

基于层级距离的朝鲜语句法处理难度研究*

华英楠 毕玉德

[提要] 依存语法框架下的依存距离和层级距离已被证实可作为句子处理难度的衡量指标。本文基于 UD、SUD 依存树库开展了朝鲜语层级距离分布规律、平均层级距离和平均依存距离相互关系的研究。研究结果显示：两树库虽然在层级距离单调性上存在一定差异，但整体上都符合长尾分布规律，呈现最小化倾向；根词及其从属词的词类分布和配价能力以及构成的依存关系类型是影响层级距离分布规律的本质原因。两树库的平均层级距离和平均依存距离相互制约，且两者均值未超过4，使句法处理难度整体处于一种动态权衡状态。

[关键词] 朝鲜语 依存树库 层级距离 依存距离 句法处理难度

一 引言

句法结构包括一维线性顺序（水平）和二维层级结构（垂直）。通常情况下，两者具有映射关系，层级结构派生线性顺序（程工 2018），即二维的层级结构可按时间序列在水平方向上形成投影，将层级结构和线性顺序联系起来（刘海涛、敬应奇 2016）。句子的处理难度不仅体现在线性维度还体现在层级维度，与语言的线性句法处理相比，句子的层级结构涉及句法结构的形成过程，层级句法处理被认为是一种更高层次、更复杂的语言处理过程，它涉及对句子根词^①的预测及对整个句子句法、语义结构的处理（牛若晨 2021:117）。对层级结构进行分析，可以更科学地揭示出句子的内部构造，对于正确理解句子的语义有重要的作用。

为了对句子的层级结构进行量化研究，刘海涛、敬应奇（2016）基于依存语法理论和依存句法树，并参照线性维度上依存距离（Dependency Distance, DD）的计算方法创新性地提出了层级距离（Hierarchical Distance, HD）这一度量指标，以期更加直观地了解句子的层级结构。在依存句法树中，层级距离表示句子中各词语到根词的垂直距离（刘海涛、敬应奇 2016；刘海涛 2022:168）。据 Yngve（1960）的深度假说和 Hudson（2007:36-41）的扩散性激活理论，层级距离越大，则词语的层级越深、离根词越远，激活难度就越大。故而层级距离的大小反映了从句子根词处出发激活一个词的难度（刘海涛、敬应奇 2016；牛若晨 2021:26）。

目前针对层级距离的研究主要包括两个方面：一方面是针对不同语言的层级距离进行量

* 本文获国家社科基金一般项目“朝鲜语语言知识结构化建模及深度语义理解研究（22BYY163）”资助。论文写作过程中，刘海涛老师、匿名审稿专家以及编辑部老师提出了宝贵的修改意见。谨致谢忱！文中错误与纰漏之处，均由笔者负责。

^① 根词，又称为根节点（root），是句子的中心节点。在依存句法树图中，根词位于最高层级，句子中所有其他子节点都直接或间接地受根词支配。

化分析,以寻求其层级距离函数分布和演化规律的异同。相关研究主要集中在刘海涛、敬应奇(2016)和刘海涛(2017)。另一方面,因为依存距离和层级距离均可以用来表征句子不同维度的处理难度,所以为了更全面地理解不同语言的句法结构在处理难度方面的相同点和差异性,有些学者(Jing & Liu 2015; Chen et al. 2021; 牛若晨 2021:32-143)联合平均依存距离(Mean Dependency Distance, MDD)和平均层级距离(Mean Hierarchical Distance, MHD)开展了对句法处理难度的影响研究。

朝鲜语在句法处理难度方面主要采用眼动追踪实验、事件相关电位等心理实验方法针对定语从句、状语从句等特定句法结构进行了处理难度的对比研究(Lee et al. 2007; Kwon et al. 2013; 白俊伍等 2019)。通过心理实验可以直观地记录句子的处理过程,观察句子的处理难度,可信度较高,但由于实验所选用的语料是人为设计的,附加了较多的限制条件,导致研究结果往往缺乏普遍性。

上述研究可以看出,近些年针对层级距离的量化分析以及联合依存距离和层级距离共同表征句法处理难度的研究已成为当前研究热点,并取得了一定的研究成果。但不可否认的是,当前针对不同语言开展的层级距离相关研究还处于初始阶段,语料库来源相对单一,且研究对象也仅集中在有限的几种语言,现有研究结论是否具有跨语言的普遍适用性,还需要在其他语言类型中进一步验证,且朝鲜语目前尚未基于层级距离开展句法处理难度的相关研究。基于以上原因,本文将以朝鲜语为对象,采用真实语料和计算指标相结合的方法开展朝鲜语句法处理难度的研究。首先,针对朝鲜语的层级距离分布规律开展研究;其次,基于前期已开展的朝鲜语依存距离研究(华英楠、毕玉德 2022),进一步探究朝鲜语 MDD 和 MHD 的相互关系,并试图回答以下问题:

①朝鲜语作为支配词居后型语言,具有丰富的格标记和相对自由的线性顺序,其层级距离分布规律如何?以层级距离作为句法处理难度指标,朝鲜语的句法处理难度如何?

②朝鲜语句子的 MDD 和 MHD 的关系及句子结构在组织过程中的构造策略是什么?

二 语料资源与方法

(一) 语料来源

本文选择 UD^①以及同构的 SUD^②依存树库作为语料来源,UD 将实词标注为头部,更倾向于语义,而 SUD 将虚词标注为头部,更倾向于句法。已有研究表明,标注方式会对 MDD 产生影响,同 UD 相比,SUD 标注方式下的句子结构具有更低的 MDD(Gerdes et al. 2018; Yan & Liu 2019)。同理,标注方式理应也会对层级距离产生影响。标注方式对 MDD 和 MHD 的影响从以下例句分析中可清晰明了:

- (1) andŏŋtejmdakl̄ hante^haŋ juŋeŋil t̄ebut^hΛ kjAndzεk̄a teingi innum teip^③
 安东炖鸡 正当-ADV 流行-V 从开始-ADV 到现在-ADV 人气 有-V 店
 从安东炖鸡流行到现在一直很受欢迎的店

① <https://universaldependencies.org/>。

② <https://surfacesyntacticud.github.io/data/>。

③ 例句(1)选自 SUD-GSD 树库中的 train-s67。

基于 UD 和 SUD 标注方式的 MDD 和 MHD 计算结果^①分别如图 1、图 2 所示。

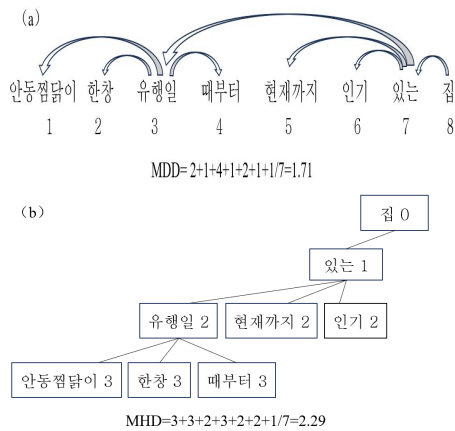


图 1 基于 UD 标注的 MDD 和 MHD

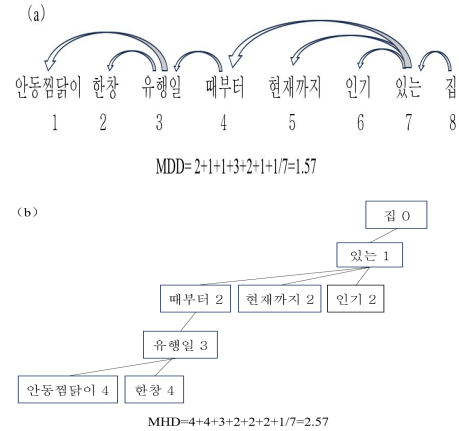


图 2 基于 SUD 标注的 MDD 和 MHD

如图 1 (b)、图 2 (b) 的依存句法树所示，最顶端为句子的根词，将其定义为 0，词语所在的层级位置便是层级距离。从根词开始，自上而下依次寻找支配词及其所支配的从属词，逐渐形成一棵完整的句法依存树。

仅从图 1 和图 2 中例句的计算结果可以看出，对于同一个句子，两种不同的标注方式对依存距离、层级距离、层级数以及部分子节点的分支方向（如例句 1 中的“때부터”）均产生影响。因此，本文将基于 UD 和 SUD 树库，进一步研究标注方式对层级距离和依存距离的影响程度以及二者之间相互关系。本研究的相关语料数据如表 1 所示：

表 1 相关语料数据

树库	标注方式	句子数	平均句长	体裁
UD-GSD	实词为头部	6339	12.67	新闻, 博客
SUD-GSD	虚词为头部	6339	12.67	新闻, 博客

(二) 层级距离计算方法

本文采用刘海涛、敬应奇 (2016) 提出的层级距离计算方法^②。具体的方法如下所示：层级距离表示子节点 (i) 在垂直方向上的投影点与根词投影点的距离 (d)。

(1) $HD=d_i$

平均层级距离为各子节点的层级距离之和除以子节点数量。一个由 N 个子节点组成的句子的平均层级距离可表示为：

(2) $MHD = \sum_{i=1}^N \frac{d_i}{(N-1)}$

^① MDD 计算方法见刘海涛 (2008)、MHD 计算公式见 (二) 层级距离计算方法。

^② 本文基于计算公式 (1) 和 (2)，利用 matlab 对层级距离开展相关的统计计算。

三 结果和讨论

(一) 朝鲜语句子的层级距离分布规律

在依存句法树图中，句子的中心节点称为根词，根词位于最高层级，句中所有其他子节点都直接或间接地受根词支配（刘海涛 2009:1-18）。从层级距离的分布规律可以看出不同语言类型的层级结构特征及其整体句法处理难度差异。从刘海涛(2017)和牛若晨(2021:32-143)基于汉语、英语和捷克语的研究结果来看，不同语言的层级距离分布规律既有共性也有差异，受限于所选的语言样本和语料标注差异，不同语言类型层级距离的分布规律还需要在更多语言中进行验证。本节我们将基于层级距离计算结果，研究朝鲜语层级距离分布规律及其背后的动因，期望得出不同语言之间层级距离分布规律的共性和差异。根据公式(2)对 UD-GSD 和 SUD-GSD 树库的层级距离进行计算，统计结果如表 2 和图 3 所示：

表 2 UD/SUD-GSD 树库层级距离分布

层级距离	SUD-GSD	UD-GSD
1	19157	19656
2	17858	20308
3	13526	14534
4	9662	9278
5	6081	5147
6	3671	2778
7	1967	1326
8	1066	566
9	512	250
10	245	96

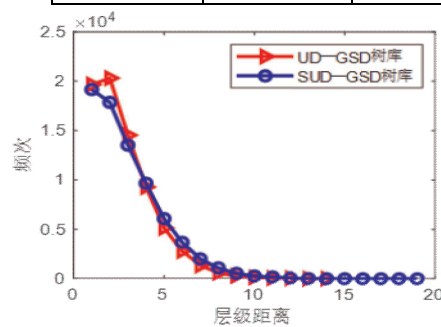


图 3 UD/SUD-GSD 树库层级距离分布图

从表 2 和图 3 可以看出，首先，UD 树库的层级距离分布呈现先升后降，在 HD=2 时达到顶点，随后逐渐降低，呈现长尾分布特点；SUD 树库的层级距离整体上单调递减，并呈现长尾分布特点。其次，UD 树库的最大层级数为 15，而 SUD 树库的最大层级数为 20，标注方式对层级数产生了影响。最后，从各层级距离的占比来看，不管是 UD 还是 SUD，当 HD

≤4^①时, 占比均达到了80%以上。基于计量语言学和心理语言学的大量研究已经证实, 受限于人类有限的记忆容量, 线性维度的依存距离倾向于最小化(Liu 2007, 2008, 2009, 2017; Liu et al. 2017; Futrell et al. 2015; 徐春山、梁君英 2015; 陆前、刘海涛 2016)。同其他语言一样, 朝鲜语的依存距离也已被证实存在最小化倾向(Futrell et al. 2015; 华英楠、毕玉德 2022; Hua et al. 2023)。本文基于层级距离的统计数据显示, 标注方式不同会对层级距离分布和层级数产生一定影响。但总体的统计结果表明, 受限于人类有限的记忆容量, 同线性维度依存距离有最小化倾向一样, 朝鲜语的层级距离也倾向于最小化。

层级距离本质代表的是句中各词语的层级位置(见图1(b)、图2(b)的依存句法树), 因此, 层级距离的占比等同于各层级词语数量的多少。在依存语法框架下, 由于词与词之间支配和被支配的非对称性关系是构成层级的基础。在依存树图中, 各层级词语的数量实质是上一层词语所支配的从属词的数量, 体现了上一层词语的配价能力。各语言层级距离分布差异来源于不同层级词语的配价差异。基于此, 我们以SUD树库为对象, 分别统计朝鲜语根词(HD=0)以及HD=1、2、3时排名前三的词类分布、依存关系类型分布及其占比(见表3)。

表3 HD≤3时排名前三的词类和依存关系类型分布及其占比

层级距离	词类分布	占比	依存关系类型分布	占比
HD=0	动词	74.5%	root	7.9%
	形容词	14.4%		
	名词	9.5%		
HD=1	名词	33.2%	mod	28.4%
	动词	23.9%	subj	18.5%
	副词	18.6%	comp:obj	10.4%
HD=2	名词	42.4%	mod	18.0%
	动词	21.0%	flat	14.5%
	副词	14.2%	comp:obj	11.7%
HD=3	名词	47.1%	mod	19.9%
	动词	17.6%	flat	17.6%
	副词	15.5%	comp:obj	11.8%

根据表3统计, 对于SUD树库, 根词所支配的从属词数量最多, 其次为第1层和第2层, 依次递减。说明根词的支配能力最强, 之后随着词语层级的加深, 支配能力也逐渐降低。从各层级距离的词类分布来看, 当HD=0时, 即位于树顶端的根词本身的词类主要为动词和形容词, 其中动词的占比高达74.5%。说明相比其他词类, 动词的配价最强, 作为句子的中心节点, 直接或间接地支配句子中的其他子节点。当HD=1时, 名词的占比超过动词, 成为最主要的词类。同时, 为了修饰根节点动词的需要, 副词的占比逐渐上升。当HD=2、3时, 占比排名前三的词类同HD=1时一样, 均为名词、动词和副词, 不同的是, 随着层级距离的增加, 名词的占比越来越多, 动词及其修饰词副词占比逐渐减少。

^① Cowan (2001) 的研究指出, 人类短时记忆的上限为4, 超出该界限往往会带来处理上的困难, 因此本文将HD≤4的层级距离定义为短层级距离。

依存语法框架下的依存关系是指词与词之间的句法关系，不同的二元依存关系构成了具有不同语义的句法结构。从表 3 中各层级距离的依存关系类型分布来看，当 HD=1 时，第 2 层级的词语作为根词的从属词，被根词直接支配，构成的依存关系类型属于全句的核心句法关系，占比前三的是 mod（修饰语）、subj（主谓）和 comp:obj（宾谓），三者的总占比达到了 57.3%。当 HD=2、3 时，占比前三的依存关系类型均为 mod、flat（扁平固定搭配）和宾谓。同 HD=1 时相比，随着层级距离的增加，subj 占比下降，mod 和 flat 占比逐渐上升。说明随着同根词距离的加大，表达句子中心语义的主谓结构逐渐减少，而修饰型句法结构的占比越来越多。当 HD=2、3 时，修饰型依存关系 mod 和 flat 的占比之和已达到 70%。

根据以上分析可知，影响层级距离分布的因素主要有两个：①是根词及其从属词的词类分布及其配价能力；②是根词及其从属词所构成的主要依存关系类型及其占比。UD 和 SUD 树库的词类分布相同，标注方式的不同造成两树库根词及其从属词构成的依存关系类型存在差异，从而使两树库的层级距离呈现出不同的分布规律。在依存句法树中，与语言类型无关，句子的根词一般由动词和形容词来充当，且同其他词类相比，动词的配价能力最强。因此，离根词越近，从属词的数量越多，这就决定了不同类型语言的层级距离分布存在共性。同时，对于不同类型的语言，由于其形态和句法等层面存在差异，就造成了主要依存关系类型存在较大差异，这也决定了其层级距离分布存在差异。两个因素相互作用，共同决定了不同语言类型的层级距离分布既有共性又存在差异。

（二）朝鲜语句子 MDD 和 MHD 的关系

MDD 和 MHD 已被证实可用作线性维度和层级维度上句法处理难度的衡量指标。具体而言，MDD 倾向于反映生成方面的处理难度，而 MHD 则倾向于反映理解方面的处理难度（Jing & Liu 2015）。依存距离之所以能够表征线性维度的处理难度，是因为在句子从左至右的增量生成过程中，句子中从属词被不断地存储于工作记忆中，只有当其支配词出现，才能构成完整的依存关系。因此，二者的线性距离越大，词被存储的时间越长，消耗的记忆也越大，处理也就越困难。层级距离可表征层级维度句法处理难度是因为层级距离的大小反映了从句子顶端的根词处出发激活一个词的难度。当一个词与根词的距离越远时，层级深度越深，预测或激活就越难。生成和理解是句法结构的一体两面，二者互相影响，缺一不可。开展依存距离和层级距离的关系研究，可以更好地理解朝鲜语句子在线性和层级维度的句法复杂度以及在句法组织结构上的异同点。

为了探究朝鲜语整体句法处理难度以及句法构造过程中的组织策略，本节我们将开展朝鲜语句子的 MDD 和 MHD 相互关系研究。基于 UD/SUD-GSD 树库的朝鲜语 MDD 和 MHD 统计结果如表 4 所示：

表 4 UD/SUD-GSD 树库整体 MDD 和 MHD 值

树库	MDD	MHD	MDD 和 MHD 联合均值
UD-GSD	3.76	2.72	3.24
SUD-GSD	2.72	2.96	2.84

据表 4 可知，①两树库的 MDD 和 MHD 存在明显差异。UD 树库的 MDD 值为 3.76，明显大于 SUD 树库的 MDD 值 2.72；UD 树库的 MHD 值为 2.72，略小于 SUD 树库的 MHD 值

2.96。②如果以 MDD 和 MHD 的均值衡量树库的整体处理难度, UD 树库的均值为 3.24, SUD 树库的均值为 2.84。③UD 和 SUD 树库的 MDD 值、MHD 值及两者的联合均值均未超过 4, 遂验证了 Liu (2008) 对 MDD 阈值和 Jing & Liu (2015) 对 MHD 阈值的研究结论。

综上所述, 结论①和②表明, 标注方式会影响 MDD 和 MHD 的整体取值, 且将实词标注为头部的 UD 树库其句法处理难度要大于将虚词标注为头部的 SUD 树库。结论③说明, 虽然标注不同, 但两树库的 MDD 和 MHD 的联合均值均未超过 4。这是因为受人类普遍认知机制的约束, 在句子的生成和理解过程中, 人们会尽量将处理难度控制在一定范围内。

为了进一步验证以上研究结论, 我们将分别对句长为 5、10、15、20、25、30、35、40 时的 MDD 和 MHD 进行统计, 其结果如图 4 所示:

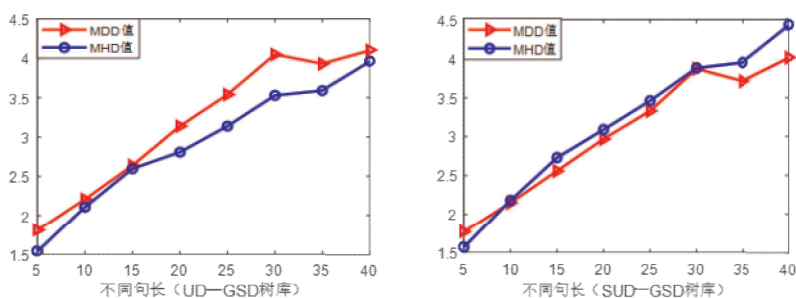
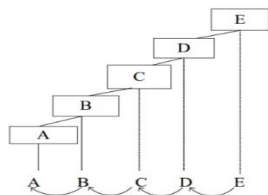


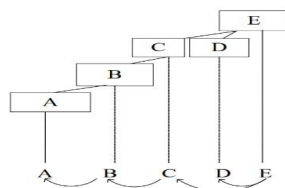
图 4 UD/SUD-GSD 树库不同句长下的朝鲜语 MDD 和 MHD 比较

从图 4 可知, 首先, 随着句长的增长, 两个树库的 MDD 和 MHD 值均呈上升趋势。其次, UD 树库的 MDD 值大于 MHD, SUD 树库的 MHD 值大于 MDD, 而两树库整体的 MDD 和 MHD 联合均值均未超过 4 (详见表 4), 说明与句长无关, 在两个树库中, MDD 和 MHD 值此消彼长, 始终保持一种权衡状态。

MDD 和 MHD 的相互关系可以从图 5 模型中得出。假设句长为 5, 句中各词语分别为 A、B、C、D、E, 其中 E 为句子的根词 (符合朝鲜语 SOV 的语言类型), 上方为句子的二维句法树, 下方为二维句法树在一维线性维度的投影, 弧线表示支配词与从属词之间的句法依存关系, 箭头方向从支配词指向从属词。随着根词所支配的从属词的增加, MDD 和 MHD 的动态变化如下图 (1) - (4) 所示^①。

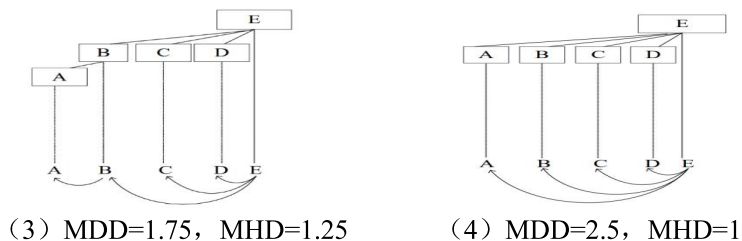


(1) MDD=1, MHD=2.5



(2) MDD=1.25, MHD=1.75

^① 该模型仅考虑了根词从属词以及从属词全部位于根词同一侧的情况。随着句长的增加, 根词从属词所支配的词语也会逐渐增加, 树的形状以及 MDD 和 MHD 也会随之发生变化。在该模型中, 当根词及其从属词各有两个从属词时, MDD 等于 MHD, 均为 1.5。



MDD 和 MHD 的数值变化可以从图 5 直观地看出：

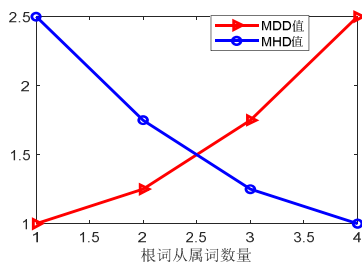


图 5 随根词从属词数量变化的 MDD 和 MHD

从依存句法树的形状来看，随着根词所支配的从属词数量的增加，依存树变得越来越宽，此时，由其投影的线性句子中的长距离依存关系也越来越多，因此，导致 MDD 越来越大。相反，随着根词从属词数量的增加，层级数越来越少，树的深度越来越浅，各词语离根节点的垂直距离降低，从而导致 MHD 越来越小。图 5 直观地反映了 MDD 和 MHD 的动态变化过程，从图中可以看出，在 MHD 最大时，MDD 倾向最小，反之亦然。如果以 MDD 和 MHD 的均值作为衡量句子整体处理难度的数据，上述 (1) - (4) 的平均值均保持在 1.5-1.75 之间。说明在句法结构的组织过程中，线性维度和层级维度的处理难度相互制约，以达到降低整体句法处理难度的目标，从而减少对人类有限认知记忆的消耗。本文基于朝鲜语的研究结果再次为复杂性权衡假设提供了实证证据 (Yan & Liu 2021; Koplein et al. 2017)。

四 结 语

本文基于 UD 和 SUD 依存树库对朝鲜语层级距离分布规律以及朝鲜语 MDD 和 MHD 的相互关系开展了研究分析。研究表明：①虽然不同的标注方式会对层级距离单调性产生一定影响，但 UD、SUD 两树库层级距离整体上都符合长尾分布规律，MHD 倾向于最小化。②影响层级距离分布差异的本质原因：一是根词及其从属词的词类分布及其配价能力；二是根词及其从属词所构成的主要依存关系类型。③本文的统计结果显示，整体上 UD 树库的 MDD 值更大，而 SUD 树库的 MHD 值更大。虽然标注方式会对 UD、SUD 树库的 MDD 和 MHD 值产生影响，但两树库的整体 MDD 和 MHD 均值均未超过 4 (详见表 4)，说明为了减少对人类有限认知记忆的消耗，朝鲜语的句法处理难度整体处于一种动态权衡状态。④通过对 MDD 和 MHD 之间的相互竞争关系进行动态模拟可以得知，根词从属词的数量是造成 MDD 和 MHD 此消彼长的主要原因。

本文基于依存语法框架下的层级距离对朝鲜语的层级结构开展计量研究,从宏观上了解朝鲜语句法结构特点及其组织策略,验证并丰富了现有的理论研究成果,为后续基于依存语法开展更多样化和更深层次的朝鲜语句法研究提供了案例。需要指出的是,本文仅基于不同标注方式的两种树库针对朝鲜语单一语言开展了层级结构计量研究,没有针对具体的句法结构开展更为详细的分析。后续我们将基于大规模平行语料库开展跨语言类型的层级距离研究,并针对层级距离的不同影响因素进行分析,探索不同语言类型在层级距离上的共性和差异。

参考文献

- 程 工. 2018. 《层级结构和线性顺序之新探》,《外语教学》第 1 期.
- 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》第 6 期.
- 刘海涛. 2008. 《基于依存树库的汉语句法计量研究》,《长江学术》第 3 期.
- 刘海涛. 2009. 《依存语法的理论与实践》,北京: 科学出版社.
- 刘海涛. 2017. 《句子结构层级的分布规律》,《外语教学与研究》第 3 期.
- 刘海涛. 2022. 《依存关系与语言网络》,北京: 科学出版社.
- 刘海涛、敬应奇. 2016. 《英语句子层级结构计量分析》,《外国语》第 6 期.
- 陆 前、刘海涛. 2016. 《依存距离分布有规律吗?》,《浙江大学学报》第 4 期.
- 牛若晨. 2021. 《线性及层级句法特征的计算认知研究》,浙江大学博士学位论文.
- 徐春山、梁君英. 2015. 《依存距离最小化是人类语言的普遍规律吗?》,《浙江大学学报》第 6 期.
- 백준오, 남윤주 & 김영주(白俊伍、南润珠、金英珠). 2019. 한국어 수식절 처리의 상대적 난이도: 시선 추적 연구(韩国语修饰从句处理的相对难度研究), *국어국문학* 제 189 호.
- Chen, Ruina, Sirui Deng & Haitao Liu. 2021. Syntactic complexity of different text types: From the perspective of dependency distance both linearly and hierarchically. *Journal of Quantitative Linguistics* 29(4): 510-540.
- Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24(1): 87-185.
- Futrell, R., K. Mahowald & E. Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33): 10336-10341.
- Gerdes, K., B. Guillaume, S. Kahane & G. Perrier. 2018. SUD or surface syntactic universal dependencies: An annotation scheme near-isomorphic to UD. *Universal Dependencies Workshop 2018*, pp. 1-9, Brussels: Association for Computational Linguistics.
- Hua, Yingnan, Jiapeng Li & Yude Bi. 2023. Research on Korean “long-before-short” preference from the perspective of dependency distance. In Qi Su, Ge Xu, Xiaoyan Yang (eds), *Chinese Lexical Semantics*, pp. 132-142, Fuzhou: Springer.
- Hudson R. A. 2007. *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
- Jing, Yingqi & Haitao Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. *Proceedings of the Third International Conference on Dependency Linguistics*, pp. 161-170, Uppsala: Uppsala University.
- Koplenig, A., P. Meyer, S. Wolfer & C. Müller-Spitzer. 2017. The statistical trade-off between word order and word structure-Large-scale evidence for the principle of least effort. *PLoS one* 12(3): 1-25.
- Kwon, N., R. Kluender, M. Kutas & M. Polinsky. 2013. Subject/object processing asymmetries in Korean relative clauses: Evidence from ERP data. *Language* 89(3): 537-585.

- Lee, Y., H. Lee & P. C. Gordon. 2007. Linguistic complexity and information structure in Korean: Evidence from eye-tracking during reading. *Cognition* 104(3): 495-534.
- Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics* 25: 1-12.
- Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2): 159-191.
- Liu, Haitao. 2009. Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics* 16(3): 256-273.
- Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews* 21: 171-193.
- Yan, Jianwei & Haitao Liu. 2019. Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand? In Xinying Chen, Ramon Ferrer-i-Cancho (eds.), *Proceedings of the First Workshop on Quantitative Syntax*, pp. 16-24, Paris: Association for Computational Linguistics.
- Yan, Jianwei & Haitao Liu. 2021. Morphology and word order in Slavic languages: Insights from annotated corpora. *Вопросы языкознания* 4: 131-159.
- Yngve V. H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104(5): 444-466.

Syntactic Processing Difficulty in Korean: A Study Based on Hierarchical Distance

HUA Yingnan and BI Yude

[Abstract] Dependency distance (DD) and hierarchical distance (HD) within the dependency grammar framework have been established as measures of syntactic processing difficulty in hierarchical dimension. We carried out a study on the distribution rules of hierarchical distance in Korean as well as the relationship between mean dependency distance (MDD) and mean hierarchical distance (MHD) based on UD and SUD dependency treebanks. Results show that different annotation methods will affect the value and distribution of HD in Korean, but HD exhibits a minimization tendency; The distribution of word classes and valency of root and their dependents, as well as the types of dependency relationships, is the essential cause for the distribution patterns of HD; Compared to the SUD, the syntactic processing difficulty of the UD is greater, but in order to minimize the consumption of limited human cognitive memory, the MDD and MHD of the two treebanks are mutually constrained, with neither of the averages exceeding 4, thus indicating a dynamic trade-off in sentence processing difficulty.

[Keywords] Korean dependency treebank hierarchical distance dependency distance syntactic processing difficulty

(通信地址: 200433 上海 复旦大学外文学院)

【本文责编 毛 选】