的不一致标注结果，我们在网页中加入提示语，并且在操作上自动排查一些常见错误，例如每条语料中至少有一个核心事件词，如检测到未标注核心事件词，在点击一下句时会自动提示返回上句标注核心事件词；对于具有对应关系的情

况，设置其对应内容，也进一步提升了标注效率。标注是以三元组标注的方式进行，但三元组集合是面向机器的表示方法，对于人并不友好，标注员不易通过三元组集合构建出整个标注对象内的语义关系，因此我们加入了根据标注三元组动态生图的功能，以更友好的方式呈现给标注员现标注情况，以便于快速排查错误。

Figure 3: 标注平台页面示例

本次标注由8位具有语言学专业背景的研究生在标注规范的指导下进行。我们本着“少量精标”的理念，采用两两对比的形式进行标注，每次任务由两位标注员独立标注同一份语料，标注完成后返回对比结果，双方对不一致结果进行讨论，确定唯一标注结果。如双方不能达成一致，则由专家（管理者）介入，确定标注结果。最终所提交的一致标注结果由管理者再次进行质检校对后入库。管理者除查看双方不一致标注结果与质检外，还阶段性参与标注，及时发现标注规范的不足与标注员存在的问题。因此，在标注平台的辅助下，同一份语料实际经过3轮、4次标注，且管理者通过多种方式深入标注工作，一定程度上保证了其标注质量。

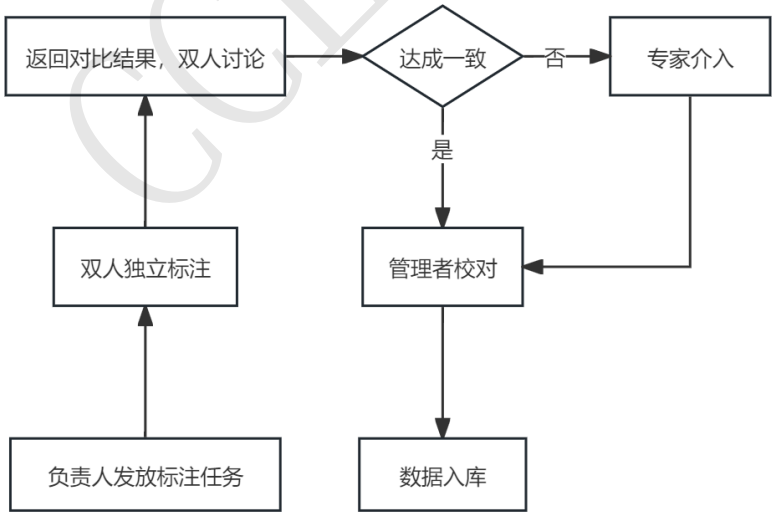

Figure 4: 标注流程图

4.2.3 标注数据与结果

标注数据共包含5000条语料，总计11.5943万字，平均每条语料有13.89个词，23.19个字。

国际中文教育领域的阅读语料平均每条12.82个词，19.22个字；来自宾州中文树库的部分新闻语料平均每条14.46个词，25.33个字。我们对标注结果进行进一步统计，其分布情况如表9、10所示。

	阅读文本	阅读文本句均	新闻语料	新闻语料句均	总体	总体句均
主体	3244	1.85	5973	1.84	9217	1.84
客体	2301	1.31	5184	1.60	7485	1.50
邻体	261	0.15	543	0.17	804	0.16
工具	47	0.03	39	0.01	86	0.02
材料	3	0.00	9	0.00	12	0.00
方式	74	0.04	191	0.06	265	0.05
依据	23	0.01	186	0.06	209	0.04
原因	43	0.02	53	0.02	96	0.02
目的	0	0.00	14	0.00	14	0.00
范围	77	0.04	390	0.12	467	0.09
数量	18	0.01	191	0.06	209	0.04
数量源点	1	0.00	20	0.01	21	0.00
数量终点	2	0.00	116	0.04	118	0.02
状态	87	0.05	144	0.04	231	0.05
状态源点	5	0.00	14	0.00	19	0.00
状态终点	33	0.02	45	0.01	78	0.02
情态	2349	1.34	2054	0.63	4403	0.88
实体结构	3139	1.79	9909	3.05	13048	2.61
核心事件词	2837	1.62	4357	1.34	7194	1.44

Table 9: 标注情况统计

	阅读文本	新闻语料	总计
时序关系	262	275	537
递进关系	30	59	89
转折关系	52	51	103
因果关系	237	183	420
条件关系	135	81	216
目的关系	144	202	346
并列关系	337	1172	1509
选择关系	37	42	79
同指关系	224	588	812

Table 10: 外部关系标注情况统计

由于意合图是以事件为中心的语义表征图，我们对标注数据中所含事件词情况进行了统计。统计得到，该标注数据中有12812个不同的事件词（不包含隐性事件词）。在国际中文教育领域的阅读语料中出现次数前五的事件词为“是”“有”“说”“到”“去”；在来自宾州中文树库的部分新闻语料中出现次数前五的事件词为“是”“说”“发展”“有”“投资”；总标注语料中出现次数前五的事件词为“是”“有”“说”“发展”“投资”。

5 结语

本文描述了意合图语义体系的具体内容，并介绍了基于意合图语义体系在展开的标注工作，对意合图语义体系进行打磨，验证了其合理性与可标注性，并构建了一批语义标注资源，为后期意合图自动生成提供数据基础。该资源也开源投入评测活动，供学界同仁使用。

在接下来的工作中，我们将进一步审视标注结果，打磨意合图理论，为意合图的应用探索可行途径，并基于后续应用需求制定2.0计划。

致谢

意合图1.0标注工作由北京语言大学李梦、何晴、胡星雨、王静怡、吴晓靖、张可芯、周书帆、朱奕瑾（按姓氏排）八位研究生完成，在此表示感谢。感谢匿名审稿人对论文提出宝贵的修改意见。

参考文献

Bin Li, Yuan Wen, Lijun Bu, Weiguang Qu. 2016. *Annotating the Little Prince with Chinese AMRs.* In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Jingyi Wang, Endong Xun. 2023. Construction and Statistical Analysis of Discourse Cohesive Components in Modern Chinese. International Journal of Asian Language Processing, 33(04).

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. *The Proposition Bank: An Annotated Corpus of Semantic Roles.* Computational Linguistics, 31(1):71–106.

Yihuan Liu, Bin Li, Peiyi Yan, Li Song, Weiguang Qu. 2016. *Ellipsis Annotation and Statistics in Chinese based on Chinese AMR.* International Journal of Knowledge and Language Processing.

陈立民. 2002. 汉语的时态和时态成分. 语言研究.

龚千炎. 2004. 现代汉语的时间系统. 语言文字应用研究论文集（Ⅱ）.

龚千炎. 2012. 汉语的时相、时制、时态. 商务印书馆, 北京.

郭梦溪, 荀恩东, 李梦, 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

黄彤, 李斌, 闫培艺, 戴玉玲, 曲维光. 2020. 基于抽象语义表示的汉语构式标注与分析. 中文信息学报.

卢露, 矫红岩, 李梦, 荀恩东. 2022. . 基于篇章的汉语句法结构树库. 自动化学报,48(12): 2911–2921.

刘亚慧, 杨浩苹, 李正华. 2020. 一种轻量级的汉语语义角色标注规范. 中文信息学报.

钱青青, 王诚文, 王贵荣, 饶高琦, 荀恩东. 2022. 基于组块分析的汉语块依存语法. 中文信息学报, 36(08): 20–28.

钱青青, 王诚文, 荀恩东, 王贵荣, 饶高琦. 2022. 汉语块依存语法与树库构建. 中文信息学报, 36(07): 50–58.

邵田. 2023. 基于大数据的汉语情态义简单状补组块研究. 北京语言大学博士论文.

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

邵艳秋, 邱立坤, 梁春霞, 毛宁. 2011. 中文语义依存树库构建及自动分析技术. 中国中文信息学会.中国计算语言学研究前沿进展（2009-2011）.

宋衡, 曹存根, 王亚, 王石. 2023. 一种改进的汉语语义角色分类体系与标注实践. 中文信息学报.

田思雨, 邵田, 荀恩东, 饶高琦. 2023. 基于结构树库的补语位形容词语义分析及搭配构建. 第二十二届中国计算语言学大会论文集,第420页-第432页,哈尔滨.

王诚文. 2021. 面向意合图的汉语动词论知识构建研究. 北京语言大学博士论文.

王诚文, 钱青青, 荀恩东, 邢丹, 李梦, 饶高琦. 2020. 三元搭配视角下的汉语动词语义角色知识库构建. 中文信息学报.

王贵荣. 2023. 意合图事件结构标注及分析研究. 北京语言大学博士论文.

文贞惠. 1999. *"N_1(*的*)N_2"*偏正结构中*N_1*与*N_2*之间语义关系的鉴定. 语文研究.

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

荀恩东. 2023. 自然语言结构计算：*BCC*语料库. 人民邮电出版社, 北京.

荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制. 语料库语言学.

萧国政. 2020. 语法事件与语义事件——面向人工智能的语言研究. 长江学术.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

张牧宇, 秦兵, 刘挺. 2014. 中文篇章级句间语义关系体系及标注. 中文信息学报.

周国光. 2002. 现代汉语的语义属性系统. 世界汉语教学.