基于依存树库的越南语依存距离研究

罗云
国防科技大学
南京
1315494900@qq.com

闫丹辉
国防科技大学
南京
diem1987@163.com

马延周
国防科技大学
南京
myz827@126.com

摘要

依存语法框架下的依存距离是衡量句法分析难度的重要指标。本文基于UD-Vietnamese依存树库对越南语依存距离的分布及影响越南语依存距离均值的因素进行分析研究。研究发现,越南语依存距离分布符合幂律分布和指数分布的混合模型;句长、长距离依存关系、依存方向均能对依存距离均值产生重要影响。该研究结果有助于从依存语法的角度揭示越南语的句法特点和规律,为提出更科学合理的依存句法分析算法提供语言学支撑。

关键词: 依存距离 ; 越南语 ; 依存树库

Vietnamese Dependency Distance: An Investigation Based on Dependency Treebanks

Luo Yun
College of international Studies,
National University of Defense Technology
Nanjing, China
1315494900@qq.com

Yan Danhui
College of international Studies,
National University of Defense Technology
Nanjing, China
diem1987@163.com

Ma Yanzhou
College of international Studies,
National University of Defense Technology
Nanjing, China
myz827@126.com

Abstract

Dependency distance within the dependency grammar framework serves as a significant metric for assessing the difficulty of syntactic parsing. This paper analyzes the distribution of dependency distance in Vietnamese and the factors influencing the mean dependency distance (MDD) of Vietnamese, based on the UD-Vietnamese dependency treebank. The study reveals that the dependency distance distribution of Vietnamese conforms to a mixed model of power-law and exponential distributions. Sentence length, long-distance dependencies, and dependency direction all exert a significant impact on the mean dependency distance. The findings of this research contribute to unveiling the syntactic characteristics and patterns of Vietnamese from the perspective of dependency grammar, and provide linguistic support for developing more scientifically sound and rational dependency parsing algorithms.

Keywords: Dependency Distance , Vietnamese , Dependency Treebank

1 引言

依存句法分析是自然语言处理的核心技术，其目标是分析句子中词语间的依存关系，输出句子的句法结构，这对于信息提取和语言生成等应用至关重要。然而，长句子的复杂性增加了句法分析的难度。为了提升越南语依存句法分析的性能，研究造成分析困难的原因非常重要。依存距离(DD) 作为衡量句子中支配词和从属词线性距离的指标，由Hudson(1995)首次提出。依存距离与认知负荷密切相关。依存距离的增加会导致词语在工作记忆中停留时间延长，增加语言加工的复杂性。依存距离最小化是语言中普遍存在的现象，已被大量研究证实，并且能有效降低认知负担。刘海涛(2008)提出的依存距离均值(MDD) 概念及跨语言研究进一步支持了依存距离最小化倾向。Futrell(2015)对多种语言的研究也印证了这一结论. 依存距离最小化与人类认知机制和“省力原则”相关。说话者倾向于使用短依存关系的句子结构，因为这类句子更易于理解和处理。影响依存距离的因素很多，包括句长、交叉依存、语法特征等(Cowan,2001)。学界已对这些因素进行了研究，例如刘海涛(2016)系统分析了影响MDD的关键因素，Temperley(2008)提出了依存距离“同分支”原则。

目前，已有研究人员运用系统功能语法(周晶晶,2016) 或对比分析方法研究越南语的句法特征(蒋跃,2021)，但较少将其有效应用于自然语言处理任务。虽然有研究关注依存句法对越南语自然语言信息处理的优化潜力，并积极构建依存树库(李英,2017)，但针对越南语依存距离的专门研究仍旧缺乏。现有研究主要从依存句法理论角度出发，且国内外都少有专门针对越南语依存距离的研究。

本文旨在填补这一研究空白，通过对越南语依存树库的计量分析，具体回答以下核心问题：第一，越南语依存距离的整体分布呈现何种数学规律？第二，哪些是影响其依存距离均值（MDD）的关键因素？为此，本研究将重点考察句长、长距离依存关系的类型与频率、以及依存方向（支配词居前vs. 支配词居后）对MDD的具体影响。

对上述问题的回答，将有助于从依存语法的角度揭示越南语的句法特点和规律(李良毅,2023)，为未来研究长距离依存的句法模式、构建更科学的句法分析特征集、并提出更优的依存句法分析算法奠定语言学基础，最终提升越南语自然语言信息处理能力。

2 资源与方法

2.1 依存树库的选择

通用依存关系树库（Universal Dependencies, UD）是一个旨在为世界多种语言提供统一标注框架的合作项目 [UD Projectn.d.]，它使得跨语言的句法研究和比较成为可能。本研究直接使用这些由前人构建和发布的公开语料资源。

目前，越南语的依存树库大多由既有的成分句法树库转换而来。本研究选取了三个公开的越南语依存树库，它们的具体来源和构建背景如下：

- **VTB 和TueCL**：这两个树库是UD 项目中UD-Vietnamese 的组成部分，其标注遵循UD框架的通用规范，VTB其数据源于越南语言与语音处理项目（VLSP）的成分树库，先以非UD风格进行手动标注，后续自动转换为UD格式。是一个原生的UD树库，其语料直接依据UD风格进行从零开始的手动标注，而非从其他格式转换而来。

- **VnDT (Vietnamese Dependency Treebank)**：该树库同样是越南语研究中广泛使用的数据集，目前可通过SEACrowd 项目获取。VnDT 原始的构建方法，包括从越南语成分树库（Vietnamese Treebank）的转换规则、中心语选择策略（head-percolation rules）以及依存关系标签的详细定义，均在Nguyen(2014) 的工作中得到了详细阐述。

表1展示了这三个树库的基本统计信息。

2.2 依存距离计算方法

依存语法关注句子中词语间的关系。句子中词语间构成非对称二元关系，即在构成关系的两个元素中，一个处于支配地位（支配词），另一个处于从属地位（被支配词）(刘海涛，2007)。依存方向由支配词指向被支配词，通过弧线连接两个具有句法关系的词，弧上标记表示依存关系类型。依存语法理论认为，词语间的依存关系是人类表达连贯思想的基础(刘海涛，2007)，也是形成句法难度的根本原因。

树库名称	句子数	词数	平均句长	体裁	领域
VTB	3323	58069	17.47	新闻, 博客	政治、经济、文化、艺术、科技、社会、财经、教育
TueCL	100	1888	18.88	演讲	教育、文化
VnDT	10200	220000	21.57	新闻	政治、经济、文化、艺术、科技、社会、财经、教育

Table 1: 越南语依存树库统计信息(VnDT数据根据其原始论文更新)

依存距离表示支配词与从属词间关系的“远近”，其值由两者间隔的词汇数量表征。通常，句子依存距离越长，处理难度越大。为减少句长对依存距离的影响，本研究采用刘海涛提出的依存距离计算方法，将句子中词汇依次编号。因此，一个由l个词构成的句子，其依存距离d可表示为公式:$d = \sum_{i=1}^{l} dd_i$其中，dd_i表示句子中第i个词的支配词序号减去从属词（本身）序号所得差值，即依存关系中两词间的序号差。该差值有正负之分，用以表示依存方向。为便于统计分析并发现语言规律，需确保所有差值为正，因此在公式中，dd_i需取绝对值。每个句子有且仅有一个根节点（root），其无支配词，或可认为其支配词为其本身，因此，本文将根节点依存距离定义为0。由l个词构成的句子，其依存距离均值$overlined$ 可表示为：$\overline{d} = \frac{1}{l-1} \sum_{i=1}^{l} |dd_i|$ 假设抽取n 个句子建立依存树库，且依存关系总数为m，用$\overline{d}$表示第i个依存关系的距离，则这n个句子的总依存距离D与依存距离均值$\overline{D}$分别为$D = \sum_{i=1}^{m} |DD_i|$和$\overline{D} = \frac{1}{m-n} \sum_{i=1}^{m} |DD_i|$。

3 结果与讨论

利用EmEditor、Excel、Altmann-Fiteer、Matlab及Matplotlib等软件对越南语三个树库的依存距离进行统计分析与量化对比研究。

3.1 依存距离的分布规律

依存距离分布受语言类型、句长、文本数量等多种因素影响，目前学界对于其是否符合单一模型尚无定论。刘海涛(2007)认为汉语依存距离分布符合离散幂律分布。而陆前和刘海涛(2016)对30种语言的真实语料进行依存距离分布研究后指出，在不限定特定句长集合等条件下，依存距离分布更符合幂律与指数的混合分布模型。华英楠(2022)基于依存树库，对朝鲜语依存距离展开研究，指出朝鲜语依存距离分布受多种因素影响。

为精确探究越南语依存距离的分布规律，本文采用了专门用于语言数据拟合的软件Altmann-Fitter，对三个树库的依存距离频次数据进行拟合。我们选用混合几何函数（Mixed Geometric Distribution）对数据进行建模。该模型假设观察到的分布是由两个不同的简单几何过程混合而成，其概率质量函数通常表示为:

$$P(k) = c \cdot p_1^k + (1 - c) \cdot p_2^k$$

其中，k是依存距离，$P(k)$是距离为k的概率。p_1和p_2是两个几何分布的参数，而c是混合比例系数。这些参数的具体数值由软件通过非线性最小二乘法进行估计。虽然因篇幅所限未在此一一列出所有参数，但模型的整体拟合效果由表2中的优度检验指标来衡量。

树库	x^2	DF	$P(x^2)$	C	R^2
VTB	321.30	47	0	0.0059	0.9990
TueCL	20.64	29	0.8721	0.0115	0.9991
VnDT	2318.20	76	0	0.0056	0.9990

Table 2: 混合几何分布模型拟合优度检验结果

对表2中的拟合优度检验结果进行分析：首先，R^2（决定系数）是衡量模型对数据解释程度的指标，其值越接近1，表示模型解释力越强。所有树库的R^2值均大于0.999，表明模型可以解释观察数据中超过99.9%的方差，拟合曲线与数据点的相关性极高。

其次，对于卡方检验（x^2 test），尽管VTB和VnDT树库的x^2值较大（导致$P(x^2) = 0$），但这在处理大样本语料时是常见现象，因为即使微小的偏差也会被样本量放大。因此，我们更关注对样本大小不敏感的差异系数C（Coefficient of Difference）。在定量语言学中，$C < 0.02$通常被认为是良好拟合的标志。本次实验中，所有树库的C值均小于0.02，其中VTB和VnDT的C值更是小于0.01，表明拟合效果非常好。

注: x^2 为卡方值, DF 为卡方检验的自由度, $P(x^2)$ 为卡方值概率, C为差异系数, R^2为拟合决定系数。

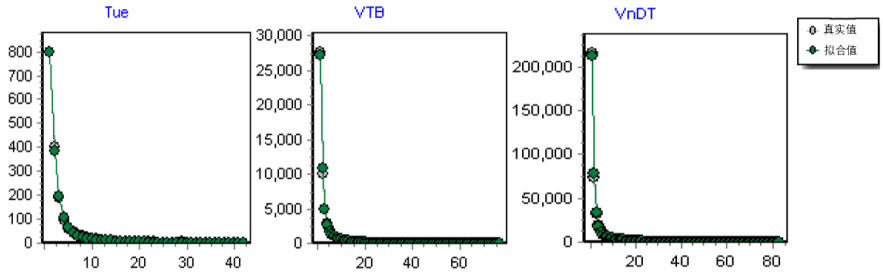

Figure 1: 越南语依存距离频次随距离增加的变化情况图(横坐标为依存距离, 纵坐标为频次)

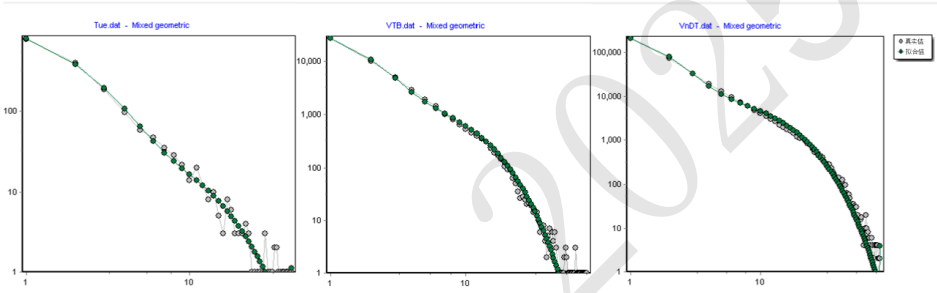

Figure 2: 越南语依存距离频次随距离增加的变化情况（对数-对数坐标图）

从图1和图2的分布形态来看，三个树库的依存距离分布在对数-对数坐标下（图2），短距离部分（左侧）近似于直线，呈现幂律分布特征；而长距离部分（右侧）则快速下降，呈现指数分布特征。这直观地验证了混合模型描述的合理性。

综上所述，无论是从拟合优度的定量指标（R^2, C）还是从分布形态的直观观察来看，越南语的依存距离分布都非常符合幂律与指数分布的混合模型。这揭示了语言系统内在的自组织性：一方面，通过少数高频的短距离依存（幂律特征）来遵循“省力原则”；另一方面，通过指数衰减的长距离依存来避免过高的认知加工负荷。

3.2 影响越南语依存距离均值的因素探讨

依存距离是依存关系的重要属性，可用于表征处理依存关系的认知成本。依存距离指词与词之间的线性距离。在比较单一语种不同语句、不同文本或不同语种之间的认知差异时，单纯考虑影响词与词之间依存距离大小的因素意义不大。基于此，刘海涛(2007) 提出了依存距离均值（Mean Dependency Distance, MDD）的概念及其计算方法。MDD能够反映语句、文本或语种的整体理解难度。对越南语的MDD及其影响因素展开研究，既有助于精确考察依存距离均值与各影响因素的定量关系，归纳总结语言的一般规律；又有助于从认知与功能层面更深入地研究影响句法难度的因素，从而更好地预测句法难度，为基于依存语法的句法分析提供更为可靠的语言学支持。

对三个SUD树库分别进行计算，得到如下结果：$MDD_{VnDT} = 3.02$，$MDD_{VTB} = 2.75$，$MDD_{GSD} = 2.94$。对现有的越南语SUD树库的计算结果显示，越南语的$MDD_{max} = 3.02$ 仍在刘海涛（2008）给出的MDD可能上限界值4以内($MDD_{max} < 4$)，越南语的MDD_{max}

在20种语言中排名第四。尽管MDD的具体数值会受到多种因素的影响，但其取值变化一般不会太大；上述计算结果初步揭示了越南语MDD的大致取值区间及其排名情况。

现有对其他语言的研究成果表明，句长、交叉依存、长距离依存关系、依存方向、组块、依存树库的选择及标注等因素均对MDD产生不同程度的影响。此外，依存方向对基于转移的依存句法分析算法的选择也是至关重要的。为此，下文基于VTB树库，重点分析句长、长距离依存关系及依存方向对越南语MDD的影响。

3.2.1 句长与依存距离均值的关系

刘海涛(2010) 通过对汉语树库的研究发现，句子的MDD与平均句长有关。Jiang和Liu(2015) 的研究结果亦显示，MDD会随着句长的增加而缓慢增加。为了探讨越南语句长对MDD的影响，下文将分别对不同句长区间的MDD及相同句长的MDD大小变化进行统计分析。在VTB树库中，句长最短为3，最长为135。基于句长将树库分为[3–34]、[35–66]、[67–98]和[99–135]这四个区间，以分析不同句长区间对MDD大小的影响。不同句长区间的句子数量及其MDD的统计结果如表3所示。图4对不同句长区间中的句子数量和MDD进行了归一化处理，其数值的相对大小不影响分布结果。

句长区间	句子数	占比	平均句长	平均依存距离(MDD)
[0–30]	2921	85.33%	17.85	2.64
[31–60]	457	13.35%	38.02	3.33
[61–90]	36	1.05%	70.94	4.15
[91+]	9	0.26%	106.78	4.42

Table 3: 不同句长区间的句子数量及其MDD统计结果(表格3)

根据图3中红色折线可知，在VTB树库中，不同句长下句子数量分布整体呈现幂律分布的长尾特征。在[0–30]区间的句子数量占总树库的85.33%左右，当句长扩大到60时，句子数量占总树库的98.68%，几乎涵盖了整个树库。这表明句长大于或等于60的句子使用频率极低，而句长位于[0–60]区间的句子使用频率较高，因此，对VTB树库MDD产生影响的句长主要集中在[0–60]区间。从图4中的蓝色折线可知，不同句长区间的MDD存在差异，但整体上MDD随着句长的增长而缓慢增大，这说明句长是影响越南语MDD的关键因素。为进一步探究句长对MDD变化的影响程度，下文将继续对相同句长的MDD进行统计分析。由于VTB树库的句子主要集中在句长为[0–60]区间，因此选取该区间的中间值，即句长为30时的句子进行MDD统计分析。在VTB树库中，句长为30的句子共有80句，按照句子的MDD升序排列后依次编号为S1–S80，计量结果如图3所示。

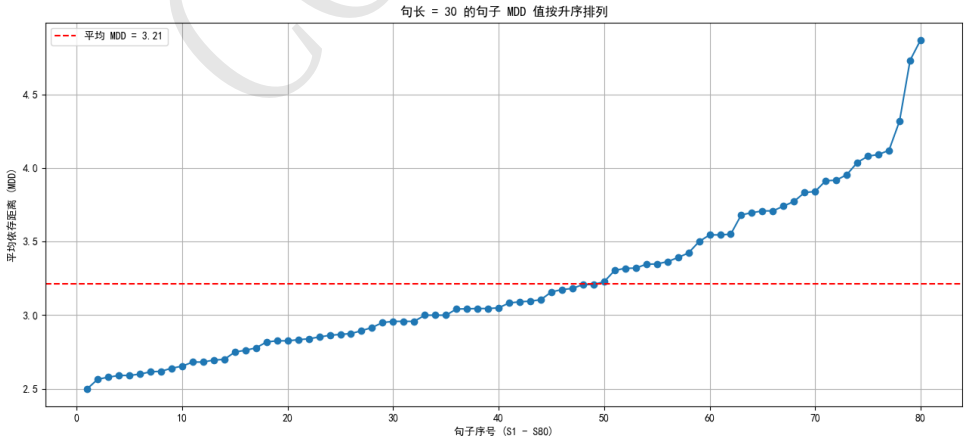

Figure 3: VTB树库句长= 20 时单个句子MDD 升序排列图

图4显示MDD随句长缓慢增长。基于原假设，即句长是影响MDD的主要因素,可推断句长为30时MDD应在较小范围内波动。然而,图??显示句长为30时,句子MDD最大值为4.87,最

小值为2.5,标准差为0.52,与图??中MDD随句长变化的标准差0.77相当,极差为2.37,大于图??中的1.53。这表明在相同句长下,越南语MDD波动较大,与原假设不符。因此,句长仅是影响MDD的众多因素之一,越南语的MDD受句长和其他因素共同作用。后续章节将从越南语句子结构角度深入分析影响MDD的因素。

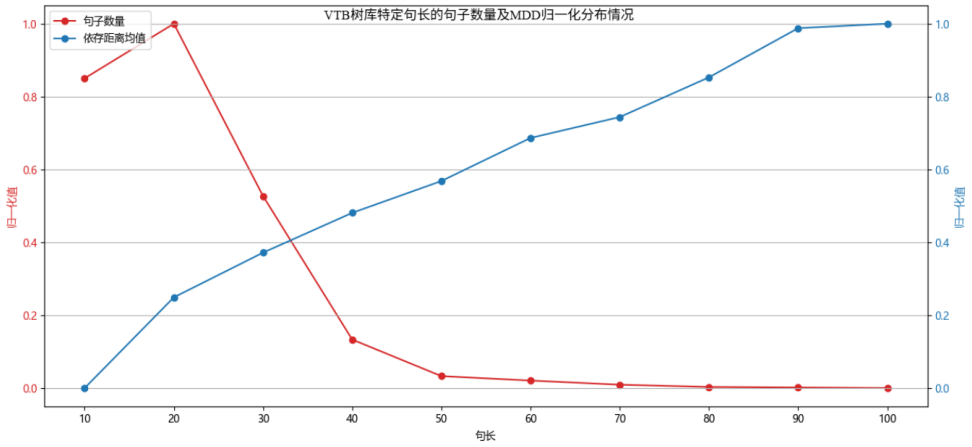

Figure 4: VTB树库不同句长区间句子数量及MDD 归一化后的分布情况

3.2.2 长距离依存关系与依存距离均值的关系

长距离依存是指句法上具有依存关系的两个成分之间插入较多其他成分，导致依存距离过大的现象。近年来，有研究表明，支配词与从属词间距离越远，加工过程需更多能量以重新激活词语。这种处理困难很可能源于工作记忆的时间衰减或干扰，而非仅仅是记忆容量问题(刘海涛，2017)。受人类认知机制和"省力原则"影响，语言存在依存距离最小化趋势(Futrell,2015)。如果语言使用者倾向于避免困难，他们会回避长距离依存关系。此外，语言若已进化到支持简单交流，理应避免造成长距离依存关系的语序。

对VTB树库的分析表明，在实际语言使用中，相邻依存关系比例约为50.62%，当依存距离扩大到2时，依存关系占比达69.04%，当依存距离扩大到3时，依存关系占比达83.40%，这表明在使用越南语时，人们也倾向于使用更易处理的短依存关系。为深入探讨长距离依存与MDD的关系，下文选取VTB树库中句长为30的句子进行统计。直观上，当句长相同时，句中长距离依存关系越多，MDD越大。下文将统计长距离依存关系类型及其比例，并结合越南语语法特征，以揭示长距离依存关系如何影响MDD数值。为此，下文对VTB树库中依存距离大于等于5（依存距离≥ 5）的长距离依存关系进行了统计，并分析了占比最高的6种依存关系类型。排名前6位的类型及其统计为：`punct` (137个, 43.49%)、`conj` (42个, 13.33%)、`obl:tmod` (23个, 7.30%)、`nsubj` (22个, 6.98%)、`obl` (17个, 5.40%)、`advcl` (10个, 3.17%)。

根据统计结果可知，句长为30 的句子中，依存距离大于等于5 的依存关系有315 个，占总依存关系数量的15.99%。在依存距离大于等于5 的长距离依存关系中，占比最高的依存关系类型依次是`punct`、`conj`、`advcl`、`nsubj`、`obl:tmod`，这5 种依存关系类型的占比约为79.67%，如图5所示。

长距依存关系类型	数量	占比
punct	137	43.49%
conj	42	13.33%
obl:tmod	23	7.30%
nsubj	22	6.98%
obl	17	5.40%
advcl	10	3.17%

Table 4: 树库长距离依存关系占比前6的类型及数量

第二十四届中国计算语言学大会论文集，第587页-第597页，济南，中国，2025年8月11日至14日。

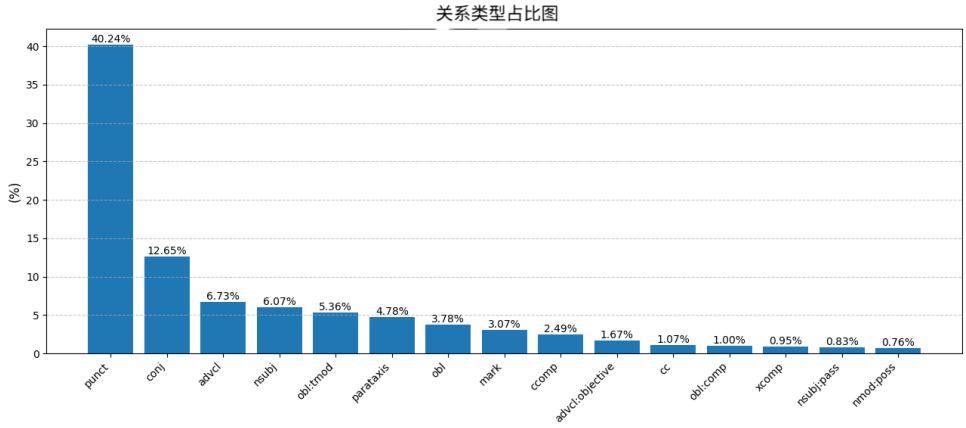

Figure 5: 树库长距离依存关系类型及数量分布图

进一步对整个VTB 树库中依存距离大于等于5 (≥ 5) 的长距离依存关系进行统计，结果显示，占比最高的5 种依存关系类型如表4。VTB树库中，依存距离大于等于5 的依存关系数量为9089个，占总依存关系数量的10.97%。在依存距离大于等于5 的长距离依存关系类型中，占比最高的依次是`punct`、`conj`、`advcl`、`nsubj`、`obl:tmod`、`parataxis`，这6 种依存关系类型总占比达到83.33%。该结果与句长为30时长距离依存关系的统计结果相一致，说明长距离依存关系占比是影响依存距离均值(MDD)的主要因素之一，且导致句子MDD 偏大的依存关系类型相对固定，主要为：`punct`、`conj`、`advcl`、`nsubj`、`obl:tmod`、`parataxis`。此外，对中长距离依存关系中依存距离值为1的依存关系类型进行了统计，总依存关系数：56534，依存距离为1的关系数：14550，依存距离为1的关系占总体比例：25.74%。依存关系类型占比最高的依次是`obj`、`compound`、`nmod`、`xcomp`、`amod`，如图6所示。

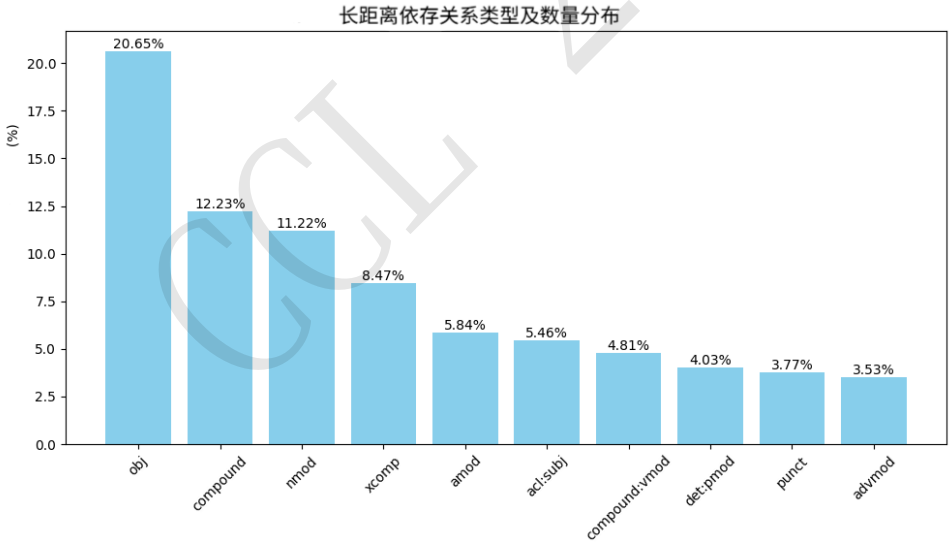

Figure 6: 长距离依存关系类型及数量分布图

下文将从越南语句法层面，针对句中长距离依存关系中占比最高的`punct`（标点符号）与`conj`（连词）两类进行深入探讨。越南语句子结构灵活、词序较自由，因而标点符号在语义分割与句子衔接中常跨越多层语法成分，形成较长依存路径；而连词在实现并列、转折、递进等语法功能时，往往跨越主句与附加信息之间的空隙，依存距离亦随之显著增大。此两类长距离依存现象不仅反映了越南语表达复杂语义时的弹性特征，也对依存句法解析构成挑战。

针对句中依存距离值为1的长距离依存关系，依存关系类型占比最高的依次

是obj、compound、nmod、xcomp与amod五种，分别对应宾语、复合词结构、名词修饰、补充说明及形容词修饰。从依存距离为1的统计现象来看，各结构成分均以紧邻方式构成依存关系，反映出核心词先行而修饰成分随后跟出的排列模式。动词与宾语、名词与定语或形容词之间往往相连，这表明越南语在传递信息时，主要内容（动词、中心名词等）先出，而用于补充限定的修饰词则后置，使语义层次逐步展开。由此可见，越南语在句法结构上具有“前正后偏”的特色：即主要信息先行，辅助或修饰信息后置。

如前文所述，句长并非影响MDD的唯一因素。在此基础上，我们可以进一步探讨句长与句子内部结构，特别是长距离依存关系之间的交互作用。句长为长距离依存的产生提供了必要的线性空间，可以说，一个句子越长，其内部出现长距离依存关系的可能性和数量就越大。这一互动关系为在上一节观察到的现象提供了深层解释：即便句长固定（如30个词），句子的MDD依然存在巨大差异。这正是因为长度相同的句子，其内部长距离依存关系的数量、类型和布局可能迥然不同。因此，可以将句长视为决定句子复杂度的宏观基础变量，而长距离依存关系则是更微观、更直接的核心影响因素。

3.2.3 依存方向与依存距离均值的关系

越南语依存方向的分布特征 依存方向是指在句子中存在依存关系的两个语言单位的相对位置。依存距离的正负可以表示依存方向的类型，依存距离为负的依存关系称之为支配词居前的依存关系，即head-initial；依存距离为正的则称之为支配词居后的依存关系，即head-final。不同的语言，其依存方向具有不同的分布特征，依存方向可以作为判定语序类型的指标。每种语言都包含或多或少的支配词居后或支配词居前依存关系，一些语言始终倾向于”支配词居后”（如日语、土耳其语），而一些语言始终倾向于”支配词居前”（如威尔士语、阿拉伯语、丹麦语），多数语言倾向于依存方向的均衡分布（如英语、法语、希腊语等）。为了探究越南语依存方向分布特征，下文对VTB树库中依存方向的分布进行统计，结果如表5所示。

依存方向类型	数量	占比
支配词居后(head-final)	24130	44.08%
支配词居前(head-initial)	30616	55.92%
总计	**54746**	**100%**

Table 5: 树库中依存方向类型及数量

根据统计结果表5可知，越南语的依存方向倾向于支配词居前(head-initial)，属于支配词居前略占优势的语言类型。从语法角度来看，越南语与汉语和英语同属主谓宾（SVO）语序结构的语言。然而，越南语的特点在于定语置于被修饰词之后，这意味着在定语结构中，支配词始终位于其前置位置。

依存方向类型	依存距离均值
支配词居前(head-initial)	2.43
支配词居后(head-final)	3.46

Table 6: 支配词居后和支配词居前的依存距离均值

依存方向对依存距离均值的影响 Dryer (1992)的研究发现，如果一个支配词有几个从属词，把它们都放在头部的同一侧会产生一种拥挤效应，如果从属词在支配词两侧保持平衡，则可以减少依存距离。实现这种平衡的一种方法是，规定一种语言的主要分支方向（如右分支），但允许一些短的依存短语向相反的方向分支(Dryer，1992) 。Hawkins 的EIC 理论预测，像英语这样以右分支为主的语言中，左分支的成分往往很短(Temperley，2018) 。

越南语作为支配词居后型（左分支）语言，支配词居前（右分支）的依存距离是否通常也较小？为此，下文对VTB树库中不同依存方向的MDD 进行统计，结果如表6所示。

根据统计结果可知，越南语支配词居后的依存距离均值为2.43，明显低于支配词居前的依存距离均值3.46。由此，产生以下两个问题：(1) VTB树库中不同句长支配词居后的MDD 是否均高于支配词居前的MDD；(1) VTB树库中不同句长支配词居前的MDD 是否均高于支配词居后的MDD；(2) 如果(1) 成立，是什么原因造成了该现象。

为了验证(1)，从VTB树库中抽取句长为10、20、30、40 的句子，分别构建3 个子库，并计算不同依存方向类型下的MDD，结果如表7所示。表7 显示，句长为10、20、30、40时，支配词居前的MDD 值均高于支配词居后的MDD 值，且相差较为明显，差值均在0.7 以上，进一步验证了表6的统计结果。

句长	依存方向类型	所占比例	依存距离均值
10	支配词居后(head-final)	43.22%	1.87
	支配词居前(head-initial)	56.78%	2.58
20	支配词居后(head-final)	43.91%	2.12
	支配词居前(head-initial)	56.09%	3.30
30	支配词居后(head-final)	44.51%	2.77
	支配词居前(head-initial)	55.49%	3.60
40	支配词居后(head-final)	46.60%	2.72
	支配词居前(head-initial)	53.40%	4.29 [?]

Table 7: 不同依存方向下依存距离均值

表7显示，支配词居前的MDD 随句长的增长缓慢增大。为探究该情况发生的原因，将通过分析VTB 树库中支配词居前的依存关系，考察其中排名前四的类型分布（如表8所示）。

类型	数量
nsubj	4045
punct	3404
case	3132
advmod	2539
nummod	1279
mark	1096

Table 8: 支配词居前中依存关系类型分布

表8显示，支配词居前中占比最高的依存关系类型是nsubj，即存在大量的名词短语结构。在越南语中，倾向于把第一个名词作为名词短语的head，且名词与名词之间的依存距离为1。conj依存关系类型倾向于把第一个成分作为支配词，当连接成分为词汇时,依存距离为1,只有当连接成分是句子时,支配词与从属词之间的介入成分较多,依存距离也随之变大。但由于支配词居前中依存关系的距离普遍为1,所以少量长距离依存未能整体提升支配词居前的MDD。限于篇幅,此处不再对其它依存关系类型的依存距离做详细分析。

综上所述，针对前述两个问题可得出以下结论：(1) 当选取不同句长时,均能得出支配词居前的MDD明显高于支配词居后的MDD; (2) 造成(1)的原因为:越南语作为左分支占优势的语言,其右分支的成分多由名词短语结构、同位语结构以及连接成分结构组成,结构简单,依存距离普遍较短。对于所有支配词居前型语言，其支配词居后的MDD是否普遍较小，仍需进一步探讨。根据表7、表8的统计结果,随着句长增大,越南语支配词居后的MDD始终保持在2以内,支配词居后的MDD是否存在阈值有待进一步研究。

4 结语

本文基于UD-V依存树库开展越南语依存距离的研究。通过对越南语依存距离分布的研究发现,不同UD树库下越南语依存距离分布具有一定规律,都符合幂律分布长尾效应;进一步对句长=20的3个子库的依存距离拟合后发现,越南语依存距离分布符合幂律分布和指数分布的混合模型。在对影响越南语MDD的因素分析中发现:句长、长距离依存关系、依存方向均能对MDD取值产生重要影响。具体表现为:(1) MDD随着句长增加而缓慢增大,但在句长固定时,MDD浮动较大,表明句长仅是影响MDD的因素之一;(2) 句子中长距离依存关系类型占比是影响句子MDD的最为关键因素,长距离依存关系类型占比越高,句子的MDD越大;(3)越南语是支配词居前型语言,因其左分支结构成分简单,使得支配词居后的MDD小于支配词居前的MDD。人类对语言的组织生产是一个复杂的系统工程。然而，本文的研究也存在一些局限性，这些局限性为未来的工

作指明了方向。首先，现有树库的体裁以新闻和演讲为主，这是一种相对正式和规划性较强的书面语体。研究结论是否能同样适用于口语、社交媒体文本、文学作品等其他多样化的语体，尚需检验。不同体裁的句法模式和复杂度差异，可能会对依存距离的分布和均值产生影响。其次，尽管本研究采用的树库均遵循通用依存关系（UD）框架，但它们由不同机构在不同时期构建和转换，在处理某些具体语言现象时（如并列结构、省略等），标注实践上可能存在细微差异。这种潜在的不一致性可能为跨库的数据整合与对比分析带来一定的噪声。因此，未来的研究可以致力于构建规模更大、体裁更多样化的越南语依存树库，并对标注规范进行交叉验证和细化，以便在更坚实的数据基础上，深化对越南语乃至更广泛语言的句法加工机制的理解。

致谢

本文的研究工作得到了国防科技大学自主创新科学基金项目“面向特定人物的语音伪造关键技术研究”（项目编号：25-ZZCX-JDZ-46）的资助，特此致谢。

参考文献

Cowan, N. 2001. The Magical Number 4 in Short-term Memory: A Reconsideration of Mental Storage Capacity. *Behavioral and Brain Sciences*, (4).

Dryer, M. 1992. The Greenbergian Word Order Correlations. *Language*, (8).

Ferreri Cancho, R. 2004. Euclidean Distance Between Syntactically Linked Words. *Physical Review E*, (7).

Futrell, R., Mahowald, K., and Gibson, E. 2015. Large Scale Evidence for Dependency Length Minimization in 37 Languages. *Proc Natl Acad Sci USA*, (3).

Futrell, R., Levy, R., and Gibson, E. 2020. Dependency Locality as an Explanatory Principle for Word Order. *Linguistic Society of America*, (2).

Jiang, J.-Y. and Liu, H.-T. 2015. The Effects of Sentence Length on Dependency Distance, Dependency Direction and the Implications Based on a Parallel English-Chinese Dependency Treebank. *Language Sciences*, (5).

蒋跃, 范璐, 王余蓝. 2021. 基于依存树库的翻译语言句法特征研究. 外语教学, 42(3): 41–46.

李良毅, 张亚飞, 郭军军, 等. 2023. 融入依存句法信息的事件时序关系识别. 计算机工程与应用, 59(7): 110–117.

李英, 郭剑毅, 余正涛, 等. 2017. 越南语短语树到依存树的转换研究. 计算机科学与探索, 11(4): 599–607.

梁君英, 刘海涛. 2016. 语言学的交叉学科研究: 语言普遍性、人类认知、大数据. 浙江大学学报*(*人文社会科学版*)*, (1).

刘海涛. 2007a. 泰尼埃的结构句法理论. 北华大学学报*(*社会科学版*)*, (5).

华英楠, 毕玉德. 2022. 基于依存树库的朝鲜语依存距离研究. 外语学刊, (6): 55–65.

Liu, H.-T. 2007b. Probability Distribution of Dependency Distance. *Glottometrics*, (5).

Liu, H.-T. 2008a. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, (2).

刘海涛. 2008b. 基于依存树库的汉语句法计量研究. 长江学术, (3).

刘海涛. 2009. 依存语法的理论与实践. 科学出版社, 北京.

Liu, H.-T. 2010. Dependency Direction as a Means of Word-Order Typology: A Method Based on Dependency Treebanks. *Lingua*, (6).

Liu, H.-T., Xu, C.-S., and Liang, J.-Y. 2017. Dependency Distance: A New Perspective on Syntactic Patterns in Natural Languages. *Physics of Life Reviews*, (21).

刘海涛. 2018a. 计量语言学进展. 浙江大学出版社, 杭州.

刘海涛, 林燕妮. 2018b. 大数据时代语言研究的方法和趋向. 新疆师范大学学报*(*哲学社会科学版*)*, (39).

陆前, 刘海涛. 2016. 依存距离分布有规律吗? 浙江大学学报*(*人文社会科学版*)*, (4).

Surface Syntactic Universal Dependencies project. n.d. Surface Syntactic Universal Dependencies. Available from: https://surfacesyntacticud.github.io/data/.

Temperley, D. 2008. Dependency-length Minimization in Natural and Artificial Language. *Journal of Quantitative Linguistics*, (3).

Temperley, D. and Gildea, D. 2018. Minimizing Syntactic Dependency Lengths: Typological/Cognitive Universal? *Annual Review of Linguistics*, (4).

Universal Dependencies project. n.d. Universal Dependencies Introduction. Available from: https://universaldependencies.org/introduction.html.

Wang, Y.-Y. and Liu, H.-T. 2017. The Effects of Genre on Dependency Distance and Dependency Direction. *Language Sciences*, (59).

Woo, Y.-M., Song, Y.-I., Park, S.-Y., and Rim, H.-C. 2007. Modification Distance Model Using Headible Path Contexts for Korean Dependency Parsing. *Information Science Society*, (34).

Yan, J.-W. and Liu, H.-T. 2019. Which Annotation Scheme Is More Expedient to Measure Syntactic Difficulty and Cognitive Demand? In: Chen, X.-Y. and Ferreri Cancho, R. (Eds.), *Proceedings of the First Workshop on Quantitative Syntax (Quasy)*. Association for Computational Linguistics, Paris.

周晶晶, 周枫, 严馨. 2016. 基于依存树的越南语新闻事件元素抽取. 计算机工程与设计, 37(8): 2233–2237.

Nguyen, D. Q., Nguyen, D. Q., Pham, S. B., Nguyen, P.-T., and Nguyen, M. L. 2014. From Treebank Conversion to Automatic Dependency Parsing for Vietnamese. In: *Proceedings of the EACL 2014 Workshop on Lexical and Grammatical Resources for Language Processing (LG-REL)*. Gothenburg, Sweden, pages 63–67.