**文章编号**：1003-0077(2020)11-0019-10

# 汉语中介语的依存句法标注规范及标注实践

肖 丹[1,2]，杨尔弘[1,2]，张明慧[1,2]，陆天荧[1,2]，杨麟儿[1,2]

(1. 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；
2. 北京语言大学 语言资源高精尖创新中心，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。依存语法分析是语言信息处理的重要步骤，英语中介语的依存语法标注语料已经有很好的应用，目前汉语中介语语料库对句法的关注度较低，缺乏一个充分考虑汉语中介语特点的依存句法标注规范。该文着眼于汉语中介语的依存句法标注语料库的建构，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，制定了汉语中介语的依存标注规范，并进行了标注实践，形成了一个包括汉语教学语法点的中介语依存语料库。

**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], YANG Erhong[1,2], ZHANG Minghui[1,2], LU Tianying[1,2], YANG Liner[1,2]

(1. National Language Monitoring and Research Center(CNLR) Print Media Language Branch, Beijing Language and Culture University, Beijing 100083, China;
2. Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China)

**Abstract**: Chinese inter-language is accompanied by Chinese international education. With growing development of Chinese language learning in the world, the scale of inter-language 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. Compared with inter-language in English with dependency grammar annotation corpus, the current Chinese inter-language corpora even have no annotation guideline for dependency syntax. Aiming to construct the corpus of inter-language dependency annotation in Chinese, this paper, develops a new dependency annotation guideline for Chinese inter-language based on the Universal Dependencies. And a corpus of Chinese inter-language annotated with dependency sturucture is finally achieved with consideration of its characteristics.

**Keywords**: inter-language; dependency grammar; annotation guideline

## 0 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，又区别于其所学的目的语，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还不够[2]。

**收稿日期**：2019-09-09 **定稿日期**：2019-10-09

**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)；国家语言资源监测与研究平面媒体中心研究经费

句法分析是自然语言处理的重要基础任务，对中介语的自动句法分析的关注度近年来不断增加，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其首要任务。

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 Dependencies①，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]。截至目前，最新版本的UD V2.6已发布了92种语言的标注数据，共163个树库。香港城市大学公开发布了基于UD的汉语中介语依存句法规范[5]。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标注过程中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。

## 1 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具有重要意义。目前国外的学习者树库构建已相对完善，但汉语中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，SALLE)是由Ragheb和Dickinson[6]构建的学习者树库，语料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，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(Treebank of Learner English②，TLE)，语料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实证语料，促进了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③、中山大学汉字偏误标注的汉语连续性中介语语料库④、暨南大学留学生汉语书面语语料库⑤等，虽然这些语料库为汉语国际教育做出过重要贡献，但是在句法方面仍稍显薄弱。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-CFL⑥。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这一情况，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做考虑。从标注对象的角度，UD_Chinese-CFL对语料进行了一定程度的修

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

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

③ 北京语言大学HSK动态作文语料库：http://202.112.195.192:8060/hsk/login.asp

④ 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/

⑤ 暨南大学留学生汉语书面语语料库：http://huayu.jnu.edu.cn/corpus3/Search.aspx

⑥ UD_Chinese-CFL的网址：https://universaldependencies.org/treebanks/zh_cfl/index.html

改，而且仅继承了“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，本文在借鉴 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。

## 2　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-GSD①、UD_Chinese-PUD②、UD _ Chinese _ HK③ 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，但仍存在某些不足。主要表现在：①考虑了汉语中的特殊词性，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这一情况，而未考虑方位词的其他用法。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。④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，包括词性和依存标签。主要创新点为：①保留了 UD V2 的 16 个词性标签，并对 UD V2 中没有说明的汉语中的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 12 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。

### 2.1　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]中认为其具有黏着性，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，做出如下处理：当方位词处于框式介词结构中时，我们认为方位词黏着性较强，具有介词的性质，因此将方位词当作介词处理，标为 ADP，如图 1 所示。在其他情况下，方位词的名词属性较为强烈，我们仍把它当作名词处理，标注为 NOUN，如图 2 所示。

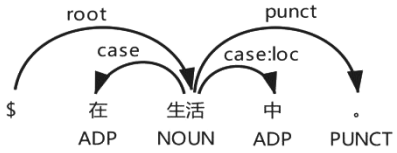


图 1　方位词处于框式结构中示例

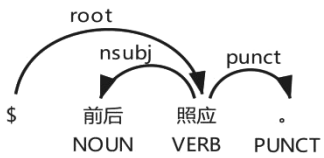


图 2　方位词处于非框式结构中示例

**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 所示，“一支笔”中的“支”标为 NOUN。

① UD_Chinese-GSD 的网址：https://universaldependencies.org/treebanks/zh_gsd/index.html

② UD_Chinese-PUD 的网址：https://universaldependencies.org/treebanks/zh_pud/index.html

③ UD_Chinese_HK 的网址：https://universaldependencies.org/treebanks/zh_hk/index.html

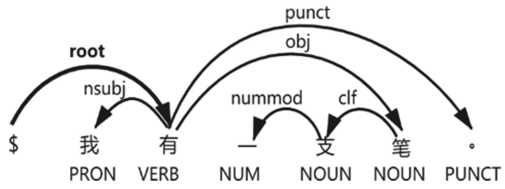


图3 量词标注示例

**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。

## 2.2 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12个次类标签，以便更好地描写汉语句法结构。删掉的标签有：expl(空语类)、fixed(固定结构)、compound(复合词)、list(列举)、parataxis(并列)、goeswith(多词连接)、reparandum(语音停顿的词)；增加的标签有：nsubj：pass(受事主语)、csubj：pass(嵌套结构充当受事主语)、mark：advb(“地”字标记)、mark：comp(“得”字标记)、mark：relcl(“的”字标记)、case：aspect(时态标记)、case：loc(方位词标记)、case：dec(定中短语“的”标记)、case：stuff(前缀)、case：pref(后缀)、xcomp：comp(补语)、dep：conj(并列小句)。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。体系化依存关系标签如表1所示。

**表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 | 句子 | 父亲很爱我们 | root<br>$→很爱 |
| | nsubj | 名词性主语 | 父亲很爱我们 | nsubj<br>父亲←很爱 |
| | nsubj:pass | 受事主语 | 观念也被带入了中国 | nsubj:pass<br>观念←带入 |
| | obj | 宾语 | 你们给我这个机会 | obj<br>给→机会 |
| | iobj | 间接宾语 | 你们给我这个机会 | iobj<br>给→我 |
| | obl | 状语，介宾短语 | 跟我约定 | obl<br>我←约定 |
| | advmod | 状语，副词 | 少喝酒 | advmod<br>少←喝酒 |
| | nmod | 定语，体词性成分 | 时代差异 | nmod<br>时代←差异 |
| | amod | 定语，形容词性成分 | 新的观念 | amod<br>新←观念 |
| | xcomp:comp | 补语 | 我吃了一会儿了 | xcomp:comp<br>吃→一会儿 |
| | appos | 同位 | 一本书《谁动了我的奶酪》 | appos<br>书←动 |

续表

| | 依存关系 | 说　明 | 例　句 | 标注示例 |
|---|---|---|---|---|
| 单句主干关系标签 | nummod | 定语，数词 | 一封信 | nummod<br>一←封 |
| | aux | 状语，能愿动词 | 能理解 | aux<br>能←理解 |
| | clf | 定语，量词短语 | 一封信 | clf<br>封←信 |
| | det | 定语，限定性成分 | 每个国家 | det<br>每个←国家 |
| | conj | 并列 | 生活与学习 | conj<br>生活→学习 |
| | punct | 标点 | 一封信。 | punct<br>信→。 |
| 单句其他关系标签 | vocative | 称呼 | 爸爸、妈妈：你们好 | vocative<br>爸爸←好 |
| | dislocated | 异位 | 出产的桃子人吃了 | dislocated<br>桃子←吃 |
| | discourse | 话语 | 这不行吧 | discourse<br>不行→吧 |
| | mark | 关联连词 | 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<br>如果←吃 |
| | mark:advb | “地”字标记 | 不断地增加 | mark:advb<br>不断→地 |
| | mark:comp | “得”字标记 | 家庭维持得很好 | mark:comp<br>维持→得 |
| | mark:relcl | “的”字标记 | 美好的事 | mark:relcl<br>美好→的 |
| | case | 介宾关系 | 在生活中 | case<br>在←生活 |
| | case:aspect | 时态标记 | 新观念也被带入了中国 | case:aspect<br>带入←了 |
| | case:loc | 方位词标记 | 在生活中 | case:loc<br>生活→中 |
| | case:dec | 定中短语“的”标记 | 外面的社会 | case:dec<br>外面→的 |
| | case:pref | 前缀 | 非正式 | case:pref<br>正式→非 |
| | case:suff | 后缀 | 我是个人主义者 | case:suff<br>者→个人主义 |
| | cc | 并列连词 | 生活与学习 | cc<br>与←学习 |
| | flat | 连接 | 弗洛伦斯·南丁格尔 | flat<br>弗洛伦斯→南丁格尔 |
| | orphan | 省略 | 他给奶奶一个手表，爷爷一个手表 | orphan<br>爷爷→手表 |

续表

| | 依存关系 | 说　明 | 例　句 | 标 注 示 例 |
|---|---|---|---|---|
| 嵌套关系标签 | csubj | 主从 | 学汉语是最好的 | csubj<br>学←最好 |
| | csubj:pass | 嵌套结构充当受事主语 | 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<br>是←应用 |
| | xcomp | 宾从，同主语 | 我去北京学习 | xcomp<br>去→学习 |
| | ccomp | 宾从，不同主语 | 我认为社会问题是教育方面的问题 | ccomp<br>认为→是 |
| | acl | 定从 | 我最尊敬的人 | acl<br>尊敬←人 |
| | advcl | 状从 | 他拍桌子保证 | advcl<br>拍←保证 |
| | dep | 小句关系 | 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 | dep<br>包括→广泛 |
| | dep:conj | 并列小句关系 | 我不仅喜欢芒果，而且喜欢香蕉 | dep:conj<br>喜欢→喜欢 |

### 2.3 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，本文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，作整个句子的谓语。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既没有关联词语，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，也不能把它当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 所示。

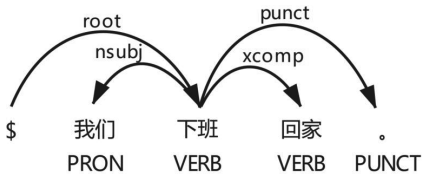


图 4 连谓句标注示例

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 5 所示。

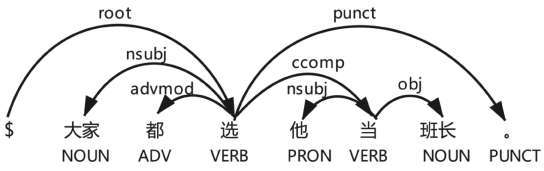


图 5 兼语句标注示例

“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(在这种情况下，“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须考虑上下文语境就可以省略时，“的”为结构助词，标为 obj，如图 6 所示；当“的”字短语的中心语在无上下文语境中不可以省略时，“的”为语气词，标为 discourse，如图 7 所示。

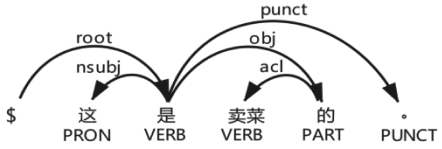


图 6 “是……的”示例一

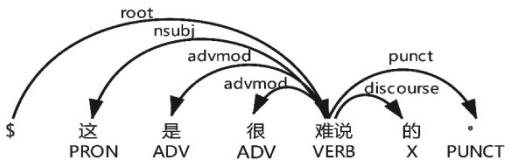


图 7　“是……的”示例二

## 3　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造成误解。②TLE 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。③数据的适应性不同。通过观察数据，我们发现英语二语学习者在核心动词方面犯的错误往往是动词时态错误；而汉语二语学习者在核心动词方面犯的较多的错误则是缺失，因而导致标注者在执行“字面标注”的原则的时候随意性较大。

因此，我们在充分吸取前人研究成果的基础上，针对上述问题提出了更为准确、细致的标注原则，即：

（1）核心标注原则

我们规定：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，这为核心标注原则。顾名思义，核心标注原则就是最重要的标注原则，即在拿到一个偏误句时，首先要尽量地根据偏误纠正后获得的目标句的句法结构进行偏误句的依存句法标注。

（2）非核心标注原则

我们规定：根据所观察到的句法结构对其进行词性标注和依存句法分析，这为非核心标注原则。非核心标注原则不是最重要的标注原则，即在拿到一个偏误句时，当用核心标注原则解决不了时，就根据非核心原则进行标注。

### 3.1　核心标注原则

核心标注原则即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。通过实际的标注，我们发现核心标注原则能够处理大部分的情况，主要包括无法判断句法结构、句法结构不符合语法、遗漏的单位不对句法结构产生影响等情况。

#### 3.1.1　无法判断句法结构

无法判断句法结构，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等。对此，我们根据核心标注原则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 8、图 9 所示。“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。

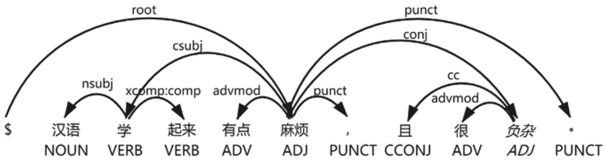


图 8　音近示例

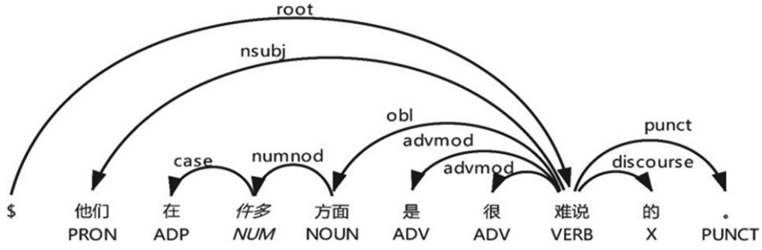


图 9　形近示例

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，如在图 10 中，“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。

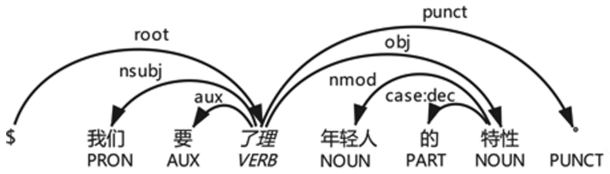


图 10 具有相同的语素示例

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 11 中，“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。

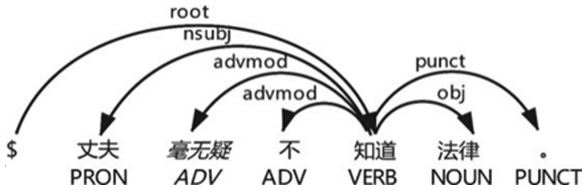


图 11 成语成分的缺失或赘余示例

3.1.2 句法结构不符合语法

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 12 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。

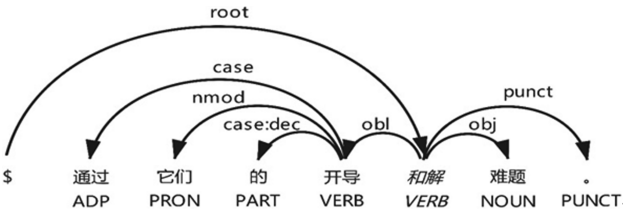


图 12 不及物动词用作及物动词示例

3.1.3 遗漏的单位不对句法结构产生影响

遗漏或误加的单位不对句法结构产生影响指的是某个大的单位在缺失或误加某个小的单位的情况下，大的语言单位的句法属性仍然未发生改变。以图 13 为例，“在那里”这个大的语言单位在整个句子中作状语，而遗漏了介词“在”的“那里”在整个句子中仍然作状语，所以根据目标句对偏误句进行依存句法标注。

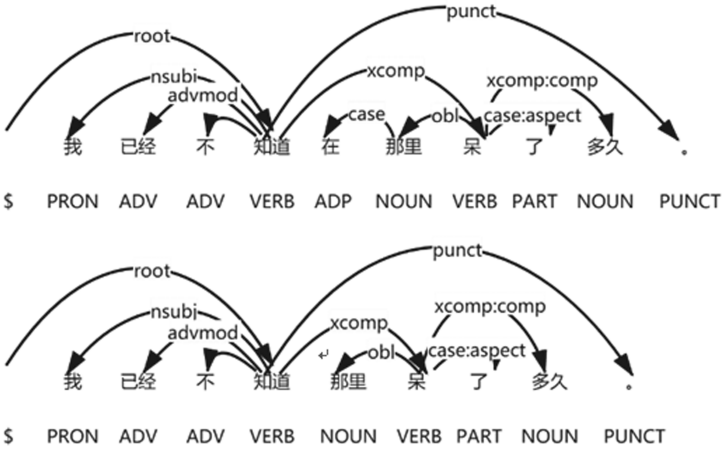


图 13 遗漏示例

## 3.2 非核心标注原则

非核心标注原则即根据所观察到的句法结构对其进行词性标注和依存句法分析。通常是在核心标注原则无法解决问题时，我们采用非核心标注原则。

如图 14 所示，如果按照核心标注原则，偏误句的 root(根节点)应该放在“深入”上。但是，如果这样的话就没有办法处理“达到”。在我们的标注体系中，表示两个动词之间关系的标签有“conj(并列)”和“xcomp(宾从)”，但是这两个标签的方向都是从第一个词指向第二个词。所以，如果按照核心标注

原则，就和 3.1 节所制定的依存标注框架相违背，因此需要根据非核心标注原则来处理偏误句，root 应该放在“达到”而非“深入”上。

由于我们是通过偏误标注得到的目标句，所以也存在一些因为偏误标注而使得我们采取非核心标注原则的情况。在基础的偏误标注阶段，标注员在标注时并没有严格遵循相关的偏误标注原则，而使得目标句和偏误句出现了不一样的句法结构。如在图 15 中，标注者在标注的过程中应该针对“是字句”的搭配不当和遗漏中心语进行修改，但是标注者直接将“是字句”改成了“是……的”句，由此造成了偏误句和目标句的句法结构存在较大的差异。

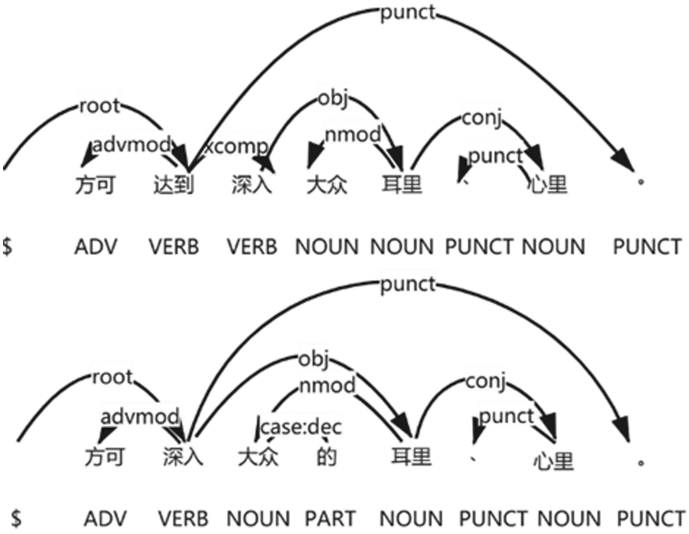


图 14　非核心标注原则示例一

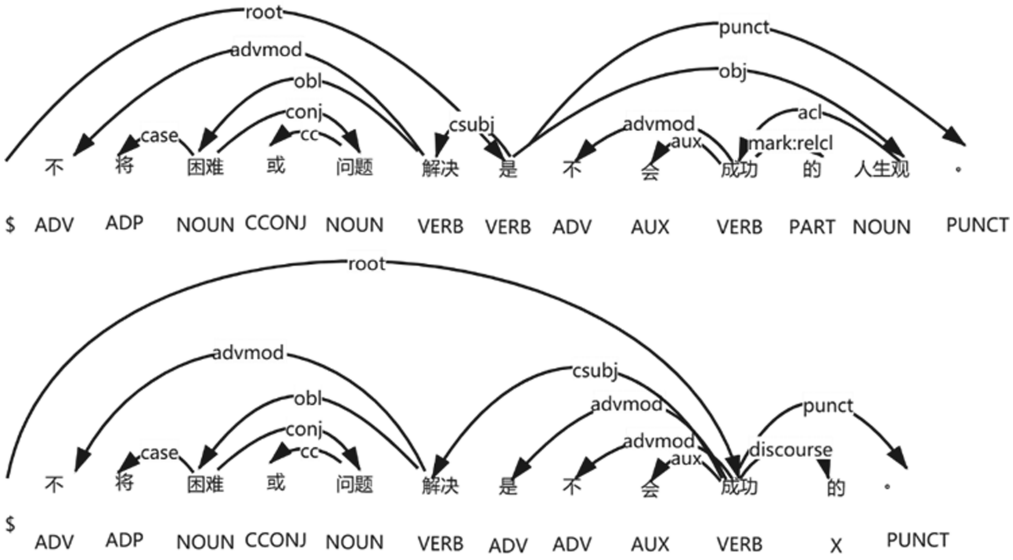


图 15　非核心标注原则示例二

## 4　标注实践

考虑到汉语中介语包含偏误的特性，存在机器没有办法切分的词，比如“仆多”，因此，我们进行两个层面上的标注，即：词层面的分词和词性标注、句法层面的依存标注。在分词和词性标注完成之后再进行依存句法标注，可有效提高标注质量，但都增加了人工成本，考虑到中介语句法标注的难度，我们采取人机结合的标注方式，基于 Arborator[①] 开发了一个在线标注平台[②]，设置了标注员和审核员两种角

① Arborator：https://github.com/Arborator/arborator-server11 https://yatlc.wenmind.net

② http://yatlc.wenmind.net

色。该平台操作简单,可以多人同时标注;标注员可在标注平台上进行分词修改、词性标签修改、依存弧以及标签标注等一系列操作;审核员能够对比多人标注的结果,极大减少了工作量。标注流程如图16所示。

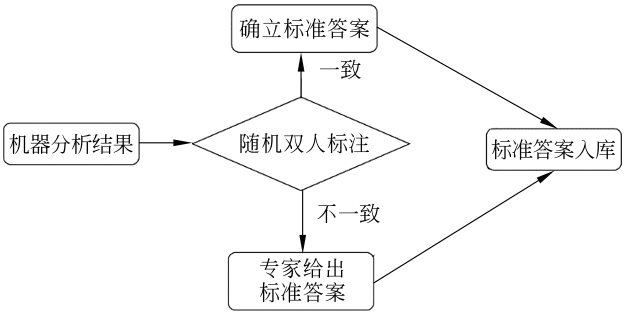


图16 标注流程示意图

我们采用人机结合的方式进行标注,先用训练好的模型对语料进行分析,给出分析结果,然后随机分配给两个标注员进行标注。标注完成后,如果两个标注结果完全一致,那么就直接认为是标准答案,答案入库,流程结束。如果两个标注结果不一致,将由专家进行审核,确定唯一答案,答案入库。

我们从HSK动态作文语料库中,依据平衡性和综合性原则抽取了9 626个句对,共覆盖了29个话题,14种语言,并招募了北京语言大学语言及应用语言学相关专业的硕士研究生作为标注人员。我们对标注人员进行了一定程度上的培训:首先,向标注人员详细介绍我们的标注规范以及标注系统的用法;然后,标注人员学习标注规范并在我们的系统上进行标注尝试;学习一段时间后,我们对标注人员进行考核,通过考核的标注人员进行真实数据的标注工作。截至目前,我们基于上述标注流程已经标注了100句对,未来我们将按照该规范持续标注汉语中介语。

## 5 总结与展望

本文在充分借鉴UD V2的基础上,制定了一个面向汉语中介语的依存句法标注规范,以指导大规模的数据标注工作。该规范充分考虑到了汉语的特点和中介语的特殊性,对汉语中的特殊现象和结构提出了相应的标注策略,同时针对中介语含有偏误的特性,提出了核心标注原则和非核心标注原则。

未来,我们会按照该规范继续标注更多的汉语学习者文本,提高依存句法分析的准确率,也为语法改错、句法复杂度等相关研究工作提供相应的数据支持。随着标注工作的不断展开,我们会将所遇到的问题进行总结,以不断完善标注规范。

## 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] Joakim Nivre, Marie Catherine de Marneffe, Filip Ginter, et al. Universal dependencies v1: A multilingual treebank collection[C]//Proceedings of the 10th International Conference on Language Resources and Evaluation. LREC,2016:1659-1666.

[5] John Lee,Herman Leung, Keying Li. Towards universal dependencies for learner Chinese[C]//Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 2017:67-71.

[6] Marwa Ragheb, Markus Dickinson. Developing a corpus of syntactically-annotated learner language for English[C]//Proceedings of the 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] Brian MacWhinney. The CHILDES system[J]. American Journal of Speech-Language Pathology, 1996, 5(1):5-14.

[9] Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, et al. Universal dependencies for learner English[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics, 2016.

[10] Diane Nicholls. 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.

(下转第36页)

[11] Auer S, Lehmann J, Hellmann S. Linkedgeodata: Adding a spatial dimension to the web of data[C]// Proceedings of the International Semantic Web Conference. Springer, Berlin, Heidelberg, 2009: 731-746.

[12] Ballatore A, Wilson D C. Geographic knowledge extraction and semantic similarity in OpenStreetMap [J]. Knowledge & Information Systems, 2013, 37 (1): 61-81.

[13] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194: 28-61.

[14] Hu W, Li H, Sun Z, et al. Clinga: Bringing Chinese physical and human geography in linked open data [C]//Proceedings of the International Semantic Web Conference. Springer, Cham, 2016: 104-112.

[15] Maxj Egenhofer, Robertd Franzosa. Point-set topological spatial relations[J]. International Journal of Geographical Information Systems, 1991, 5(2): 161-174.

[16] 马雷雷. 空间关系本体描述与推理机制研究 [D],郑州: 信息工程大学硕士学位论文, 2012.

[17] Clementini E, Sharma J, Egenhofer M J. Modelling topological spatial relations: Strategies for query processing[J]. Computers & Graphics, 1994, 18(6): 815-822.

[18] Goyal R K. Similarity assessment for cardinal directions between extended spatial objects[D]. Ph D Thesis, Maine: The University of Maine, 2000.

[19] 王东旭, 诸云强, 潘鹏, 等. 地理数据空间本体构建及其在数据检索中的应用[J]. 地球信息科学学报, 2016, 18(4): 443-452.

[20] 张雪英, 张春菊, 杜超利. 空间关系词汇与地理实体要素类型的语义约束关系构建方法[J]. 武汉大学学报(信息科学版), 2012, 37(11): 1266-1270.

刘俊楠(1991—),博士研究生,主要研究领域为时空数据挖掘与知识图谱相关研究。
E-mail: 6929423@qq.com

陈晓慧(1983—),博士,副教授,主要研究领域为时空数据可视化。
E-mail: cxh_vrlab@163.com

刘海砚(1970—),通信作者,博士,教授,主要研究领域为时空数据挖掘相关研究。
E-mail: liuharry2020@163.com

(上接第 28 页)

[15] Gerdes Kim, 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.

肖丹(1995—),硕士研究生,主要研究领域为自然语言处理。
E-mail: 1273450007@qq.com

张明慧(1996—),硕士研究生,主要研究领域为自然语言处理。
E-mail: zmh19960206@126.com

杨尔弘(1965—),通信作者,博士,教授,主要研究领域为自然语言处理。
E-mail: yerhong@126.com