汉语学习者依存句法树库构建

师佳璐， 罗昕宇， 杨麟儿， 肖丹， 胡正升，
王一君， 袁佳欣， 余婧思， 杨尔弘
北京语言大学，北京 100083

摘要

汉语学习者依存句法树库为非母语者语料提供依存句法分析，可以支持第二语言教学与研究，也对面向第二语言的句法分析、语法改错等相关研究具有重要意义。然而，现有的汉语学习者依存句法树库数量较少，且在标注方面仍存在一些问题。为此，本文改进依存句法标注规范，搭建在线标注平台，并开展汉语学习者依存句法标注。本文重点介绍了数据选取、标注流程等问题，并对标注结果进行质量分析，探索二语偏误对标注质量与句法分析的影响。

关键词： 汉语学习者 ； 依存句法树库 ； 语料标注 ； 偏误分析 ； 依存句法分析

Construction of a Treebank of Learner Chinese

SHI Jialu, LUO Xinyu, YANG Liner, XIAO Dan, HU Zhengsheng,
WANG Yijun, YUAN Jiaxin, YU Jingsi, YANG Erhong
Beijing Language and Culture University, Beijing 100083, China

Abstract

A dependency treebank of learner Chinese provides parse trees of non-native sentences, which could promote the teaching and researching of Chinese as a second language, as well as support related researches such as syntactic analysis of learner language and grammatical error correction. However, few treebank of learner Chinese has to be seen, and there are still some problems in annotation guideline. So far, we improve the annotation guideline, develop an online annotation platform, and build the Treebank of Learner Chinese. This thesis describes the details in data selection and annotation workflow, evaluates the quality of annotation, and explores the impact of errors on annotation quality and syntactic analysis.

Keywords: Chinese learners , dependency treebank , data annotation , error analysis , dependency analysis

1 引言

基金项目：北京语言大学语言资源高精尖创新中心项目(TYZ19005)；国家语委信息化项目(ZDI135-105)；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）项目成果(20YCX141)

树库作为一种记录每个句子句法分析结果的标注语料库(黄昌宁and 靳光瑾, 2013)，融合了分词、词性、句法等各种信息。一方面能为语言学、句法学研究提供实例，另一方面也能为句法分析、机器翻译、语法改错等自然语言处理领域的相关任务提供训练数据。与其他类型树库相比，依存句法树库具有以下优点：1）存储空间小，便于大规模储存；2）依存句法的形式简洁，易于理解(刘挺and 马金山, 2009)，更加适合人工标注；3）更加突出中心词的地位，侧重于反映语义关系，有助于语义角色标注、信息抽取等上层应用(刘挺and 马金山, 2009)；4）依存句法在表示交叉关系时更有优势(刘挺and 马金山, 2009)，特别适合于分析语序灵活的汉语(郭丽娟, 2019)。

汉语学习者语料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，语料的规模也不断增长，所构建的汉语学习者语料库也越来越多。汉语学习者语料与一般语料相比有其独特性，包含着大量的偏误，即中介语与目的语规律之间的差距，只有学习外语的人才会产生(鲁健骥, 1984)。正是由于这些语料在语言使用上的独特性，使得汉语学习者语料成为语言信息处理和智能语言辅助学习领域的独特资源。目前的汉语学习者语料库在“字”、“词”的偏误标注上较为深入，但对句法结构的关注度不够(李娟et al., 2016)。

香港城市大学构建了汉语学习者依存句法树库[0]（UD_Chinese-CFL）。它在一定程度上弥补了现有汉语学习者语料库的不足，但是该树库对汉语特殊词性和句式考虑不够周全，标注标签种类过多，且未充分考虑学习者语料中的偏误对标注原则和标注结果的影响。

鉴于此，肖丹et al. (2019)制定了面向汉语中介语的依存句法标注规范，并考虑到了汉语特殊词性、句法结构和汉语中介语的特性等问题，然而该标注规范在标注原则和依存关系标签上仍需要进一步完善。本文对该标注规范进行了改进，使之更符合汉语及汉语学习者语料的特点，并搭建了在线标注平台，对含有偏误句（汉语学习者原始语料）和目标句（纠偏后的句子）的平行句对进行标注，初步构建了汉语学习者依存句法树库，并以此探讨偏误对依存句法的影响。

2 相关研究工作

学习者语料库是第二语言或外语学习者产生的语言数据的电子文本库(Granger, 2012)。它记录的是非母语学习者在使用目的语的过程中产生的语言。这种语言既不同于母语，也不同于目的语，并且带有语法偏误信息。构建带有显性句法信息的学习者语料库对自然语言处理领域具有重要意义，能够为语法改错、句法复杂性研究等任务提供帮助。目前，国外学习者依存句法树库的建设工作已逐渐走向成熟，而相比之下国内汉语学习者依存句法树库建设的相关研究尚处于起步阶段，进展相对缓慢。

英语学习者依存句法树库发展迅速，最具规模。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)(Ragheb and Dickinson, 2014) 和英语学习者树库（The Treebank of Learner English，以下简称TLE）(Berzak et al., 2016)是英语上影响力最大的两个学习者树库。

SALLE是学习者树库的先驱，由Ragheb和Dickinson等人于2014年构建。它在SUSANNE Corpus(Sampson, 2011)的词性标签集和儿童语言数据交流系统（Child Language Data Exchange System）(MacWhinney, 2014)的依存标签集基础上构建标注规范，对大学生的英语作文语料进行标注。SALLE关注到了句子的表层结构，对推进句法标注在学习者语料库中的发展具有重要意义。但由于SALLE只对学习者语料表现出的语言学特征进行标注，而没有进行偏误标注和修改，因而难以应用到语法错误识别、语法改错等任务中。

TLE是2016年由Berzak等人构建的英语学习者树库。与SALLE相比，它有两大特征：1）在国际通用依存标注体系（Universal Dependencies，以下简称UD）[1] (de Marneffe et al., 2014; Nivre et al., 2016)之下建立标注规范，对多语言对比分析具有极大意义。2）对语法错误进行标注，便于应用到自然语言处理领域之中。UD是目前拥有语言种类最多的通用依存标注体系，它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(Nivre et al., 2016)。截止目前，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。TLE结合英语学习者的语言特征，对英语母语者语料标注规范进行一定修订，形成了基于UD的英语学习者语料标注规范。TLE的语料来源于剑桥学习者语料库（Cambridge First

[0]UD_Chinese-CFL：https://universaldependencies.org/treebanks/zh_cfl/
[1]UD V2：https://universaldependencies.org/guidelines.html

Certificate in English learner corpus）(Yannakoudakis et al., 2011)。TLE对语法错误进行了标注，在一定程度上弥补了SALLE缺乏偏误标注的问题。TLE的构建影响很大：在二语习得领域，它为偏误分析研究提供了语料支持，促进了第二语言教学与量化研究的发展；在自然语言处理领域，TLE为句法分析器提供了大量的训练语料，并通过实验验证了基于L1和L2的平行依存句法树库对提升句法分析器准确率的影响。

与此同时，汉语学习者树库的构建尚在初探过程中。北京语言大学HSK动态作文语料库[2](张宝林, 2009; 张宝林, 2010) 、全球汉语中介语语料库[3] (张宝林and 崔希亮, 2013)等汉语学习者语料库的关注重心主要在字、词等层面的偏误标注上，而对句法结构信息关注度不够。

已有的汉语学习者依存句法树库构建工作也有待完善。香港城市大学制定了面向汉语学习者的标注规范(Lee et al., 2017)，并构建了汉语学习者依存句法树库UD_Chinese-CFL。CFL遵循TLE“字面标注”的原则，结合汉语学习者语料特点，提出了基于UD的汉语学习者依存句法标注规范。但它也存在一些不足之处：标注原则不够清晰，将许多难以满足标注原则的语料当作例外情况处理；标注过程为了适应规范对语料做了一定程度的修改；对于一些汉语的特殊词性、结构未作详尽考虑。

为解决以上问题，肖丹et al. (2019)基于UD V2(Nivre et al., 2020)提出了一个新的面向汉语中介语的依存句法标注规范，考虑了汉语特殊结构的标注方法，并进一步细化了标注原则。然而，在应用该规范对学习者语料进行标注时，我们发现该规范的适应性不够强，在标注原则和依存关系标签上仍需进一步改进，以增强对汉语学习者语料的适应性。

3 汉语学习者依存句法树库构建

本章介绍了汉语学习者依存句法树库的构建工作。首先，我们对肖丹et al. (2019)制定的依存句法标注规范进行改进，提出了更适用于汉语学习者语料的标注规范。其次，我们从HSK动态作文语料库中筛选出带语法偏误信息的语料构建语料库。最后，我们搭建了在线标注平台，通过人机结合的方式，对分词、词性和依存句法进行标注。

3.1 标注规范

在使用肖丹et al. (2019)制定的标注规范进行标注的过程中，我们发现其存在以下问题：

1）从标注原则上看，面对汉语学习者语料的适应性不够强。它参考TLE制定的“字面标注”原则，未考虑到汉语学习者语料有不同于英语学习者语料的特点。如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上犯的较多的则是缺失错误。核心动词缺失会对句意理解造成很大影响，导致标注者在执行“字面标注”原则的时候，根据个人理解对原意进行判断，随意性较大。

2）从标注框架上来看，其框架未涵盖所有语言现象，导致一些汉语独特的语言结构只能与其他结构共用标签，难以区分，并且可能导致结构关系模糊不明。

针对以上问题，本文对该标注规范进行调整，提出了更细致、更符合汉语特点和汉语学习者语料特点的标注规范。制定标注原则时，充分考察汉语学习者语料，以期增强该规范对汉语学习者语料的适应性。现标注原则包括核心标注原则和非核心标注原则。

核心标注原则：根据纠偏后的目标句进行分词、词性标注和依存句法分析。核心标注原则是最重要的标注原则，即在拿到一个偏误句时，我们首先要尽量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。如图1，偏误句中“共供”为别字，同时遗漏了介词“对”，但大的语言单位的句法属性未发生改变，“公共场所的卫生”仍在句中作状语，对于这样的无法判断其句法结构、句法结构不合法或遗漏的单位不会使句法结构发生改变等情况，按照核心标注原则进行标注。

非核心标注原则：根据所观察到的句法结构进行分词、词性标注和依存句法分析。在核心标注原则不适用时，采用非核心标注原则。如图2，如果依照核心标注原则，偏误句中root应在“深入”上，但是，这样就没有办法处理“达到”。因为在我们的标注体系中，表示两个动词之间关系的标签有“conj”和“xcomp”，而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标注的话，就会和所制定的依存标注框架相违背，因此需要根据观

[2]北京语言大学HSK 动态作文语料库：http://hsk.blcu.edu.cn
[3]全球汉语中介语语料库：http://qqk.blcu.edu.cn/

察到的偏误句句法结构，将root放在“达到”上。对于此类句法结构发生改变或其他原因导致核心标注原则不适用的偏误句，按照非核心标注原则进行标注。

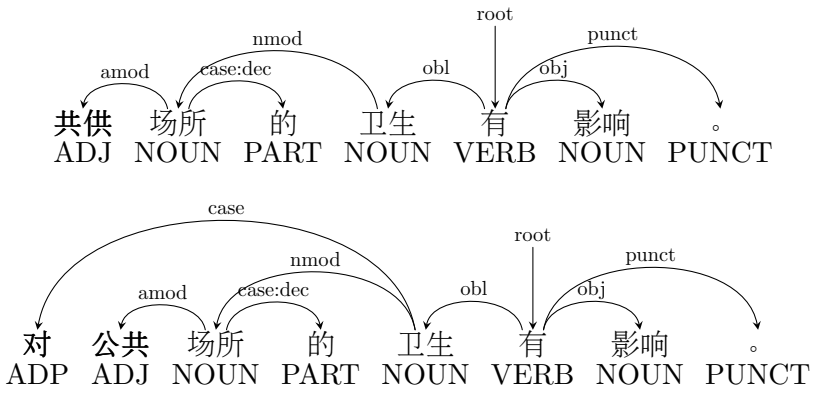

Figure 1: 核心标注原则示例

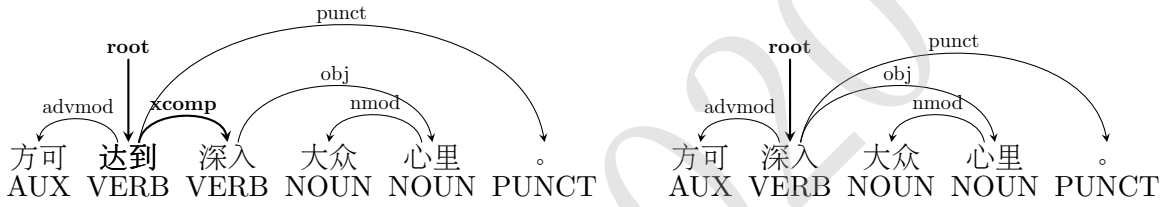

Figure 2: 非核心标注原则示例

制定依存关系标签时，充分借鉴UD体系，在标注语料的过程中弥补现存框架的不足，新增了2个标签类型。具体新增标签如下：

补语标签“xcomp:comp”

在原标注框架中，标签xcomp既用于联结相同主语的两个动词，又用于联结中补结构中的补语与中心语。为了加以区分，现规定xcomp仅用于联结相同主语的两个动词，新增标签xcomp:comp来联结中补结构中的补语与中心语，如图3。

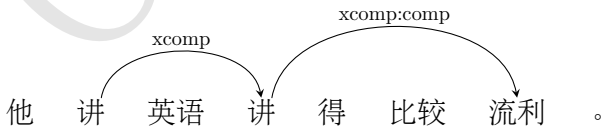

Figure 3: xcomp:comp用法示例

并列关系复句标签“dep:conj”

在原标注框架中，标签dep联结复句中的两个小句。然而当两个小句是并列关系时，就可能导致关系层次模糊不清。例如图4(a)与(b)的复句层次关系不同：(a)中的C小句与A、B小句分别构成并列关系，而(b)中的C小句与A小句非并列关系。但是它们的标签与依存弧方向完全一致，光看图难以区别这两种层次关系。为解决这一问题，为复句间的并列关系另设标签dep:conj，使得复句层次关系更加清晰，如图5。

3.2 数据选取

对汉语学习者语料进行依存句法分析，可以使我们直观地看到汉语学习者语料的，特别是存在偏误等语言现象的语料的句法结构。本文选取带有偏误信息的汉语学习者语料，通过对典型语料的分析窥探汉语学习者的语言特征，并期望对自动句法分析等研究有所帮助。

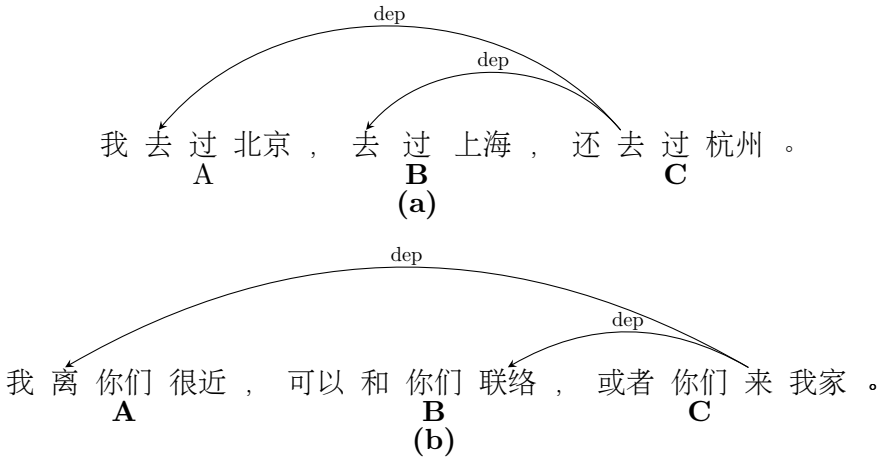

Figure 4: 原标注规范标注示例

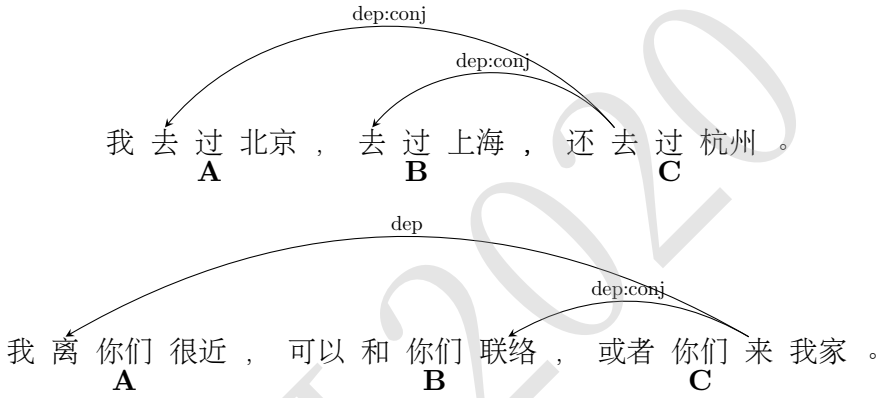

Figure 5: 现标注规范标注示例

首先，从北京语言大学构建的HSK动态作文语料库中选取了9626个带有偏误标注的句子。HSK动态作文语料库是目前国内最大的汉语中介语语料库，它总共收录了11569篇1992-2005年不同母语的留学生作文试卷的语料，并且对字、词、句和篇章进行了偏误标注和修改。为了保证语料的广泛性，以及抽取语料的平衡性和综合性，从包括叙事、议论、应用等不同文体的29个话题中均衡地抽取语料。选取语料的作者母语各异，包括韩语、日语、英语、俄语、法语、蒙古语等14种语言，遍布五大语系，含有孤立语、屈折语、黏着语三种语言形态类型，这可为语言迁移研究提供数据。在选取的篇章上进行分句时，由于学习者语料的标点可能存在偏误，因此按照偏误修改后的语料进行分句，以“。”“！”“？”等标点进行切分。

其次，以语法点为切入点，试图选取具有有效、典型偏误信息的语料。在对外汉语教学当中，语法点教学是常用的一种教学手段，它是对系统语法进行切分之后的结果。本文借鉴了北京师范大学的汉语国际教育动态语料库（CTC）[4](谭晓平et al., 2015)中的语法点信息，并进行了增补和筛选，总结了包括复句、固定结构、介词及介词结构、特殊句型四大类的137个语法点。随后采用人机互助的形式对语料进行语法点标注，共提取出了1056句带有语法点偏误信息的句子。

3.3 标注流程

首先，在标注之前对语料进行一定处理，得出适于标注并易于进行对比分析的偏误句与目标句的平行句对；其次，开展两个层面的标注：词层面的分词及词性标注，句法层面的依存句法标注。在这两个层面上，都采用人机结合的方式标注，并设置了标注员和审核员两种角色，

[4]汉语国际教育动态语料库：http://www.aihanyu.org/basic.aspx

既可以提高标注的效率，也能保证标注的质量。

3.3.1 语料预处理

本树库面向汉语学习者，并期望探索汉语学习者偏误对依存分析的影响，因此树库中的语料均以一组平行句对的形式显示，即汉语学习者生成的偏误句与母语者修改后正确的目标句。HSK动态作文语料库已对偏误句的字、词、标点偏误进行了一定的修改，对于此类已给出修改结果的偏误标注，可以通过程序直接获得偏误句和目标句。

然而，对于句式方面的偏误，HSK语料库只进行了偏误标注，而没有进行修改。比如“人类是有精神方面的追求{CJsd}，必须满足其需求。”一句中，{CJsd}表示此句是“是.....的”句式错误，但没有明确表明应如何修改。此外，HSK语料中还有一些错误存在漏标的情况。针对这两类句子，我们对其进行人工修改，将原句按照最小改动的原则修改为符合汉语语法的句子，得到正确的目标句。在492句偏误句当中，共对227句语料进行了人工修改。

3.3.2 语料标注

标注过程包括分词、词性标注和依存标注。分词及词性标注是做好依存标注的前提，其标注质量直接关系着依存句法的标注质量，一旦分词和词性出现错误，会使得句法层面的标注难以进行，增大标注员的负担。

UD_Chinese-GSD[5] 是谷歌在UD上发表的汉语树库，本文参考了它的分词及词性标准，在实际标注时采用字标注的方式进行标注，如“苍白”一词为形容词（ADJ），则对这两个字分别标注“B-ADJ”和“I-ADJ”两个标签。但在标注过程中发现GSD的分词和词性标注规范仍有一定不足之处，主要有：1）部分出现在相同语言环境中的同一个词的词性不同。2）部分出现在相同语言环境中的同一个词的分词情况不同。3）部分词的分词情况与词典有差别。4）相同成分构成的词，有的被分开列为两个词，有的合起来标为一个词。如同样是“ADV+‘是’”的结构，“都是”、“不是”被标为一个词，“总是”则被分开标为“总”和“是”两个词。针对这样的一些问题，我们参考了宾州中文树库CTB的分词(Xia, 2000b)及词性规范(Xia, 2000a)，对GSD的分词和词性标注规范进行了部分修改，如：1）现阶段不对词缀、类词缀进行单独分词，在后期工作中如需拆分词缀，则针对词缀、类词缀的判别进行规定后制定词缀表，再对词缀进行拆分。2）把成语和俗语当作一个整体来标注。因为成语和俗语是一种相沿习用的固定短语，具有结构定型、意义完整两个基本特点。它们在语句中是作为一个整体来应用的(王兴全and 方忠, 2017)，其意义通常不是字面意义的简单相加。如果拆分进行标注的话，会破坏其所要表达的意义。分词与词性标注完成后，即可进行依存分析。

为了提高标注的效率和一致性，我们采用人机结合的方式进行标注。标注流程如下：

1）用GSD数据训练分词、词性和依存句法分析模型，并得出语料的分析结果。

2）将每条句子的分析结果随机分配给两位标注者标注。标注完成后，若标注结果完全一致，则确立为标准答案，结束流程；否则进入步骤3）。

3）如果两个标注答案不完全一致，则将这条句子分配给专家进行审核，由专家给出标准答案，结束流程。

3.4 标注平台

为了提高标注的效率，降低标注的难度，本文基于Arborator [6]开发了一个在线标注平台[7]。该平台操作简单，可以多人同时进行标注，也能减少标注管理者的工作，便于查看标注的进度和管理标注结果。标注界面如图6所示。

标注平台的主要功能有：

1）修改句子的分词错误和词性标签。当存在分词错误时，标注员或审核员可通过删除、增添、替换等操作对分词情况进行修改；当存在词性错误时，标注员和审核员可在预设好的词性标签中选择正确的标签进行修改，避免人工输入时可能产生的误操作。

2）标注依存弧及标签。标注员和审核员可通过将依存弧从弧起始端的词语拖拽至弧末端对应的词语来标注依存弧，并通过预设的依存标签栏选择正确的依存标签。

[5]UD_Chinese-GSD：https://universaldependencies.org/treebanks/zh_gsdsimp/

[6]Arborator：https://github.com/Arborator/arborator-server

[7]https://yatlc.wenmind.net/

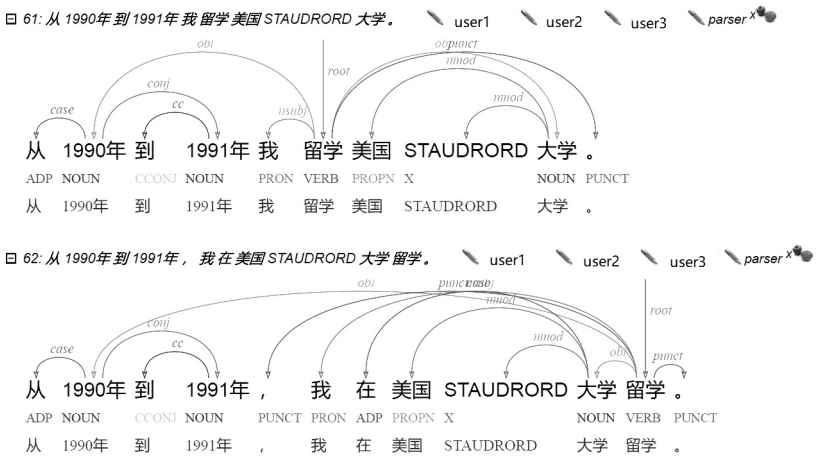

Figure 6: 标注平台界面

3）对比多人标注结果。审核员可以自由选择想要对比的多人标注结果，标注一致的显示为灰色，不一致的弧或标签则以不同的颜色显示不同的标注结果，极大地方便了审核工作。

4 标注数据分析

基于以上标注流程，本文对100条语料进行了依存标注，其中每条语料都至少含有一个非标点符号的语法错误，平均句长为35字，平均每条语料带有4处偏误。首先，统计整体依存标签的分布情况，并尝试从语言学的角度来解释。其次，通过分析两位标注员之间的一致性来观察标注的质量，并通过对比偏误句中带偏误标记的词与无偏误标记的词的一致性来分析偏误对标注质量的影响。最后，分析偏误类型对依存句法分析的影响。

4.1 依存标签分布情况分析

本文对语料库中的依存标签的数量和分布情况进行了统计，总共包含35类、4554个标签。由于个别标签出现的次数非常少，所以只呈现了出现次数在100次以上的依存标签的分布情况，如图7。经过对各标签情况的观察、分析，可以获得以下信息：

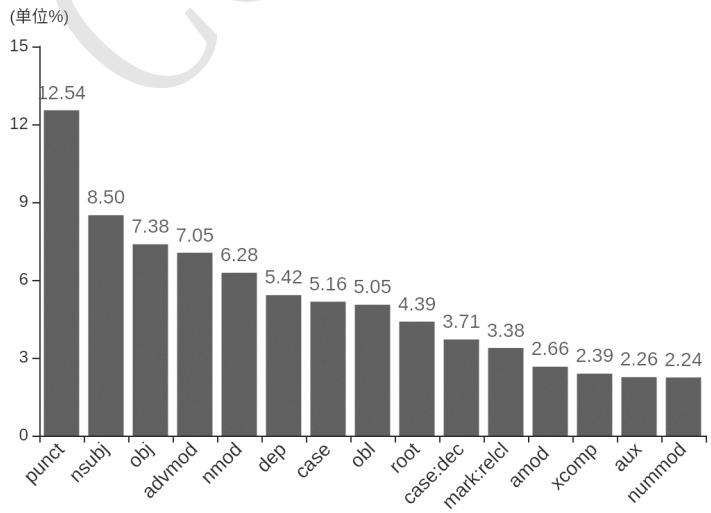

Figure 7: 高频标签分布情况

1）主干成分占比高。句法结构中的主干成分，如表示主语的“nsubj”、表示宾语的“obj”、表示状语的“advmod”“obl”、表示定语的“nmod”“amod”“nummod”等占比明显高于其他成分占

比，基本符合人们的语言认知。这说明，从总体上看我们的标注规范是科学、合理的。在高频标签中没有出现补语标签“xcomp:ccomp”，补语是汉语与印欧语言的一大区别，印欧语言中没有补语这种句法成分，因此二语学习者在语言迁移的过程中受到母语的影响，较少地使用补语成分。

2）复句较多。排名前十的标签中，除了句子的主干成分以外，还包括“punct”、“dep”两类标签。“dep”是表示复句关系的标签。“punct”表示标点，排名第一，占比远高于排名第九表示根节点的标签“root”，说明许多句子包含多个标点符号，符合复句的特点。这是因为语料主要来源于考试中的作文语料，学习者为了得到较高的分数，会更多地使用长难句。这反映了语料的特点：句子长度大，复句较多，各小句间的层次关系比较复杂。

4.2 语法偏误对标注的影响

本文通过计算两位标注员在偏误句、目标句、偏误句中的偏误词语和无偏误词语的一致性来进一步分析语法偏误是否会对标注质量产生较大的影响。

对于一致性和准确率的计算方法如下:

依存弧的一致性（C_A）：假设一个句子有m条依存弧，a和b两个人都对其进行标注。如果两人的标注结果中一致的依存弧为i_A条，则$C_A = i_A/m$;

依存标签的一致性（C_L）：假设一个句子有m条依存弧，a和b两个人都对其进行标注。如果两人的标注结果中有i_L个依存标签相同，则$C_L = i_L/m$;

带标签的依存弧的一致性（C_{LA}）：假设一个句子有m条依存弧，a和b两个人都对其进行标注。如果两人的标注结果中有i_{LA}条依存弧相同且依存标签也相同，则$C_{LA} = i_{LA}/m$;

我们统计了两位标注员间的一致性，如表1所示。首先，以句为单位，对两位标注员的一致性进行统计。“所有句”表示两位标注员标注的所有数据，“偏误句”表示带有偏误的句子，“目标句”是对偏误进行修改后的句子。其次，我们作出一个假设：语法偏误对标注的一致性有影响。为此，通过编辑距离将偏误句与目标句进行比对，提取出偏误句中经过“替换”、“删除”操作的词，将之称为“偏误词”，其他未发生改变的词，称之为“无偏误词”。

类别	C_A	C_L	C_{LA}
所有句	91.59	92.75	87.77
目标句	92.14	93.40	88.59
偏误句	91.02	92.09	86.93
无偏误词	91.07	92.49	87.09
偏误词	90.64	89.14	85.77

Table 1: 一致性分析

由表1中的数据可得到以下结论:

1）标注员总体的标注一致性较高。这在一定程度上肯定了标注规范与流程的合理性。

2）在依存弧、依存标签和带标签的依存弧上，偏误句的一致性都明显低于目标句，偏误词的一致性都低于非偏误词。这可以表明语法偏误在一定程度上增加了标注的难度，对标注的一致性造成影响。

3）比较“C_A”和“C_L”两列数据，在“所有句”、“目标句”、“偏误句”、“无偏误词”中，依存标签的一致性都高于依存弧。其原因是存在某些会产生连锁反应的标签，如表示根节点的“root”，当这些标签指向的词不一致时，会导致这些词与子节点间的弧也发生改变，而其上的一些依存标签却不发生改变。当“root”所指的词不一致时，指向语气词的弧会发生改变，但“discourse”标签不发生改变。在“偏误词”中，依存标签的一致性却低于依存弧，并且与“无偏误词”相比，相较于依存弧，依存标签的一致性差别明显更大，说明语法偏误对依存关系的影响比对依存结构的影响更大。

4.3 语法偏误对依存句法分析的影响

鲁健骥 (1994)认为偏误类型可分为四种：“误加”、“遗漏”、“误代”、“错序”。遗漏偏误是指由于在词语或句子中遗漏了某个/几个成分导致的偏误；误加偏误是指当某些语法形式发生某种变化时，在通常情况下可以/必须使用的某个成分变为一定不能使用这个成分，而汉语学习者

往往不了解这种条件的变化仍然使用这个成分，从而导致的偏误；误代偏误是指从两个或几个形式中选取了不合适于特定语言环境的一个词造成的偏误；错序偏误指的是由于句中的某个或几个成分放错了位置造成的偏误(鲁健骥, 1994)。本文按照这种分类，探索偏误类型对句法分析的影响，以帮助训练二语的句法分析器，提高二语语法纠错任务的准确率。

在对语料中出现的偏误考察后发现:

1）当出现“误加”或“遗漏”偏误，即词语冗余或缺失时，如果该偏误在句中承担核心句法成分或者和其他词语构成某个结构共同承担核心句法成分时，那么该偏误对句法分析的影响较大，即平行句对的标注结果大有不同，具体表现为根节点与依存关系标签发生变化，如图8；反之则适中，即平行句对的标注结果略有不同。略有不同表现在多一个弧、少一个弧、弧长度有所不同或弧交叉，但根节点与依存关系标签未发生变化，如图9。

2）当出现“误代”偏误，即词语使用错误时，如果该词与修改后的词具有相同的词性，且出现的句法环境也相同，那么该偏误对句法分析的影响较小，即平行句对的标注结果基本一样，如图10；反之则较大。

3）当出现“错序”偏误，即词语的位置发生了改变时，如果所修饰的核心词没有发生改变，仅仅是弧的距离发生改变，那么该偏误对句法分析的影响适中；如果所修饰的核心词发生改变，即弧的父亲节点发生改变，那么该偏误对句法分析的影响较大。

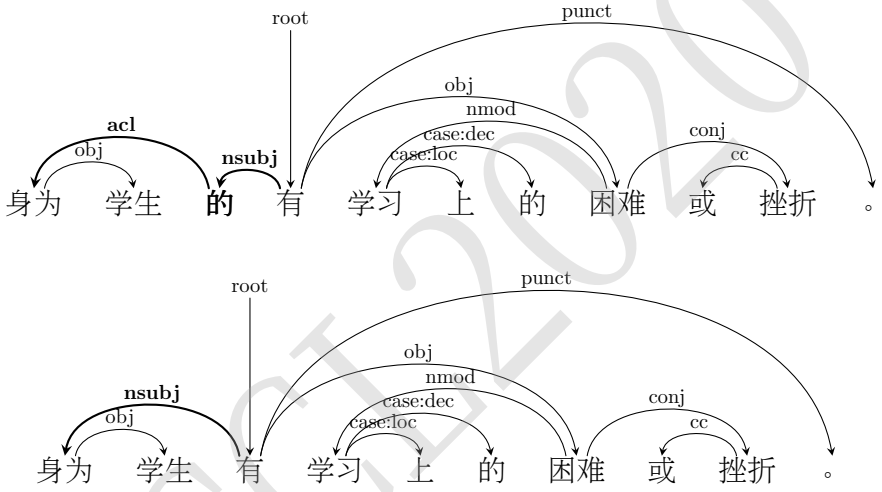

Figure 8: “误加”偏误对句法分析影响较大示例

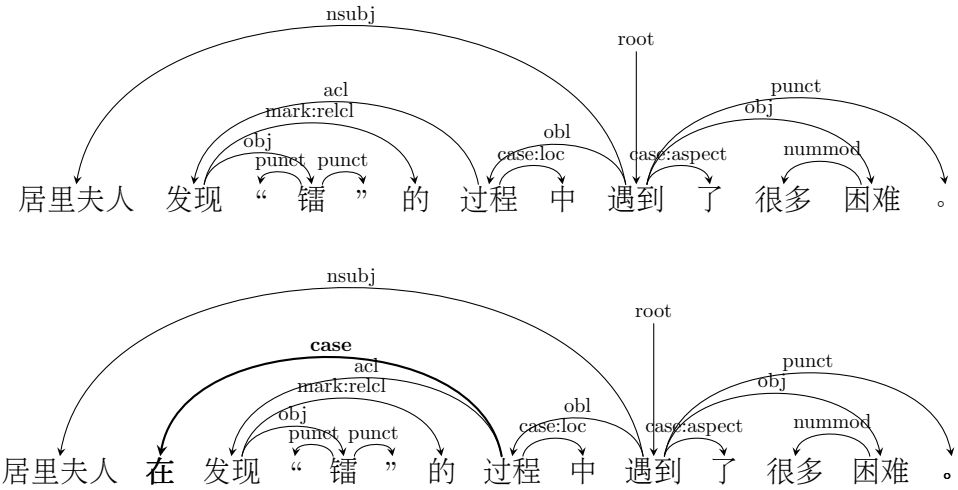

Figure 9: “遗漏”偏误对句法分析影响适中示例

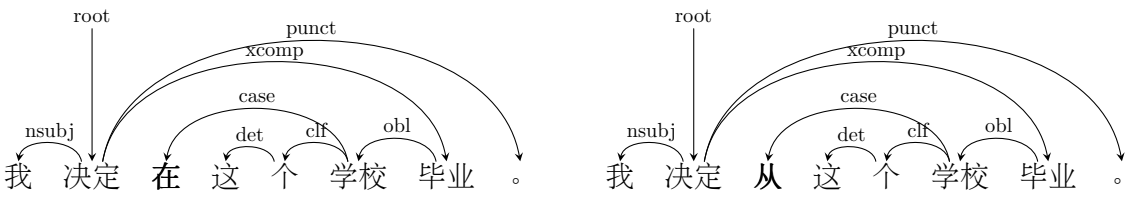

Figure 10: “误代”偏误对句法分析影响较小示例

5 总结与展望

本文介绍了我们在汉语学习者依存句法树库构建上所做的一些工作，包括标注规范的改进、数据选取、标注流程等，并对依存关系标签分布情况及偏误对标注质量及依存句法的影响情况进行了分析。本文的创新之处在于：1）对依存句法标注原则进行了改进，并弥补了现有标注框架的不足，考虑到了汉语及汉语学习者语料的特点，增强了对汉语学习者语料的适应性。2）在严格的标注流程控制后，通过统计与分析发现语法偏误对标注质量和依存句法分析都有一定的影响，对待不同的偏误要采取不同的标注策略，以降低标注难度，节约标注时间。目前我们树库的规模还比较小，未来我们将继续完善标注规范，在标注规范的指导下扩大树库规模，为二语教学与研究提供更多帮助，也为句法分析器、语法改错等相关研究提供更多数据支持。

参考文献

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany, August. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Sylviane Granger. 2012. Learner corpora. *The encyclopedia of applied linguistics*, pages 1–8.

John Lee, Herman Leung, and Keying Li. 2017. Towards universal dependencies for learner chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, F. Ginter, Jan Hajivc, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and D. Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *LREC*.

Marwa Ragheb and Markus Dickinson. 2014. Developing a corpus of syntactically-annotated learner language for english. *CLARIN-D*, pages 292–300.

Geoffrey Sampson. 2011. Susanne-a deeply analysed corpus of american english. *New Directions in English Language Corpora: Methodology, Results, Software Developments*, 9:171.

Fei Xia. 2000a. The part-of-speech tagging guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*, page 38.

Fei Xia. 2000b. The segmentation guidelines for the penn chinese treebank (3.0).

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

刘挺and 马金山. 2009. 汉语自动句法分析的理论与方法. 当代语言学, 011(2):100–112.

张宝林and 崔希亮. 2013. "全球汉语中介语语料库建设和研究"的设计理念. 语言教学与研究, (05):27–34.

张宝林. 2009. "HSK动态作文语料库"的特色与功能. 国际汉语教育, (4):71–79.

张宝林. 2010. 汉语中介语语料库建设的现状与对策. 语言文字应用, 000(3):129–138.

李娟, 谭晓平, and 杨丽姣. 2016. 汉语中介语语料库应用及发展对策研究. 曲靖师范学院学报, 35(2):86–91.

王兴全and 方忠. 2017. 现代出版物语言文字使用规范. 电子科技大学出版社.

肖丹, 杨尔弘, 张明慧, 陆天荧, and 杨麟儿. 2019. 面向汉语中介语的依存句法标注规范. 第十八届中国计算语言学大会（CCL 2019）.

谭晓平, 杨丽姣, and 苏靖杰. 2015. 面向汉语(二语)教学的语法点知识库构建及语法点标注研究. 中文信息学报, 29(6):54.

郭丽娟. 2019. 汉语依存句法分析树库构建与应用研究. Ph.D. thesis, 苏州大学.

鲁健骥. 1984. 中介语理论与外国人学习汉语的语音偏误分析. 语言教学与研究, (03):44–56.

鲁健骥. 1994. 外国人学汉语的语法偏误分析. 语言教学与研究, (01):49–64.

黄昌宁and 靳光瑾. 2013. 从宾州中文树库观察三个汉语语法问题. 语言科学, 12(2):178–192.