面向汉语中介语的依存句法标注规范*

肖丹[1,2,3]，杨尔弘[1,2]，张明慧[1,2,3]，陆天荧[1,2,3]，杨麟儿[1,2,3]

（1. 北京语言大学 语言资源高精尖创新中心，北京 100083；
2. 北京语言大学 国家语言资源监测与研究平面媒体中心，北京 100083；
3. 北京语言大学 信息科学学院，北京 100083）

摘要：汉语中介语是伴随着汉语国际教育产生的，随着汉语学习在全球的不断开展，汉语中介语的规模不断增长，由于这些语料在语言使用上有其独特性，使得中介语成为语言信息处理和智能语言辅助学习的独特资源。目前，依存分析是语言信息处理和智能语言学习的重要步骤，依存语法以其形式简洁、易于标注、便于应用等优点，被广泛应用于语料标注中。面向英语中介语的依存语法标注语料已经有很好的应用，而现有汉语中介语语料库对句法的关注度普遍较低，并且缺乏一个充分考虑汉语中介语特点的依存句法标注规范。本研究面向中介语的依存句法分析，建构汉语中介语依存标注语料库。本文探讨依存标注规范，在充分借鉴国际通用依存标注体系（Universal Dependencies）的基础上，制定了面向汉语中介语的依存标注规范，以期解决现有研究的不足。

关键词：汉语中介语；依存句法；标注规范

中图分类号：TP391　　**文献标识码：**A

Dependency Annotation Guideline for Chinese Inter-language

XIAO Dan[1,2,3], YANG Erhong[1,2], ZHANG Minghui[1,2,3], LU Tianying[1,2,3],YANG Liner[1,2,3]

(1. Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China;
2. Chinese National Language Monitoring and Research Print Media Sub-center, Beijing Language and Culture University, Beijing 100083, China;
3. School of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract : Chinese interlanguage is accompanied by Chinese international education. With growing development of Chinese language learning in the world, the scale of interlanguage in Chinese has been expanding. Considering the uniqueness of using inter-language, it has become a unique resource for language information processing and intelligent language assisting learning. At present, dependency analysis is an important step in language information processing and intelligent language learning. Owing to the simple form, labeling simplicity and user-friendliness, dependency grammar has come to be widely used for corpus annotation. Dependency grammar annotation corpus for inter-language in English has been well applied. However, the current Chinese inter-language corpora pays little attention to syntax and lacks an annotation guideline for dependency syntax that fully considers the characteristics of Chinese inter-language.This study aims to construct the corpus of inter-language dependency annotation in Chinese by analyzing the dependency syntax of inter-language. Discussing the dependency annotation guideline, this paper, based on the Universal Dependencies, develops a new dependency annotation guideline for Chinese inter-language, with the aim to make contribution to existing research.

Key words：inter-language; dependency grammar; annotation guideline

***收稿日期:；定稿日期:**

基金项目：北京语言大学校级项目（中央高校基本科研业务费专项资金）（18YBB20）；语言资源高精尖创新中心项目（TYZ19005）

作者简介：肖丹（1995-），女，硕士研究生，主要研究方向为自然语言处理；杨尔弘（1965-），女，通讯作者，教授，博士生导师，主要研究方向为自然语言处理；张明慧（1996-），女，硕士研究生，主要研究方向为自然语言处理；陆天荧（1995-），女，硕士研究生，主要研究方向为自然语言处理；杨麟儿（1983-），男，博士，主要研究方向为句法分析。

1 引言

中介语指的是由于学习外语的人在学习过程中对于目的语规律所做的不正确的归纳与推论而产生的一个语言系统，这个语言系统既不同于学习者的母语，又区别于他所学的目的语，在这个过程中就产生了“偏误”，即中介语与目的语规律之间的差距[1]。汉语中介语是汉语学习者在学习汉语的过程中产生的一种特殊的语言系统，包含大量不规范语言。汉语中介语作为一种独特的语言资源，在汉语国际教育中的作用日益凸显，目前的汉语中介语语料库在“字”、“词”的标注上较为深入，但是对句法结构的关注度不够[2]。

句法分析是自然语言处理的重要基础任务，中介语的自动句法分析也已成为该领域的重要研究内容，汉语中介语句法分析是计算机辅助汉语作为第二语言学习的重要部分，可以应用到句法复杂度分析、辅助写作等多个任务中。因此，构建具有句法标注的汉语中介语语料库，是从一个新的视角探索语言信息处理、智能辅助汉语学习的基础型数据研究，制定面向汉语中介语的句法标注规范成为首要任务。

依存句法用于分析输入句子的句法结构，将词语序列转化为树状的依存结构[3]，来捕捉句子内部词语之间的修饰或搭配关系，描写句法结构。依存句法以其形式简洁、易于标注、便于应用等优点，被广泛应用于资源建设的语料标注中。国际通用依存标注体系（Universal Dependencies[1]，以下简称“UD”）是目前拥有语言种类最多的通用依存树库。它通过构建跨语言树库，捕捉不同语言之间的相似性和特殊性，为所有语言提供统一的标注方案，来解决句法分析器在跨语言分析上效果不佳的问题[4]。截止目前，最新版本的 UD V2.4 已发布了 83 种语言的标注数据，共 146 个树库。香港城市大学[5]公开发布了基于 UD 的汉语中介语依存句法规范。通过实际语料分析，我们认为该规范：1）对汉语中的特殊结构未做考虑；2）标注过程为了适应规范对语料做了一定程度的修改；3）没有充分考虑中介语中的偏误对标注原则和标注结果的影响。

本文充分借鉴 UD V2，提出一个新的面向汉语中介语的依存标注规范，包括标注框架和标注原则两大部分。规范的制定是基于对 HSK 动态作文语料库的文本进行标注实践不断完善的。

2 相关工作

中介语作为一种特殊的语言系统，尤其是带有显性句法信息标记的中介语语料库对语言信息处理和智能语言辅助学习来说具有重要意义。目前国外的学习者树库构建已相对完善，但是汉语上相关研究的进展则较为缓慢，尚存在一些问题。

英语学习者句法标注项目（The Project on Syntactically Annotating Learner Language of English，以下简称“SALLE”）是由 Ragheb 和 Dickinson[6]等人构建的学习者树库，语料来源于大学生所写的英语短文，采用 SUSANNE Corpus[7]的词性标签集和儿童语言数据交流系统[8]（Child Language Data Exchange System，简称 CHILDES）的依存标签集进行标注。SALLE 关注句子表层结构，是学习者树库的先驱，对于二语句法研究具有重要意义。但该树库未关注“语法错误”，因而不能应用于语法错误识别、语法改错等相关任务中，并且无法满足跨语言的对比分析。为此，Berzak[9]等人构建了英语学习者树库（The Treebank of Learner English[2]，以下简称“TLE”），语料来源于剑桥学习者语料库[10]（Cambridge First Certificate in English learner corpus）并采用 UD 框架进行标注，使之能够进行多语言的对比分析。一方面 TLE 树库为二语习得领域的错误分析提供了大量的

[1] UD V2 网址：https://universaldependencies.org/guidelines.html

[2] The Treebank of Learner English 的网址：http://esltreebank.org/

实证语料，促进第二语言教学与研究的发展，另一方面 TLE 树库为学习者语料的句法分析器提供了大量的训练语料，并且通过实验表明基于 L1 和 L2 的平行依存树库可以使中介语句法分析的准确率得到显著提升。

与构建相对完善和应用较为广泛的英语学习者树库相比，汉语学习者树库的建设尚处于起步阶段。目前，国内相关汉语中介语语料库缺乏句法结构信息，主要关注“字”、“词”等偏误标注。例如，北京语言大学 HSK 动态作文语料库[1]、中山大学汉字偏误标注的汉语连续性中介语语料库[2]、暨南大学留学生汉语书面语语料库[3]等，虽然这些语料库为汉语国际教育做出重要贡献，但是在句法方面就稍显薄弱。香港城市大学制定了面向汉语学习者的标注规范，并构建了汉语学习者树库 UD_Chinese-CFL[4]。该规范考虑到中介语的特点，在继承 TLE 和 SALLE“字面标注”的基础上对汉语的词目、词性、依存关系等进行了解释说明。但是，由于基本沿用面向汉语普通话的依存标注框架[11]，该规范存在两方面的问题：从标注框架的角度，1）对汉语特殊词性考虑不够周全，只考虑到方位词处于介词与方位短语构成的介宾结构中，而未考虑到方位词的其他用法；2）删掉“虚位（expl）”标签,增加了 18 个小类标签，在一定程度上增加了标注的负担；3）对汉语中的特殊结构未作考虑。从标注对象的角度，UD_Chinese-CFL 对语料进行了一定程度的修改，而且仅继承“字面标注”原则，未全面考虑偏误对于标注方式和标注结果的影响。

鉴于此，为了更加充分地刻画汉语中介语的句法结构,我们在借鉴 UD V2 的基础上，对汉语特殊词性、句法结构、汉语中介语的特性以及标注一致性等问题做了全面考虑，制定了面向汉语中介语的依存句法标注规范，主要包括标注框架和标注原则两大部分。

3 标注框架

本节提出适应汉语特点的标注框架，包括词性和依存标签两部分。UD_Chinese-GSD[5]、UD_Chinese-PUD[6]、UD_Chinese_HK[7]分别是谷歌、CoNLL 2017、香港城市大学在 UD 上发表的汉语树库。与主要考虑印欧语言特点的 UD 相比，这些树库在词性和依存关系上都做了一定的调整，但仍存在一定的不足。主要表现在：1）考虑了汉语特殊词性，但不够全面。例如，香港城市大学只考虑到方位词处于介词与方位短语构成的介宾结构中,而未考虑到方位词的其他用法。2）保留的标签和增加的标签不能充分地刻画汉语句法结构。例如，谷歌和 CoNLL 2017 都增加了“定语从句：关系从句（acl:relcl）”这个次类标签，但是通过语料对比分析，发现“acl:relcl”标签和“定语从句（acl）”标签没有区别。3）增加的标签存在一定的问题。例如，谷歌和 CoNLL 2017 增加了“辅助：致使关系（aux:caus）”标签表示引出受事宾语的“把”和谓语动词之间的介宾关系。根据次类标签应当符合主类标签基本属性的原则，应当采用“介宾关系（case）”的次类标签，而不是“辅助关系（aux）”。4）树库的质量存在一定的问题。例如，谷歌发布的 UD_Chinese-GSD 中，“acl”标签使用不明确，既用来表示嵌套结构作定语修饰名词性成分，也可以用来表示方位短语中方位词与名词的关系。5）没有关注到汉语中的特殊句式结构。

因此，我们在充分借鉴 UD V2 的基础上，制定了更加适应汉语特点的标注框架，包括词

[1] 北京语言大学 HSK 动态作文语料库：http://202.112.195.192:8060/hsk/login.asp
[2] 中山大学汉字偏误标注的汉语连续性中介语语料库：http://cilc.sysu.edu.cn/
[3] 暨南大学留学生汉语书面语语料库：http://huayu.jnu.edu.cn/corpus3/Search.aspx
[4] UD_Chinese-CFL 的网址：https://universaldependencies.org/treebanks/zh_cfl/index.html
[5] UD_Chinese-GSD 的网址：https://universaldependencies.org/treebanks/zh_gsd/index.html
[6] UD_Chinese-PUD 的网址：https://universaldependencies.org/treebanks/zh_pud/index.html
[7] UD_Chinese_HK 的网址：https://universaldependencies.org/treebanks/zh_hk/index.html

性和依存标签。主要创新点在：1）保留了 UD V2 的 16 个词性标签，并对 UD V2 中没有说明的汉语特殊词性现象以及上述汉语树库中没有充分考虑的现象提供了独特的处理方式。2）为避免上述树库在保留和增加标签时出现的问题，在一定的考量下保留了 29 个主类标签，增加了 8 个次类标签。3）对所有的标签根据汉语的句法理论体系分为三大类，包括单句主干关系标签、单句其他关系标签、嵌套关系标签，使之体系化。4）针对汉语中的独特结构，提出了特殊的标注策略，以便更加充分地刻画汉语句法结构。

3.1.汉语中几种特殊词性的标注方法

方位词是汉语中一种特殊现象，属于名词类别，表示位置和方向，例如：“上、下、以前、以后”等。虽然张斌在《现代汉语描写语法》[12]中认为它具有黏着性，不能单独充当句法成分，经常依附在一些词语后面构成方位短语，但是在某些情况下，方位词也可以充当句法成分，例如：前后照应、前后矛盾等。除此之外，方位词也可以用于诸如“在……里、因……上、在……中”等结构中充当后置词。针对这种现象，刘丹青[13]引入“框式介词”的概念，即在名词短语前后由前置词和后置词一起构成的介词结构，认为处于框式结构中的方位词因而具有了介词的性质。因此，在充分考虑了方位词的特殊性之后，我们做出如下处理：

当方位词处于框式介词结构中，我们认为方位词粘着性较强，具有介词的性质，因此将方位词当作介词处理，标为 ADP，如图 3-1。其他情况下，方位词的名词属性较为强烈，我们仍当作名词处理，标注为 NOUN,如图 3-2。

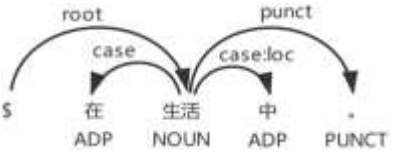

图 3-1 方位词处于框式结构中示例

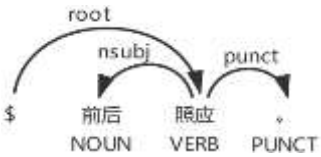

图 3-2 方位词处于非框式结构中示例

量词，通常放在数词或代词后面，表示计量单位[14]，例如：一张纸、一桶水等。在 UD 框架中，没有表示量词的标签。对此，香港城市大学[11]将量词归入名词类别当中。我们认同这一做法，也将量词当作名词处理，标为 NOUN。如图 3-3 所示,“一支笔”中的“支”标为 NOUN。

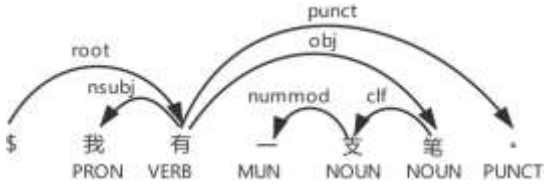

图 3-3 量词标注示例

连词分为组合连词和关联连词[12]，主要起连接作用。组合连词用于连接词或短语，关联连词用于连接小句或句子。UD 框架中有并列连词的标记（conj）和从属连词的标记（sconj）。在理论上，将关联连词的标记直接对应到从属连词的标记上是可行的，但由于在汉语中，连词、副词都可以充当关联词语，所以在实际标注中往往会出现将连词、副词混淆的问题。因此，我们将关联连词当作副词处理，标为 ADV。

3.2 依存关系标签介绍

UD V2 一共包含 36 个依存关系标签，根据汉语的句法结构特点，我们保留了 29 个主类标签，增加了 8 个次类标签，以便更好地描写汉语句法结构。删掉的标签有：cop、expl、fixed、compound、list、parataxis、goeswith、reparandum；增加的标签有：nsubj:pass、csubj:pass、mark:advb、mark:comp、mark:relcl、case:aspect、case:loc、case:dec。在此基础上，结合汉语句法理论我们将上述依存关系标签分为三大类：单句主干关系标签、

单句其他关系标签、嵌套关系标签，解决了 UD 缺乏理论体系的问题[15]，使之体系化。如表 3-1 所示。

表 3-1：体系化依存关系标签

	依存关系	说　明	例　句	标注示例
单句主干关系标签	root	sentence root	父亲很爱我们	root $ → 很爱
	nsubj	nominal subject	父亲很爱我们	nsubj 父亲 ← 很爱
	nsubj:pass	passive nominal subject	观念也被带入了中国	nsubj:pass 观念 ← 带入
	obj	object	你们给我这个机会	obj 给 → 我
	iobj	indirect object	你们给我这个机会	iboj 给 → 机会
	obl	oblique nominal	跟我约定	obl 我 ← 约定
	advmod	adverbial modifier	少喝酒	advmod 少 ← 喝酒
	nmod	nominal modifier	时代差异	nmod 时代 ← 差异
	amod	adjectival modifier	新的观念	amod 新 ← 观念
	appos	appositional modifier	一本书《谁动了我的奶酪》	appos 书 ← 动
	nummod	numeric modifier	一封信	nummod 一 ← 封
	aux	auxiliary	能理解	aux 能 ← 理解
	clf	classifier	一封信	clf 封 ← 信
	det	determiner	每个国家	det 每个 ← 国家
	conj	conjunct	生活与学习	conj 生活 → 学习
	punct	punctuation	一封信。	punct 信 → 。
单句其他关系标签	vocative	vocative	爸爸、妈妈：你们好	vocative 爸爸 ← 好
	dislocated	dislocated elements	出产的桃子人吃了	dislocated 桃子 ← 吃
	discourse	discourse element	这不行吧	discourse 不行 → 吧
	mark	marker	如果产妇吃了化学污染的食品，后代的孩子会发生健康问题	mark 如果 ← 吃
	mark:advb	manner adverbializer	不断地增加	mark:adv 不断 → 地
	mark:comp	manner complementizer	家庭维持得很好	mark:comp 维持 → 得
	mark:relcl	manner adjectivalizer	美好的事	mark:relcl 美好 → 的
	case	case marking	在生活中	case 在 ← 生活
	case:aspect	aspect	新观念也被带入了中国	case:aspect 带入 ← 了
	case:loc	postpositional localizer	在生活中	case:loc 生活 → 中
	case:dec	manner nominalizer	外面的社会	case:dec 外面 → 的
	cc	coordinating conjunction	生活与学习	cc 与 ← 学习

	flat	flat multiword expression	弗洛伦斯▪南丁格尔	弗洛伦斯 $\xrightarrow{flat}$ 南丁格尔
	orphan	orphan	他给奶奶一个手表，爷爷一个手表	爷爷 $\xrightarrow{orphan}$ 手表
嵌套关系标签	csubj	clausal subject	学汉语是最好的	学 $\xleftarrow{csubj}$ 最好
	csubj:pass	clausal passive subject	烧荒肥田是一种原始的农业技术，曾被广泛应用于世界大部分地区	csubj:pass 是 ← 应用
	xcomp	open clausal complement	我去北京学习	去 $\xrightarrow{xcomp}$ 学习
	ccomp	clausal complement	我认为社会问题是教育方面的问题	认为 $\xrightarrow{ccomp}$ 是
	acl	clausal modifier of noun	我最尊敬的人	尊敬 $\xleftarrow{acl}$ 人
	advcl	adverbial clause modifier	他拍桌子保证	拍 $\xleftarrow{advcl}$ 保证
	dep	Unspecified dependency	日本的漫画读者包括了所有的年龄层,因此日本漫画的题材非常广泛	包括 $\xrightarrow{dep}$ 广泛

（依存标签详细信息见附录）

3.3 汉语特殊结构的标注策略

汉语中存在一些不同于印欧语言的特殊结构，如连谓、兼语、“是……的”等。为了准确刻画这些结构，我们提出了面向汉语特殊结构的标注策略。

由连谓短语充当谓语或独立成句的句子叫连谓句[14]，例如：肇事者开着车跑了。“开着车跑”是一个连谓短语，做整个句子的谓语。连谓句具有以下特征：连用的动词或动词性短语之间不能有语音停顿，书面上不能有逗号隔开；而连用的动词或动词性短语之间没有关联词语也没有分句间的逻辑关系;介词短语和动词或动词性短语连用是非连谓结构[12]。考虑到上述连谓句特点，我们既不能把它当作并列结构（conj）看待，也不能把它看作一个介宾结构修饰动词，即 obl。因此，在充分考虑连谓句特点的基础上，我们规定：把后一个动词或动词短语所带的结构看作是前一个动词或动词短语的补充成分，即 xcomp，如图 4-1 所示。

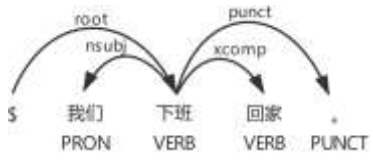

图 4-1 连谓句标注示例

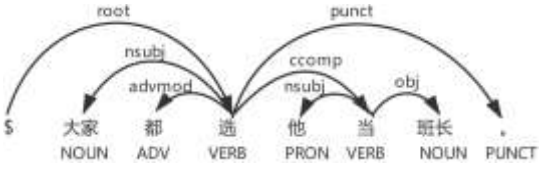

图 4-2 兼语句标注示例

由兼语短语充当谓语或独立成句的句子叫兼语句[14],例如：他有个妹妹很能干。“妹妹”既是“他”的宾语又是“能干”的主语。如果采取直接标注的方式，会造成一个词有两条入弧。因此，为了解决这种特殊现象，我们规定：把兼语短语看作是对前一个动词的补充说明，标为 ccomp（不同主语），如图 4-2 所示。

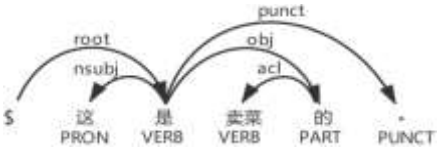

图 4-3 “是……的”示例一

图 4-4“是……的”示例二

“是……的”特殊结构在汉语中通常表示强调，其中“是”是起强调作用的副词，“的”是句末表示语气的助词。但是这种结构经常会与含有“的”字短语的判断句（在这种情况中，“是”是谓语动词，“的”是“的”字短语中的结构助词）发生混淆。以“这是很难说的”为例，我们既可以认为“这是很难说的事情”，也可以认为“这很难说”。因此，我们做如下

规定：当“的”字短语的中心语在无需考虑上下文语境就可以省略时，“的”为结构助词，标为 obj，如图 4-3；当“的”字短语的中心语在无上下文语境不可以省略时，“的”为语气词，标为 dicourse，如图 4-4。

4 标注原则

本节给出的标注原则，旨在处理汉语中介语的不规范现象。

TLE[9]提出“字面标注”的标注原则，强调根据观察到的语言使用现象进行句法分析；香港城市大学[5]遵循了 TLE 提出的“字面标注”原则，并在词目、词性和依存关系上做了说明。“字面标注”原则遵循了二语习得研究领域的基本原则，即客观、准确地描述学习者语言。但是，我们认为“字面标注”存在以下问题：1）概念过于含糊，没有一个明确的界定，在实际的标注过程中容易产生误解；2）TLE 和香港城市大学在面对一些不符合“字面标注”原则的语言现象时，都采用“例外”进行解释。这表明在实际标注过程中“字面标注”原则不能涵盖所有语言现象。

因此，我们在充分吸取前人研究成果的基础上，提出了更为准确、细致的标注原则，即：对语法上不具有可解释性的中介语，根据偏误纠正后获得的目标句进行词性标注和依存句法分析；对语法上具有可解释性的中介语，根据观察到的句法结构对其进行词性标注和依存句法分析。

4.1 对语法上不具有可解释性中介语，根据目标句进行词性标注和依存句法分析

语法上不具有可解释性，我们分为两种情况：1）无法判断所观察到的语言现象的句法结构；2)可以判断所观察到的语言现象的句法结构，但是其句法结构不符合语法规则。

4.1.1 无法判断其句法结构

无法判断其句法结构，是指由于书写错误或用词错误等导致的无法正常理解其句法结构的情况，主要包括音近或形近、具有相同语素、成语成分缺失或赘余。对此，我们根据目标句进行词性标注和依存分析。

音近或形近，指的是偏误和目标句具有语音或书写形式上的相似性，如图 4-5 和图 4-6。“复杂”和“负杂”具有相同的声母和韵母；“许多”和“仵多”在书写形式上近似，但是偏旁不一样，一个是单人旁，一个是言字旁。面对“负杂”、“仵多”等这样不合法且无法从字面获取有效信息、判断其句法结构的语言现象，我们根据目标句进行标注。

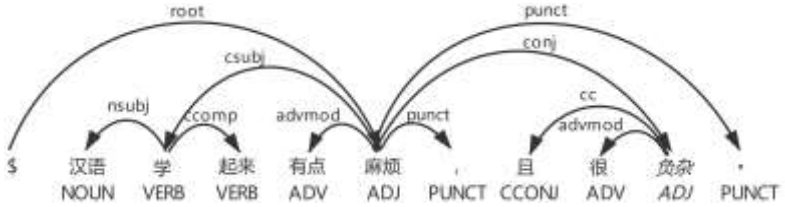

图 4-5 音近示例

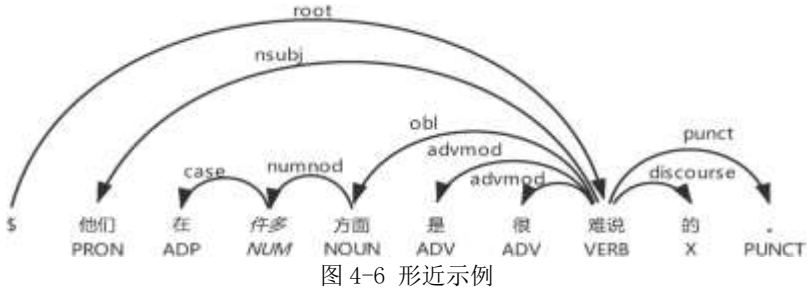

图 4-6 形近示例

具有相同的语素，指的是偏误和目标句有一个或多个相同语素。如图 4-7，“了”是“了理”和“了解”的共有语素，但“了理”是不合法且无法从字面获取有效信息、判断其句法结构的语言现象，我们同样根据目标句进行标注。

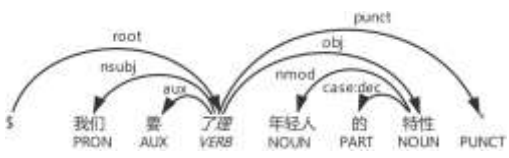

图 4-7 具有相同的语素示例

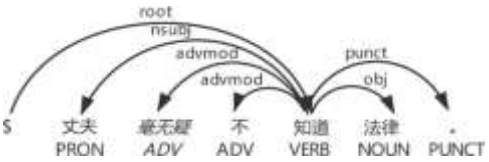

图 4-8 成语成分的缺失或赘余示例

成语成分的缺失或赘余是指二语学习者在书写过程中将成语的某些成分漏写或多写，如图 4-8，“毫无疑问”写成了“毫无疑”，但是仍然表达原词的意思。我们采取同样的标注方法。

4.1.2 可以判断句法结构，但不符合语法

可以判断句法结构但其句法结构不符合语法，指的是我们可以通过语言现象判断其句法结构，但是由于用词错误或句式杂糅等导致其不符合语法规则，我们根据目标句进行词性标注和依存分析。例如，不及物动词用作及物动词，如图 4-9 所示。“和解”是一个不及物动词，后面不能带宾语，但是可以判断这句话是一个带宾语的谓语句。所以，我们规定：根据目标句进行标注，即把“和解”当作“化解”，“难题”作为“化解”的宾语，标为 obj。

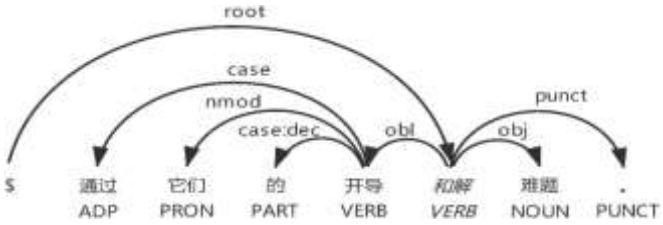

图 4-9 不及物动词用作及物动词示例

4.2 对语法上具有可解释性中介语，根据所观察到的语言现象进行词性标注和依存句法分析

在语法上具有可解释性，是指能够从所观察到的语言现象中获取有效信息来判断句法结构，并且句法结构符合语法规则。此时，我们根据所观察到的语言现象进行词性标注和依存句法分析。

如图 4-10 所示，学习者将“挤”写成了“偷”，但“偷”在语法上具有可解释性，“偷”和“挤”具有相同的句法环境，但是语义环境不同。我们就根据所观察到的语言现象进行词性标注和依存句法分析。“我”是“偷”的名词性主语，标为 nsubj；“过去”作为“偷”的补充成分，标为 xcomp。

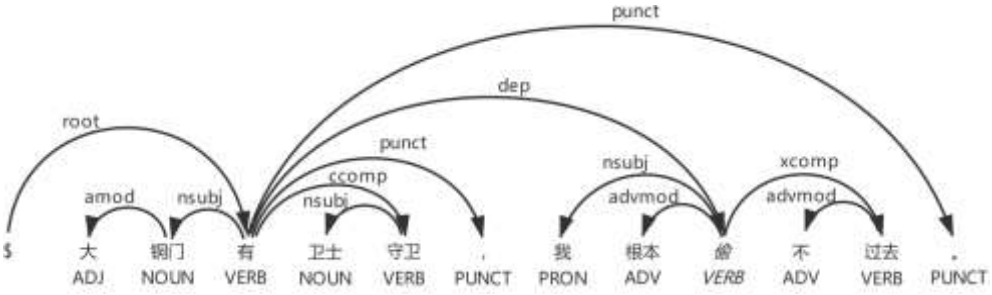

图 4-10 句法环境相同，语义环境不同示例

如图 4-11 所示，定中结构的中心语冗余，使得 “忆起”和“回忆”在语义上不能搭配，但是合乎语法规则。因此，“回忆”作为“忆起”的宾语，标为 obj；“童年”作为名词修饰语修饰多余的中心语“回忆”，标为 nmod。

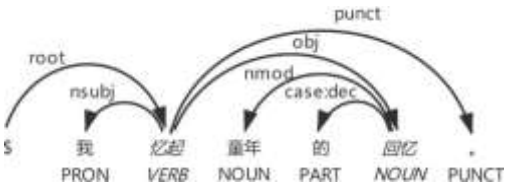

图 4-11 定中结构中心语冗余示例

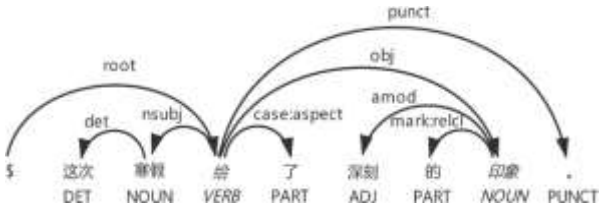

图 4-12 双宾句缺少直接宾语示例

双宾句缺少直接宾语，我们规定：将双宾句看作单宾句。如图 4-12 所示，“印象”标记为“给”的 obj 而不是 iobj。

5 结语

制定面向汉语中介语的依存句法标注规范对于汉语国际教育以及构建服务于自然语言处理领域的语言资源来说具有十分重要的意义。本文在 HSK 动态作文语料库提供的学习者作文文本基础上，制定了面向汉语中介语的标注规范，并抽样一部分语料进行人工标注，所有语料都保持了原始面貌，没有经过任何处理。本文的创新工作如下：

（一）在充分借鉴 UD V2 的基础上制定了面向汉语中介语的标注规范，包括标注框架和标注原则；（二）结合 UD V2 提出了更能够刻画汉语句法结构的依存句法标注规范。在词性上，保留了 UD V2 的 16 个词性标签，并对 UD V2 中没有说明的汉语特殊词性现象做出了特殊的处理方式。在依存关系标签上，首先根据汉语的结构特点删掉了不符合汉语的 8 个主类标签，增加了 8 个次类标签；其次对所有标签根据汉语的句法理论体系分为三大类，包括单句主干关系标签、单句其他关系标签、嵌套关系标签，使之体系化；最后，针对汉语中的独特结构，提出了特殊的标注策略，以便更加充分地刻画汉语句法结构。（三）提出了面向汉语中介语的依存标注原则，即：在语法上具有不可解释性的语言现象，根据偏误纠正后获得的目标句进行词性标注和依存句法分析；在语法上具有可解释性的语言现象，根据句法结构对其进行词性标注和依存句法分析。

参考文献

[1].鲁健骥. 中介语理论与外国人学习汉语的语音偏误分析[J]. 语言教学与研究，1984(3):44-56.
[2].李娟,谭晓平,杨丽姣. 汉语中介语语料库应用及发展对策研究[J]. 曲靖师范学院学报，2016(2):86-91.
[3].李正华. 汉语依存句法分析关键技术研究[D]. 哈尔滨: 哈尔滨工业大学博士学位论文，2013.
[4].Nivre,Joakim, Marie-Catherine de Marneffe,et al. Universal Dependencies v1: A Multilingual Treebank Collection[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation. LREC,2016:1659-1666.
[5].John Lee,Herman Leung and Keying Li. Towards Universal Dependencies for Learner Chinese[C]// Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies. 2017:67-71.
[6].Marwa Ragheb and Markus Dickinson. Developing a Corpus of Syntactically-Annotated Learner Language for English[C]// In Proc.13th International Workshop on Treebanks and Linguistic Theories (TLT). 2014.
[7].Geoffrey Sampson. English for the Computer: The SUSANNE Corpus and Analytic Scheme[M]. UK: Clarendon Press，1995.
[8].MacWhinney B. The CHILDES system[J]. American Journal of Speech-Language Pathology, 1996, 5(1): 5-14. [9].Yevgeni Berzak,Jessica Kenney, et al. Universal
[9].Yevgeni Berzak,Jessica Kenney, Carolyn Spadine,et al. Universal Dependencies for Learner English[C]// Annual Meeting of the Association for Computational Linguistics. 2016.
[10].Nicholls D. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT[C]// Proceedings of the Corpus Linguistics 2003 conference. 2003(16): 572-581.
[11].Herman Leung,Rafaël Poiret,Tak-sum Wong,et al. Developing Universal Dependencies for Mandarin Chinese[C]// Proceedings of the 12th Workshop on Asian Language Resources (ALR12). 2016:20-29.

[12].张斌. 现代汉语描写语法[M]. 北京: 商务印书馆，2005.
[13].刘丹青. 汉语中的框式介词[J]. 当代语言学, 2002(4):241-253.
[14].黄伯荣,廖旭东. 现代汉语（增订四版）[M]. 北京: 高等教育出版社，2007.
[15].Gerdes,Kim,and Sylvain Kahane. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies[C]// Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016. 2016: 131.

附 录

（一）单句主干关系标签

单句主干关系标签主要用于标注由不同词性的词或短语充当汉语句子中的主谓宾定状补等结构。

nsubj（名词性主语）：名词性主语是谓语的陈述对象,依附于谓语。

nsubj:pass（受事主语）：在被动句中，受事主语是谓语的受事，依附于谓语。

obj（宾语）：宾语是谓语的实施或受事，通常位于谓语的后面。

iobj（间接宾语）：“送、给、授予、称呼、叫”等动词后面可以跟两个名词性宾语，为了区分，第二个宾语称为间接宾语，依附于谓语核心。

obl（状语，介宾短语）：由介宾短语构成的状语，依附于中心语。

advmod（状语，副词）：副词作状语，修饰谓词性成分，依附于中心语。

nmod（定语，体词性成分）：体词性成分做定语修饰名词性成分，依附于中心语。

amod（定语，形容词性成分）形容词性成分作定语修饰名词性成分，依附于中心语。

nummod（定语，数词）：数词作定语修饰名词性成分，依附于中心语。

aux（状语，能愿动词）：能愿动词做状语，修饰谓词性成分，依附于中心语。

det（定语，限定性成分）：限定性成分作定语，修饰名词性成分，依附于中心语。

clf（定语，量词短语）：量词短语作定语，修饰名词性成分，依附于中心语。

conj（并列）：两个成分并列，由第一个成分指向第二个成分。

appos（同位）：表示一个体词性成分作另一个名词性成分的同位语，依附于本位语。

punct(标点)：标点符号以 punct 关系修饰核心词。

注：中补结构是汉语特有的现象，例如“跑得快”，在不增加主类标签的基础上，我们采用 xcomp 表示。

（二）单句其他关系标签

单句其他关系标签用于辅助标注汉语句子的其他结构关系。

vocative（称呼）：用于说话者对受话者的称呼，与句内其他成分不发生结构关系，一般由标点隔开，以 vocative 关系修饰谓语。

dislocated（异位）：表示某个句法成分脱离其正常的句法位置，位于它所处句子的前面或者后面，以 dislocated 关系修饰谓语。

discourse（话语）：表示语气词、叹词等在句中与其他成分之间的关系，以 discourse 关系修饰其他句法成分。

mark（关联连词）：表示关联连词与连接的两个小句之间的关系。关联连词起连接作用并表示一定语气，以 mark 关系修饰谓语核心。

mark:advb（“地”字标记）：出现在状中结构中，“地”用来标记状语和中心语的关系。

mark:comp（“得”字标记）：出现在中补结构中，“得”用来标记中心语和补语的关系。

mark:relcl（“的”字标记）：出现在谓词做定语的定中结构中，其中“的”用来标记定语和中心语的关系。

case（介宾关系）：介词和名词性成分构成介宾短语，介词以 case 关系来修饰宾语。

case:aspect（时态标记）：时态标记表示动作或变化的状态，通常位于动词或形容词之后，一般体现为动态助词“着”、“了”、“过”等与动词或形容词之间的关系。

case:loc（方位词标记）：在介宾结构中，依附于名词性成分之后，构成框式结构。

case:dec（定中短语“的”标记）：出现在名词性成分做限定性定语的定中结构中，其中“的”、“之”是定中关系的标记。

cc（并列连词）：表示“和”、“与”等并列连词起连接作用，我们规定“和”、“与”等连词以 cc 关系修饰后一个并列成分。

flat（连接）：用于表示无法分析或者不需要分析的句法结构内部之间的关系。

orphan（省略）：用于表示动词从前省略现象。

（三）嵌套关系标签

嵌套关系标签主要用于标注汉语中的嵌套结构。

csubj（主从）：表示含动词的嵌套结构做主语，相当于主语从句。

csubj:pass（嵌套结构充当受事主语）：在被动句中，嵌套的复杂结构做主语。

xcomp（宾从，同主语）：表示含动词的嵌套结构作谓语的补充成分，嵌套结构的主语与句子的主语不一致。

ccomp（宾从，不同主语）：表示含动词的嵌套结构作谓语的补充成分，嵌套结构的主语与句子的主语一致。

acl（定从）：表示含动词的嵌套结构作定语修饰名词性成分，相当于定语从句。

advcl（状从）：表示含动词的嵌套结构作状语修饰谓词性成分，相当于状语从句。

dep（不确定关系）：表示无法确定的关系时，标为 dep，dep 要尽可能少用。

作者联系方式：
肖丹　北京市海淀区学院路 15 号 100083 18810511621 1273450007@qq.com
杨尔弘 北京市海淀区学院路 15 号 100083 13661127882 yerhong@126.com
张明慧 北京市海淀区学院路 15 号 100083 18813161625 zmh19960206@126.com
陆天荧 北京市海淀区学院路 15 号 100083 19801255265 ltytianyinglu@163.com
杨麟儿 北京市海淀区学院路 15 号 100083 13426012083 lineryang@gmail.com