**文章编号**：1003-0077(2018)11-0008-08

# 维吾尔语依存树库构建及统计分析

麦热哈巴·艾力，吐尔根·依布拉音，加米拉·吾守尔

（新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046）

**摘　要**：本着构建维吾尔语依存树库的目的，该文根据黏着性语言的结构特点及其在依存属性中对依存角色的影响，提出构建维吾尔语依存树库时需要考虑的几点要素。其包含依存粒度的确定、维吾尔语依存关系、标注原则、依存树结构以及标注工具的设计与实现。然后根据《维吾尔语依存树库标注手册》人工标注了 3 400 多条句子并从三个角度对依存树库信息做了统计分析。

**关键词**：依存句法；依存树库；维吾尔语

**中图分类号**：TP391　　**文献标识码**：A

## Construction of Uyghur Dependency Treebank and Its Statistical Analysis

Mairehaba Aili, Tuergen Yibulayin, Jiamila Wushouer

(College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China)

**Abstract**: As an agglutinative language, the complex word structure in Uyghur affects its dependent role. This paper presents a few important factors in Uyghur that should be considered when building Uyghur dependency treebank. These factors include (1) the granularity of dependency, (2) dependent relations, (3) the annotation guidelines, and (4) the annotation tool. More than 3 400 Uyghur sentences are annotated manually based on the "Uyghur dependency tree bank annotation manual". A statistical analysis have been made on that Uyghur dependency Treebank for three aspects.

**Keywords**: dependency syntax; dependency treebank; Uyghur

## 0　引言

依存句法以形式简单、易于理解、侧重于反映句子中各成分之间的语义关系等特点，一直备受国内外研究者的青睐。构建相应句法树库无疑是句法分析处理的基础，虽然耗时、耗力，但却是利在当今、功在千秋的工作。目前国内外相应研究机构相继构建并发布了很多相关语言的依存树库，包括瑞士语[1]、希腊语[2]、俄语[3]、日语[4]、汉语[5]、捷克语[6]、土耳其语[7]、德语[8]及斯洛文尼亚语[9]等。

近年来，维吾尔语的信息处理研究也得到了快速发展，包括维吾尔语词法分析器、校对系统、维汉机器翻译以及维吾尔语短语句法分析等方面。本文在已有的研究成果之上探讨了现代维吾尔语依存树库的构建过程。即：从维吾尔语的词法、句法特点出发，分析维吾尔语的结构特性，选定维吾尔语依存树库中依存单元粒度，确定依存关系及标注原则，设计树库存储结构等，并从不同角度对人工标注的依存树库做了统计分析。

## 1　维吾尔语依存树库的构建

作为黏着性语言，维吾尔语的词法、句法特性主要体现在以下几点：(1)在词干后缀接词尾构成不同的形态，附加语素数量多且携带一定的语义信息；(2)维吾尔语属于主语可省略(pro-drop)型语言，省略的主语可从动词词尾判断出来；(3)虽然维吾尔语的主要句法结构为 SOV 型，但也有 SVO 型(如

**收稿日期**：2017-11-10　**定稿日期**：2018-05-30

**基金项目**：新疆少数民族科技人才特殊培养计划(201423120)；国家 973 计划(2014CB340506)

诗词等)的情况。同时,维吾尔语的语序比较灵活。如：ikki yash yigit (两个年轻人)和 yash ikki yigit (年轻的两个人)是同一个意思,但所强调的内容不一样：前者强调年纪,后者强调人数。

构建依存树库的目的之一是为计算机的语义分析打基础,服务于计算机“理解”自然语言。所以构建依存树库时,根据维吾尔语的特点尽可能地保留或体现语义信息是本文构建依存树库时所重视的要点。维吾尔语中某个单词在句子中起到的语法功能往往受到缀接其后词尾的影响,从而也影响它的依存角色：支配词(head word)或从属词(dependent word),如图 1 所示。

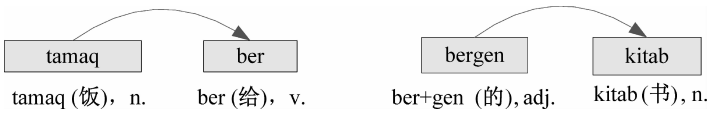


图 1　同一个词不同形态的不同依存角色

图中左边词对“tamaq ber”(给饭)中“ber”是支配词,右边词对“bergen kitab”(给的书)中的“bergen”(ber＋gen)则为从属词(此处词尾“-gen”使其语法功能发生了变化)。所以我们认为词尾在决定依存角色时起到不可忽视的作用。

同时,一个质量高、规模大、覆盖面广的标注树库为相应语言的信息处理可以提供丰富的信息。我们期望所构建的树库能够受到更多“消费者”的青睐。既为语言学家提供研究语言形态结构及分布、句法结构、语序等方面的信息,又可为计算语言学家建立语言模型、研究句法分析评价标准等提供便利。

考虑到以上因素,本文的维吾尔语依存树库的构建过程采用了以下流程：(1)依存单元粒度的确定及表示；(2)依存关系的确定及标注原则；(3)依存树库存储结构的确定；(4)依存标注工具的开发。

### 1.1　维吾尔语依存粒度的确定及表示

维吾尔语中,在词干后接词尾使其具有不同的形态且具有不同的语法功能。例如,ket(走,v.),ket＋ken(走的,adj.)；kongül(心灵,n.),kongül＋lük(开心的,adj.)。通常,一个发生形态变化的词在句子中只有的语法功能往往是由它最后的词尾来决定的。例如,“zamaniwilashturush (现代化)”一词的结构为‘zamaniwilash(现代化,v.)＋tur＋ush’,是最后词尾“-ush”使其具有名词功能。中间词尾虽不具有最终语法功能决定权,但却搭载着某种语义信息。如果把词尾简单地看成是词干后面的追随者,则不仅造成数据稀疏问题,更严重的还会失去一些重要的语法语义信息。例如,分析短语“balilarning eng kichiki(孩子们中最小的)”的依存情况的分析结果如图 2 所示。

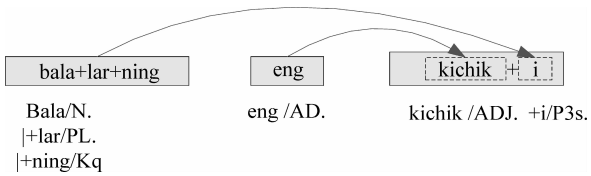


图 2　词尾在依存关系中的作用

图 2 中可看出,最后一词“kichiki(小的)”有两层组成：“kichik(小,adj.)＋i(词尾,第三人称单数)＝kichiki (n.)”。仔细分析不难发现,副词“eng”(最)修饰的不是 kichiki 而是 kichik,因为副词只能修饰动词或形容词,不能修饰名词；词“balilarning”(注：中间有个弱化音位)中“＋ning”从属于“＋i”。可见,依存关系中充当支配词或从属词的不仅是某个单词也可能是某个词尾。

经过以上分析,本文认为在维吾尔语依存树中以词素(即词干及词尾)而不是单词来表示依存单元更能体现此语言的结构特性,且以这种方式表示的依存关系更具有代表性。例如,分析以上实例“balilarning”与“kichiki”之间的依存关系时,可获得依存模板“n.＋ning　n.＋i”(表示谁的什么,n.表示名词)及其之间的依存关系,此模板能代表其他具有类似结构的两个词之间的依存关系,例如,“Alimning topi (阿里木的球)”,“gülning yopurmiqi (花儿的花瓣)”等。

维吾尔语中还有很多固定搭配词,也称作多词表达,包含对偶词、复合词、习语、成语等。例如,bara(去)和 barmay(不去)构成的对偶词 bara-barmay(一去就…)。维吾尔语中多词表达的结构很灵活,其中包含的词可有多种形态,词之间还可以插入其他词,对有些语块是否为多词表达还存在争议。对此我们采取以《维吾尔语详解词典》(2008 年版)为主要依据,以词典中出现的语块视

为多词表达并以下划线连接成一词，再判断它与其他词之间的依存关系；若不在词典中，则视为一般词处理。

### 1.2 维吾尔语依存关系

句法分析的输入是一串词，输出的是词串按某种关系呈现的结构。依存语法中的关系就是句子中词串之间的依存关系。语言不同，则其词性集以及依存类型、数量也不同。捷克语依存树库 PDT 采用了 7 种依存类型，德语依存树库 TIGER 采用了 49 个依存类型。汉语依存树库中，清华大学计算机学院依存树库一开始采用了 106 个依存类型，后来减为 44 个；同时，依存数量的多少也很重要，数量多，对计算机的识别及分析带来一定的影响，增加时间、空间复杂度；数量少，则不足以描述语言现象，降低模型的表现能力。

传统的维吾尔语句法指定的关系主要为主语、宾语、定语、状语、补语及谓语。显然，它们具有重要的参考价值。但它们不能全面地刻画句子中词之间的关系，需要指定既能覆盖维吾尔语句子的语法结构，又能符合计算机处理的依存关系集。我们借鉴国内外发布的依存树库，特别是一些黏着性语言，例如，日语、韩语、土耳其语在这方面的研究成果，再结合维吾尔语的特性制定了维吾尔语依存关系并标注句子以尝试其合理性。通过反复几次修改，最后制定了 23 种依存关系，分别为以下所示：

ABL(Ablative Adjunct) —起因关系
APPOS(Apposition)—同位关系
ATT(Attributive Modifier)—定中关系
ADV(Adverbal modifier)—状中关系
AUX(Auxiliary Verb)—体助关系
CLAS(Classifier)—分类关系
COLL(Collocation)—词串关系
CONJ(Conjunction)—连词关系
COORD(Coordination)—并列关系
DAT(Dative Adjunct)—指向关系
INST(Instrumental Adjuncts)—工具关系
LOC(Locative Adjunct)—时位关系
OBJ(Object)—宾动关系
POSS(Possessor)—领属关系
POST(Postpositions)—后置关系
QUOT(Quation)—引用关系
ROOT(Root of Sentence)—主管关系
PRED(Predicate)—表语关系
SUBJ(Subject)—主谓关系
CL(CLause)—从句关系
IND(Independent component)—独立关系
COP(Copula)—表系关系
COM(Comparison)—对比关系

为了给标注者提供可参考的标注依据、保证标注的一致性，我们还制定了《维吾尔语依存树库标注手册》，其中讲述了每一种依存关系的定义、出现形式、判断依据以及相应的实例。同时，又根据在测试过程中所遇到的情况以及国内外相应树库建设中所积累的经验，制定了在标注语料时必须遵守的几点原则。

(1) 单纯节点规则：依存树中不允许存在游离于集体之外的成分。

(2) 独根节点规则：一个依存树只能有一个根节点(Root)，它是依存树中唯一没有父节点的节点，根节点支配着其他的所有的节点。

(3) 单一父节点规则：在依存树中，除了根节点没有父节点之外，其他节点都只有一个父节点。

(4) 非交规则：依存树中的树枝不能彼此相交。

(5) 有向条件：依存关系本质上是一种非对称的，有向的同现关系。在维吾尔语依存关系标注中，从属成分(子节点)的方向指向支配成分(父节点)，而不是相反。

(6) 起点规则：依存关系的起点是从属(子节点)词语的最后一个附加成分，结束点是支配(父节点)词语的词干(或在词根上附加词尾的部分)

所有参与依存树标注人员需在参加按此手册进行的培训、测试后方能标注。

### 1.3 维吾尔语依存树的存储结构

为了让维吾尔语依存树包含更多的语言信息以及为以后的扩展提供便利，借鉴国内外相关树库的构建经验，本文规定维吾尔语依存树中每个单词应包括表 1 所示的信息。

**表 1 依存单元属性表**

| 标记 | 意思 | 标记 | 意思 | 标记 | 意思 |
|---|---|---|---|---|---|
| ID | 词位 | Inf | 屈折变化标记 | Lem | 词汇 |
| Morph | 形态结构 | Rel | 依存关系 | Word | 词本身 |

表 1 中，ID 表示当前词索引(下标从 1 开始)，

Morph 表示当前词的形态结构，其格式为：Stem＋$Af_1$＋$Af_2$＋…＋$Af_n$（其中 Stem 为词干，Af 为词尾，下标为词尾索引，从 1 开始）；Lem 表示当前词的词典形式；Inf 表示当前词在缀接词尾后的屈折组信息；Rel 为当前词与支配词之间的依存关系，它是三元组＜HeadID，Inf_pos，Type＞组成。其中，HeadID 表示当前词所从属的支配词索引，若当前词是根节点（Root），则为 0；Inf_pos 表示支配词中起支配作用的屈折组位置；Type 表示依存关系类型；Word 表示当前词本身。

根据表 1，以上短语 balilarning eng kichiki 的依存信息如表 2 所示。

**表 2　依存单元属性描述**

| word | ID | Morph | Lem | Inf | Rel |
|---|---|---|---|---|---|
| balilarning | 1 | bala＋lar＋ning | bala | bala＋N，＋Pl，＋Kg | 3,2,CLAS |
| eng | 2 | eng | eng | eng＋Adj | 3,1,ATT |
| kichiki | 3 | kichik＋i | kichik | kichik＋Adj，＋P3s | [] |

表 2 中第一行最后 Rel 字段的值为三元组＜3，2，CLAS＞，表示当前词“balilarning”从属于第三个词“kichiki”的第二个屈折组（此处为-i），“CLAS”表示其依存关系。

最后本文以 XML 文件形式保存了依存树库，树库中每个句子作为一个节点＜S＞ ＜/S＞保存，它由每一个词的以上信息构成，例如：

```
＜S＞
＜w ID＝1　Lem＝“bala” Morph＝“ bala＋lar＋ning” Inf＝[(1，“bala＋N.”)，(2，“＋Pl”)，(3，“＋Kg”)] Rel＝[3,2，“CLAS”]＞ balilarning ＜/w＞
＜w ID＝2　Lem＝“eng” Morph＝“eng”　Inf＝[(1，“eng＋Adj.”)] Rel＝[3,1，“MOD”)]＞ eng ＜/w＞
＜w ID＝3Lem＝“kichik” Morph＝“kichik＋i” Inf＝[(1，“kichik＋N.”)，(2，“＋P3s”)] Rel＝[]＞ kichiki ＜/w＞
...
＜/S＞
```

## 1.4　维吾尔语依存树标注工具

依存树的标注是一件费时、费力的过程。为了减少语言学家或标注者的标注工作量以及加快标注的速度，开发一个可视、可控的标注工具很有必要。虽然也有一些开源的标注工具，但语言特性、标注要求等方方面面的不同，这促使我们开发一个适合维吾尔语依存树的标注工具。根据维吾尔语依存树标注规范提及的相关要求及实际需求，我们设计了维吾尔语依存树库标注工具软件应有的功能模块：可单独/批量输入、查重、带箭头的链接线来表示依存角色（箭头表示支配词，另一头表示从属词）、不同颜色表示不同依存关系、以 XML 格式保存结果、重现已标注树以便修改。最后我们以 C＃作为开发工具，实现了可在 Windows 平台下工作的维吾尔语依存树库标注工具。图 3 为此工具的截面图一例，为节省篇幅此处省略了工具的工作流程及更多使用实例。

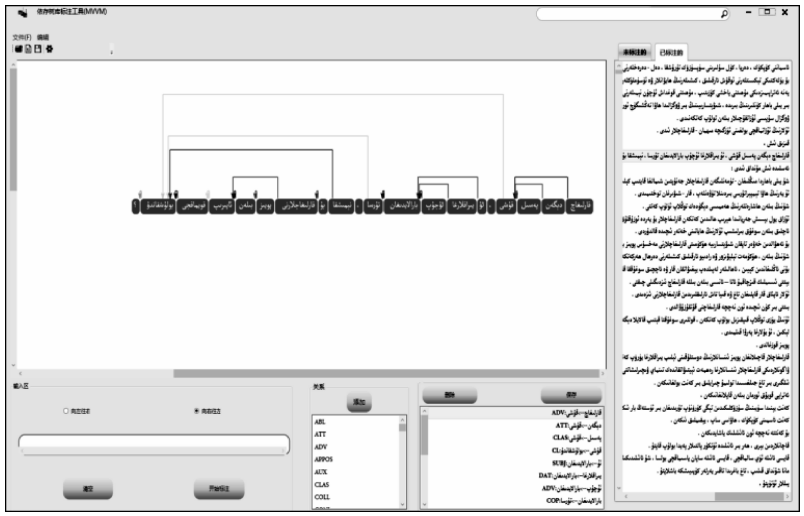

图 3　依存树库标注工具软件

## 2 维吾尔语依存树库结构分析

维吾尔语依存树库的构建过程与其他树库的构建过程一样，首先确定语料的来源，然后标注者根据《维吾尔语依存树库标注手册》予以相应培训后进行人工标注。本文选定新疆维吾尔自治区中小学所使用的双语教材“语文”(维吾尔文)中故事、新闻、事件类文本作为生语料。选择中小学生课文的原因是其语言通俗、简单、易懂、层次显明、具有一定的普遍性。这使标注者能从中得出具有一定代表性的结构，以从简单到复杂、从浅层到深层的原则做标注。

本文使用新疆大学新疆多语种信息技术重点实验室开发的维吾尔语词性标注器做了单词的形态分解。根据前面的考虑，句子中每个词应该按其词干、词尾的形式出现。但实际操作时本文受到了一些客观条件的限制：(1)目前维吾尔语自动词性标注器对词尾的分解程度不够理想(准确率达不到要求)；(2)维吾尔语词尾特别是动词词尾的分解上还存在一定的争论；某个词的词性以词干为主，不能对形态变化后的部分做标注(例如，bergen(给的)一词是词干“ber”(给，V.)和词尾“-gen”结合而成。对它标注词性时以词干为主标为“V.”，忽略了添加“-gen”后变成了形动词即具有形容词的特点；(3)当句中每个单词分解后，例如，“kitabliringlardiki ＝kitab＋lar＋inglar ＋diki”(在你们书中的)，句子的长度变得过长，特别是词数比较多的长句中此情况变得尤为突出。分析实际情况，最后决定现阶段采取对句中每个词暂时不考虑形态，等条件成熟时再把这些细节信息加进来的方案。

目前构建的树库所包含的句数为 3 456 条，其中句子长度最长的为 54(以空格作为自然分界符)，最短的为 2，句子平均长度为 11.6。为了了解当前树库结构相关的信息，本文从以下三方面做了统计分析：(1)树库中不同依存关系之比；(2)依存距离及依存方向的计量分析；(3)词性与依存关系之间的关系。

(1) 依存关系比例分析

统计得到目前树库所包含的依存关系总数是 31 356 个。其中，每一种依存关系及其所占的比例如图 4 所示。(注：为了处理标注时所遇到的疑难依存关系，我们增加了依存关系 OTHER，所以图中共显示 24 种依存关系)。

从比例图可以看出，所占比例排在前面的 5 个

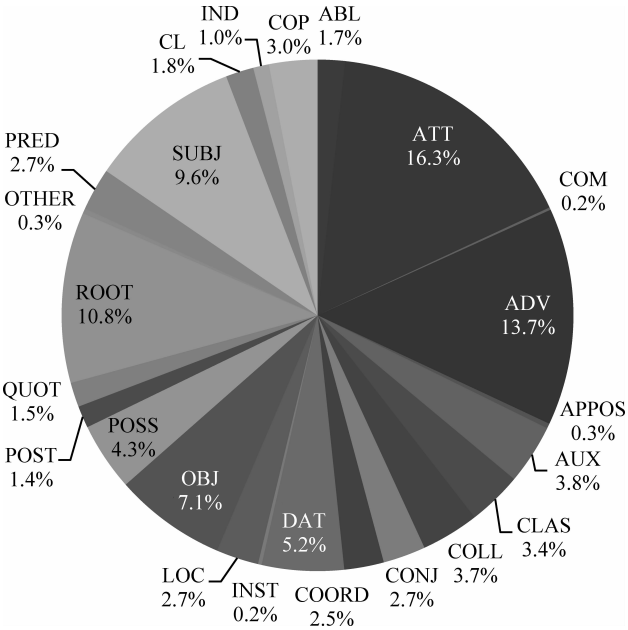


图 4 依存关系比例图

依存关系依次为 ATT、ADV、ROOT、SUBJ 及 OBJ。出现这种比例关系是合理的。因为，ROOT 是依存树的根节点。凡是依存树肯定有一个 ROOT 关系，而 ROOT 的数量大于 SUBJ 的数量也能说明树库中存在没有主语的句子，这又与维吾尔语作为主语可省略(pro-drop)语言的特性相吻合；同时，OBJ 的数量少于前几种关系也是可想而知。ATT 与 ADV 的数量远大于 ROOT、SUBJ 与 OBJ。原因是某个句子中可以出现多个 ATT 及 ADV，使得其数量增加。虽然，修饰关系在语义层面不像 ROOT、SUBJ 及 OBJ 一样决定句子关键内容。但它进一步描述事物之间精细的关系，传达更深、更广的信息，是符合人类使用语言的特性。

图 4 显示关系 COP 和 AUX 的数量并不在少数，这两种依存关系是由 N-V 和 V-V 型词对构成。其中，第二个动词在此处失去自己的语义，起助动词的作用。维吾尔语中很多主动词都可作为助动词，应用非常灵活。这种灵活性往往也是确定其依存关系的难点。维吾尔语助动词、轻动词等也是目前维吾尔语语法研究中争议最多的部分，语料中出现的数量之多又说明对其研究的紧迫性。

(2) 依存距离及依存方向的分析

本文对树库做了句法计量的分析，包括依存距离、依存方向等。虽目前树库规模及覆盖面还很小，得出的结论还未具有代表性。但计量分析可以基本反映目前语料库中句子结构、依存倾向等信息，也为

以后相关研究提供可比的信息。

本文中依存距离(标示为 $Ds_r$)定义为支配词与从属词之间的词数，如式(1)所示。

$$Ds_r = Ind_h - Ind_d \quad (1)$$

用式(1)来计算依存距离值，其中 $Ind_h$ 表示支配词的索引值，$Ind_d$ 表示从属词的索引值。根据支配词出现在从属词之前($|Ind_h| < |Ind_d|$)或之后($|Ind_h| > |Ind_d|$)，依存距离就有正负值，可用于表示依存方向。若支配词出现在从属词之后，则 $Ds_r$ 得正值，定为正向依存，标示为 $Rel^+$；若支配词出现在从属词之前，则 $Ds_r$ 得负值，定为反向依存，标示为 $Rel^-$；用依存距离的绝对值计算出一个句子的依存距离均值。同时，若树库中有 $n$ 个句子，$m$ 个依存关系，则整个树库的依存距离如式(2)所示。

$$\frac{1}{m-n}\sum_{m-n}^{i=1} | dis_{R_i} | \text{[10]} \quad (2)$$

由式(2)得出，其中 $R_i$ 为第 $i$ 个依存关系。对依存树库做了以下若干项的统计，其计算结果如表 3 所示。

**表 3　依存树库的统计结果**

| Num | $Av_{dis}$ | $Rel^+$ | $Rel^-$ | $NB_{Rel}$ | $NNB_{Rel}$ | $Av_{dis+}$ | $Av_{dis-}$ |
|---|---|---|---|---|---|---|---|
| 31 356 | 2.76 | 99.08% | 0.92% | 61.3% | 38.7% | 2.44 | 3.14 |

表 3 中“Num、$Av_{dis}$”指的是树库中依存关系总数及树库平均依存距离；“$Rel^+$、$Rel^-$”指的是不同方向依存所占百分比；“$NB_{Rel}$、$NNB_{Rel}$”分别指相邻词之间依存关系百分比和非相邻词之间依存关系百分比；“$Aver_{dis+}$、$Aver_{dis-}$”分别指不同方向依存关系的平均依存距离

从表 3 中可知，目前依存树库的平均依存距离 $Av_{dis}$、正向依存关系的平均依存距离 $Av_{dis+}$、反向依存关系平均依存距离 $Av_{dis-}$ 值分别为 2.76、2.44、3.14，即维吾尔语依存树中支配词与从属词之间一般有 1 到 2 个词；表中依存方向百分比值得注意及分析，其中 $Rel^+$ 的百分比为 99%远大于 $Rel^-$ 的百分比。虽然目前树库规模不足以对维吾尔语句子结构的倾向性下定义，但我们认为这么大比例的反差可以说明维吾尔语是支配词倾向于靠后的语言，即：一个句子中几乎所有的词都倾向依存于其后出现的词。另外，维吾尔语句子结构属于 SOV 型，我们认为 $Rel^+$ 数量如此之高不仅仅是由于谓语(谓词)出现在句末的原因，它还说明句子中其他像名词短语等语块的依存关系也是倾向于后面。需要注意的是，维吾尔语依存树库标注规范中所规定的与后置词、助动词、系动词等非内容词(non-content word)构成的依存对中，这些词标为支配词，例如，oqup boldi (读完了) 中，主要内容为 oqup (读)，boldi 为助动，但根据目前依存关系标注规范 oqup 标为从属词，boldi 标为支配词。从语言的流利度出发这种规定是可行的，但在标注过程中发现它的一些不合理性。与其他相关语言在处理这种情况的方式作比较后，我们决定应该对此加以修改。若这些依存关系的方向由原来的 $Rel^+$ 改为 $Rel^-$，显然会增加 $Rel^-$ 的数量，但新增加的数量不足以改变 $Rel^+$ 与 $Rel^-$ 之间如此悬殊的差别，所以我们认为对维吾尔语句子依存倾向性的结论仍然成立。另外，相邻依存关系(即相邻两个词之间形成的依存关系)百分比大于非相邻依存关系，但非相邻依存关系的百分比也在 40%左右，表示树库中不相邻词之间形成的依存关系不是少数。此数量是树库平均依存距离接近 3 的原因之一，同时说明“只采用相邻同现来构建人类语言复杂网络的方法可能是不恰当的”[10]。

目前的树库中我们还未发现交叉依存(即：$w_i \rightarrow w_n$，$w_j \rightarrow w_m$，其中 $i<j<n<m$ 是词在句子中出现的位置)，但这还不能说明维吾尔语依存句法中不出现交叉依存，还需要进一步验证。

(3) 依存关系与词性之间的关系

作此统计的意图是试图将依存关系描述为两种词性之间的关系，从而能否制定或列出某种依存关系的模板。下面的分析中使用了词性二元组一词，其意思是：由构建当前依存关系的两个词词性构成的二元组，结构为$<POS_1, POS_2>$(其中 $POS_1$ 是从属词的词性，$POS_2$ 是支配词的词性，$POS_1$ 与 $POS_2$ 可相同)。统计数据结果以 $Rel(Cnt_1/Cnt_2)$ 的形式给出，其中 Rel 表示依存关系名、$Cnt_1$ 表示当前依存关系在树库中出现总数、$Cnt_2$ 表示构成当前依存关系的不同词性二元组的个数。如图 5 中图例 ABL(540/25)表示树库中关系 ABL 出现次数为 540，共有 25 种不同的词性二元组构成了 ABL 关系。统计结果显示树库中出现的不同词性二元组总数为 112 个(树库使用的维吾尔语词性总数为 14 种)，构成某个依存关系的词性二元组数量范围比较广，少则 10 (INST(72/10))，多则达到 65 (COLL(1 149/65))。构成 INST 关系的不同词性二元组数量虽然为 10 种，但仅出现了 72 次，所以相比起来词性二元组还不够收敛。根据统计结果，本文认为在目前的树库规模下用词性二元组来刻画某种依存关系为时还早，仍需要扩大树库规模。

此统计中本文还观察了不同词性二元组在不同依存关系中所占的比例即二元组在依存关系中的分布。比较后发现虽然在树库中总共出现 112 种不同词性二元组，但很大部分二元组的出现频率极低，而构成依存关系最活跃的词性二元组主要集中到三种，分别为＜N，V＞，＜V，V＞，＜N，N＞（其中 N 为名词、V 为动词）。它们在 17 种依存关系（共 23 种）的构造中占据了首位，其中比例在 50％以上的有 11 种，如图 5 所示。我们认为就算考虑标注错误及其他不可预料的人为情况，比例达到 60％以上的词性二元组能够体现出此依存关系的结构倾向性，而且名词和动词本来就是某个语言的最主要部分。

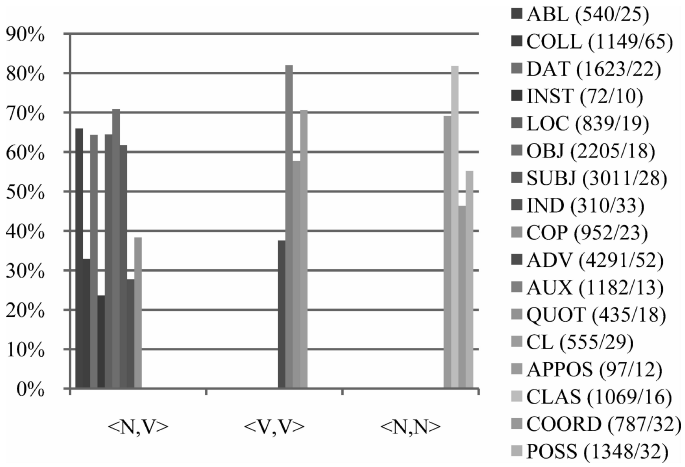


图 5 不同词性在依存关系中比例

与此同时，本文注意到了一些奇怪的现象。例如，构成 ATT 关系的词性二元组中＜V，N＞不是首位，但占的比例也不少（由于篇幅的原因，这些统计信息没在此处列出）。从语言学角度无法解释动词修饰名词的现象。通过分析发现，出现此情况的原因是目前使用的词性标注器采用的标注方案有关。即词性以当前词的词干为主，不考虑其缀接词尾后的变化，例如，chiqqan（出来的）一词的词性以词干 chiq（出来）为主标注为 V.。于是，chiqqan adem（出来的人）和 yaxshi adem（好人）的依存关系是一样的，但构成此关系的词性二元组就不一样，且出现了动词修饰名词的情况，如图 6 所示。

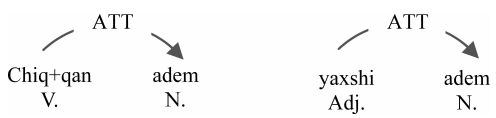


图 6 词性二元组与依存关系

可见，词尾是加以区分依存关系非常重要的特征之一。若要用词性二元组来描述依存关系即构造依存关系的模板在不考虑词尾的情况下是不完整的。这再次强调了词尾在依存关系中的重要性。但词尾层次的完全分离导致的词长度也是需要考虑的问题，特别是长句中此现象尤为突出。那么词尾的分离用什么标准为合理呢？要每层分离还是选择性的分离（只分离对依存关系起作用的词尾）？若是，那么哪些词尾对依存关系起至关重要的作用？这些都是我们需要进一步研究的问题。

## 3 总结与展望

本文阐述了维吾尔语依存树库的构建过程，包括维吾尔语依存粒度的确立、依存关系的制定、依存树库标注原则、数据库存储结构的制定及标注工具的设计与实现，最后对人工标注的 3 000 多条维语依存树从三个方面做了统计，并对结果做了相应分析。目前以自动句法分析的方式扩大了树库，其规模达到了一万多条依存树，需进一步评价其准确率。同时，对上一节中所分析及留下的问题做进一步探讨、找出适合于维吾尔语依存句法分析的方法以便快速扩建维吾尔语依存树库规模是我们下一步研究的重点。

## 参考文献

[1] Nivre J. Theory-Supporting Treebanks[C]//Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories，2003：117-128.

[2] Prokopidis P，et al. Theoretical and practical issues in the construction of a Greek dependency corpus[C]//Proceedings of the 4th Workshop on Treebanks and

Linguistic Theories (TLT-2005). Barcelona, Spain, 2005.

[3] Boguslavsky I Grigorieva S. Dependency treebank for Russian: Concept, tools, types of information[C]// Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2000: 987-991.

[4] Lepage Y, et al. An annotated corpus in Japanese using Tesnière's structural syntax[C]//Processing of Dependency-Based Grammars, 1998: 109-115.

[5] TBY L, Huury C. Dependency-based syntactic analysis of Chinese and annotation of parsed corpus[C]// Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000: 255-262.

[6] Hajičová E. Prague Dependency Treebank: From analytic to tectogrammatical annotation[C]//Proceedings of the 1st Workshop on Text, Speech, Dialogue, 1998: 45-50.

[7] Oflazer K, "Chapter 1 BUILDING A TURKISH TREEBANK[M] Build. Exploit. syntactically Annot. corpora, 2003: 1-17.

[8] Brants S, Hansen S. Developments in the TIGER Annotation Scheme and their realization in the Corpus [C]//Proceedings of the 3rd International Conference on Language Resources and Evaluation, 2002: 1643-1649.

[9] Džeroski S, Erjavec T, Ledinek N. Towards a Slovene dependency treebank[C]//Proc. Fifth Intern. …, 1388-1391, 2006.

[10] 刘海涛. 基于依存树库的汉语句法计量研究[J]. 长江学术, 2008(3): 120-128.

[11] 阿孜古丽·夏力甫. 维吾尔语动词附加语素的复杂特征研究[J]. 中文信息学报, 2008, 22(3): 105-109

[12] Zeman D, et al. HamleDT: Harmonized multi-language dependency treebank[J]. Language Resources and Evaluation, 2014, 48(4): 601-637.

[13] Haverinen K, et al. Building the essential resources for Finnish: the Turku Dependency Treebank[J]. Language Resources and Evaluation, 2014, 48(3): 493-531.

麦热哈巴·艾力(1973—),博士,副教授,主要研究领域为自然语言处理、资源建设。
E-mail: marhaba@xju.edu.cn

加米拉·吾守尔(1969—),硕士,副教授,主要研究领域为计算语言学、资源建设。
E-mail: jiamila@xju.edu.cn

吐尔根·依布拉音(1958—),通信作者,教授,主要研究领域为自然语言处理、社会计算。
E-mail: turgun@xju.edu.cn