基于依存句法标注树库的汉俄语对比研究

——以 2021 年俄罗斯国情咨文为例

张　嵘

（辽宁大学，沈阳 110036）

提　要：以计量语言学为研究方法，以依存语法为理论框架，分别对 2021 年俄罗斯国情咨文俄语版、汉语版进行 MDD、NDD 的计算，并统计依存关系正负值占比，以及相邻词间的依存关系占比。得出结论，一是数据验证出 MDD 与 NDD 变化趋势并非完全相同，二是用数字更加明确了俄语和汉语的语言类型，三是指出了俄语词序是相对自由的，受到位置范围的限定，并发现总结出新的问题：句法关系并非是孤立存在的，其配阶的实现受到几个方面的影响，其中包括：语义，语法，语言实践中的配阶使用价值，这几个层面相互作用影响。

关键词：依存语法；依存距离；国情咨文；MDD；NDD

中图分类号：H08　　**文献标识码：A**

一、引言

在字典当中，词与词都是孤立存在的，没有任何关联，但当它们一旦有组织地合成一个句子，它们之间就产生一种关系，这些关系又层层叠叠地构成了句子的框架。绝大多数的词在自己身边都有一些“空位”，这些空位在词孤立状态下是不可见的，当收到大脑指令时，空位被填充，词与词按照一定的规则进行结合，或者说基于语言经验进行“配阶”从而形成更大的语言单位。（刘海涛 2009：72）这种词与词的结合力、配阶的过程就是我们这里所说的依存关系。依存关系就是依存语法研究的对象，具体地说，依存语法研究的是词之间的从属关系，它在句法层面上构建出了句中词之间支配与被支配的地位关系，是建立在二元词间关系基础之上的语言理论。正如刘海涛所述：依存语法认为，句子是有组织的单位，其基本组成元素是词，词与词之间相互联系。（刘海涛 2009：20）

这种联系是抽象存在的，为了更直观的研究它，可以通过显示的图示方式将句中各成分之间的关系和功能进行标注形成句法图。一般认为，最早的句子结构图式出现在 19 世纪中叶的美国。（刘海涛 2009：4）到了 20 世纪，现代依存语法创始人 Tesnière 首次提出用句法树来分析现代依存句法，之后在美国语言学家 Hays 的依存树图中不仅仅强调了句子的结构词序，而且标注了线性次序，美国的语言学家 Van Valin 进一步使用了有向弧来体现句法关系中的从属性，在前辈的基础上刘海涛使用类似“World Grammer”的一种复杂特征结构依存树的有向图对依存关系进行区分。它能清楚直观地显示出一个节点的多个从属者，有向地反应出句子的线性次序、层次次序以及其它必要的句法信息。（刘海涛 2009：19）哈工大社会计算与信息检索研究中心的语言技术平台 LTP 也采用该标注方法（本文汉语语料标注使用该平台代码），具体见图 1：

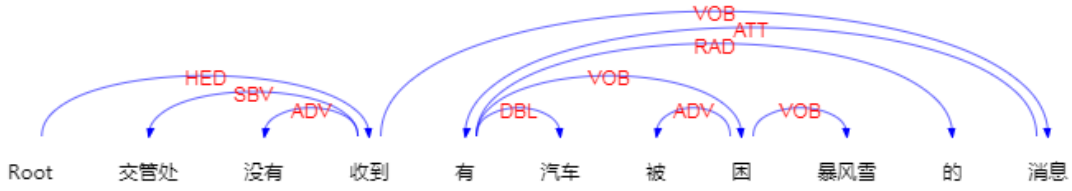

图 1. LTP 语言技术平台生成依存树图

注：如上依存树图由 LTP 语言技术平台生成，平台将输入的线性词串转变为复杂特征结构的有向图。有向弧由支配词出发指向从属词，弧上方为核心关系、主谓关系、动宾关系等依存句法关系标签。+

二、研究现状与理论简介

刘海涛提出依存关系的计算方法，$\overline{DD}=\frac{1}{n-s}\sum_{i=1}^{n-s}|DD_i|$，（刘海涛 2007，2008：163）用数学直观地计算出词与词的依存距离均值 MDD，证明了人类语言依存距离最小化的倾向。在该理论基础上，雷蕾提出了依存距离同时也受到 Root 动词位置和句长的影响，提出了标准化依存距离 NDD 的计算公式，$NDD=abs\left(\ln\left(\frac{MDD}{\sqrt{rootdia\tan ce\times sentencelength}}\right)\right)$。（雷蕾，Jockers 2018：6）刘丙丽对真实语料进行句法成分分析，发现汉语不同语体中充当相同句法成分的各词类所占比重有较大差异。（刘丙丽等 2012：134）王巧林对两任美国总统的就职演说进行量化对比，发现语言的计量分析有助于对文本复杂度的判断。（王巧林，李雯雯 2019：58）多语言对比，大数据计算机技术的应用，为人类语言的认知探索开启了新的研究之路。俄罗斯 A. И. Мельчук 与 A. К. Жолковский 提出意义 ⇔ 文本语言学模型，据此理论由俄罗斯科学院信息传送问题研究所计算语言学实验室开发出 ETAP-3 机器翻译系统，并于 2009 年完成了俄语句法标注树库（STR- СинTar Рус）的建设，（王永，刘海涛 2013：176）该树库采用依存树的形式对句子中的依存语法进行了标注，并显示出词的形态和句法功能等信息。本文便采用 STR 的句法划分标准，部分例句来源于此语料库。

依存距离指的是支配词和从属词之间的线性距离，即一个句子中存在依存关系的两个词之间的词位置差。（刘海涛 2009：252）基于刘海涛提出的依存距离计量方法，将具有依存关系的两个相邻单词之间依存距离基准值定为 1，并为专注于词与词之间的关系而取消句中句末标点，避免其影响而使依存距离增大，对 Информация о застрявших на перевалах машинах в диспетчерскую не поступала.（交管处没有收到有汽车被困暴风雪的消息）进行标注得出以下数据：

表 1. 支配词与从属词线性距离数据表

Dependents	Dependent Pos	Position of the dependent	Governors	Position of Governors	Dependency relations	DD
поступала	глаг	10	ROOT	0	root	/
Информация	сущ	1	поступала	10	nsubj	9
о	пред	2	информация	1	prep	1
застрявших	прилаг	3	машинах	6	amod	3
на	пред	4	застрявших	3	prep	1
перевалах	сущ	5	на	4	prep	1
машинах	сущ	6	о	2	prep	4

в	пред	7	поступала	10	prep	3
диспетчерскую	сущ	8	в	7	prep	1
не		9	поступала	10	det	1
MDD						2.66
NDD						1.27

这句话的根是动词 поступать，它与主语名词 информация 是名词性主语谓语关系，依存距离为 10-1=9，依次向下进行运算 $MDD=\frac{9+1+3+1+1+4+3+1+1}{9}=2.66$。

$NDD=abs\left(\ln\left(\frac{2.66}{\sqrt{10*9}}\right)\right)=1.27$。这句话的汉语表述 MDD 为 3，NDD 为 0.55。

有趣的是，俄语词序比较自由，打乱这三个部分：1）Информация о застрявших на перевалах машинах，2）в диспетчерскую，3）не поступала 并不影响对句意的理解，得到 123，132，312，321，231，213 六种组合进行计算。如下图对比所示：

表 2. 词序自由组合后支配词与从属词线性距离数据表

序列	组合	分词数	动词位置	MDD	NDD
1	Информация о застрявших на перевалах машинах в диспетчерскую не **поступала**. (123)	10	10	2.66	1.27
2	Информация о застрявших на перевалах машинах не **поступала** в диспетчерскую. (132)	10	8	2.33	1.29
3	Не **поступала** информация о застрявших на перевалах машинах в диспетчерскую. (312)	10	2	2.22	0.65
4	Не **поступала** в диспетчерскую информация о застрявших на перевалах машинах. (321)	10	2	1.77	0.87
5	В диспетчерскую не **поступала** информация о застрявших на перевалах машинах. (231)	10	4	1.77	1.22
6	В диспетчерскую информация о застрявших на перевалах машинах не **поступала**. (213)	10	10	3.11	1.12
7	交管处没有收到有汽车被困暴风雪的消息	10	3	3	0.55

从图表数据的对比可以看出，句中词序进行调整，这里的改变主要是动词位置移动，词组的搭配仍然保留，其依存距离保持不变，但不可避免的是与谓语动词有依存关系的词与动词依存距离值会发生改变。当谓语动词移动时，MDD 与 NDD 的值明显随之改变，MDD 也会受到动词移动的影响，但 NDD 的曲线变化走向更加贴合动词的移动轨迹，NDD 受到动词位置的影响更大。将 NDD 用于分析类似俄语这种词序灵活的语言时，其变化幅度更加的明显。在句 4、句 5 中出现 MDD 同值，均为 1.77，但语序并不相同。当计算公式中引入的因素相对较少时，出现相同数值的比例大大高于考虑因素多的计算所得，所以我们有必要计算 NDD，扩大对依存关系有影响的考量因素的分析范围。依存距离值大的 123 与 213 形式，谓语动词处在句末，是认知消耗最大的方式，但 123 恰恰出现在报纸上，而非其他依存距离小的句子。笔者推测书面语用眼获取信息，与口语用听觉获取信息的方式不同，认知消耗也是不同的。用眼可以快速略过次要信息，或者再回看次要信息，所以依存距离大小对于书面语的理解的难易程度影响不大。而口语则偏向使用依存距离小的句序排列方式。 这里应引起注意的是为何俄语 MDD 比汉语小，NDD 反而比汉语大呢？这里受影响最大的就是动词位置。俄语表述中动词在句末，动词位置为 10，而汉语表述中动词在 3，其 NDD 就大大的

小于俄语 NDD。也即俄语话语接近结束才出现句子的中心动词：消息没有收到。收到主干信息后，大脑继续接收一系列的限定修饰，关于什么的消息？什么样的汽车？到句末重新整合得到完整信息。该俄语句的认知消耗大于汉语，因此可以认为此句 NDD 更加贴合地体现句子理解的难度。虽然依存距离可以作为一种对句子理解难度的计量指标，MDD 距离值越大表明支配词和从属词之间的距离就越大，也就越难理解，大脑对句子的处理时间越长。但是仅仅凭一次计算无法下结论说汉语比俄语复杂。只有语料库越大，语料越整齐，得到的数据才能越可靠。对单一句子分析后得出的数据说服力不足，并考虑到国情咨文在传媒语言研究中具有的严谨性和代表性，所以我们选取 2021 年国情咨文作为语料。

数据涉及到的参数包括：第一，MDD 与 NDD，考虑到俄罗斯动词位置的灵活性，以及俄语的句子普遍较长，而 NDD 更加明显地体现动词与句长的参数变化对依存距离的影响，所以统计时将 MDD 与 NDD 同时计算。第二，因为支配词可在从属词之前，也可置后，这使得依存关系带有方向性，所以依存距离是有正负的。比如支配词 информация 与从属前置词 о 之间依存距离是 2-1=-1 负值，这是一个支配词居前的依存关系。而 В диспетчерскую поступала 中支配词 поступала 与从属词 в 之间的距离是 3-1=2，正值表明支配词置后。通过依存距离这种测量方法可以在真实的语料中推断出这是一种偏向什么类型的语言，是支配词置后还是居前，亦或是混合型；有助于习语者避开母语语序的影响，从而入乡随俗地组装出正确的习惯语序。因此将说明支配词位置在前还是在后的正负依存距离值占比纳入统计考量。第三，研究方向是在依存距离的基础上进一步统计，以求得俄语语序相对自由的程度，依存距离大于 1 的绝对值占比多少？频率最高的依存距离是多少？用数字来求证俄语词序的习惯排列。第四，同时将平均句长作为影响 MDD 值的主要因素列入统计范围内。本文从以上几个方面对 2021 年俄罗斯总统普京的国情咨文以及汉语版翻译进行量化的对比。

三、统计步骤与标准

本文采取刘海涛提出的 MDD 计算方法与 2018 年雷蕾提出的 NDD，对 2021 年的俄语版国情咨文及其汉语译文进行统计，基于 Lancsbox 的统计分词后得出：俄文样本包括 7905 词、515 句，中文在分词计算后为 7803 个词，16508 字，502 句。中文标注采用现代汉语词类体系以及现代汉语依存关系层级图，标注词类与依存关系。

1. 首先需要对两种语料进行标准化处理，变为 TXT 纯文本之后进行分词。

Lancsbox 的统计俄文样本包括 7905 词，经人工检查后发现实际为 7694 个词。原因在于：第一，句中做谓语的动词是依存关系的根本，当谓语省略动词存在时，依存关系无法划分，所以句中人为地加入该动词，它被称为幻影动词（фант）。如国情咨文写道：Искренние слова признательности и учителям школ, преподавателям вузов, других образовательных учреждений. 该句在句法关系划分时需在 искренние слова 前加入 выразить。但幻影动词仅仅在依存关系划分时存在，并不计入词数统计中。同时，现在时态的主动句根据动词变位后的词尾可以推断出主语，有时出于各种原因将主语省略，这种情况在咨文中较常见，但划分句法关系时并不需要加入幻影主语，因为主语并非是关系划分的出发点，此时，仅将主谓关系去除。第二，与汉语不同，俄语存在复合前置词，如 по отношению к，计算机会自动计算为三个词，但使用中无需进行词法变化，它们是固定存在的，而且无法将其分开, 所以人工标注为一个词。并非所有的复合前置词或词组都被计算为 1 个词，当构成词可以拆分或是可以进行词法变化时被认为附加补充关系（支配词为 X，从属词为 Y），如 потому[X] что[Y], и так[Y] далее [X] 等等。第三，带有连字符的复合词并非一个单词，如 научно-технологический（科技的）计为两个词，句法关系是复合关系（композ），以及咨文中出现的类似 бизнес-инкубатор, учебно-лаброторный, транспортно-логистический 等等。

2. 在分词无疑且明确定义后，使用安装俄语语言包的 python 软件对俄语文本进行划分

计算，用 LTP 平台代码导入 python 对汉译文本语料进行计算，最终生成运行结果，之后人工校对（本语料逐句校对效果不佳），同时说明依存关系标注的参考标准。

因为研究对象包括俄汉两种语言，其语言体系不同，依存关系的划分也不尽相同，再者其他学者的研究也普遍使用英语缩写词（一般以斯坦福大学的 stanfordnlp 为准）。俄语最大的语料库 STR 将依存关系分为四大类，并向下细分出 67 种依存关系。按照出现的频次依次为：定语关系、第一补足语关系、前置词短语关系、述谓关系（王永，刘海涛 2013：178）。所以结合三种语言，以对应的方式列出如下主要关系（见表 3）。需提及的是，俄语说明复句的依存关系划分时，连接词引导的从句，STR 将连接词划为从属词（如 Мне нравится [X], что [Y] вы больны не мной..），关联词引导的从句，谓语动词作为与其发生句法阶 1 的直接关系（Любопытно [X], куда он пошел [Y].）。原因在于，从属词的准确断定是得出 MDD 与 NDD 数据准确的根本，如关系划分有异议，那么数据统计得出的结论也不可信，统计应在统一的划分框架下进行。

表 3. 国情咨文例句标注

例句	俄语名称	俄语缩写	汉语名称	汉语缩写
Граждане[Y] действовали[X].	主谓	Предик	主谓	SBV
Распространить[X] эту меру[Y1] на[Y2] будущих первоклашек.	第一至第四补足语	Комплетивные СинтО (1-компл,2-компл）	动宾/宾语前置/兼语/双宾语/依存分句	VOB/FOB /DBL/DOB/DC
Киплинг великий писатель[Y] был [X]. в[X]последнее время[Y]	表语 前置词	Присвязочное Предложное СинтО	介宾	POB
если[X] кто-то воспринимает[Y] наши добрые намерения	关联词	подч-союзн	关联词	CNJ
в России должен быть меньше[X], чем[Y] в Евросоюзе	比较	сравнит	比拟	SIM
Это сложная[Y]задача[X].	限定	Опред	定中/状中/动补	ATT/ADV/CMP
Мы обязательно поможем людям[X], которые здесь живут[Y]...	定语从句	релят		
повышению[X] качества[Y] жизни наших людей	非一致定语	Общеатрибутивные		
принцип[X] «загрязнитель платит[Y]» должен	同位语	Аппозитивные СинтО	连谓/同位	VV/APP
45[Y] тысяч[X] бюджетных мест	数量	Количественные СинтО	数量	QUN
Здесь нужно действовать[X] жёстко[Y].	状语	обстоятельственное СинтО	时间关系/处所关系	TMP/LOC
«Справедливая Россия»[X], ЛДПР, КПРФ и[Y] «Единая Россия» – их поддержат.	并列	Сочинительное СинтО	并列	COO
Рынок труда и доходы граждан обязательно будут восстановлены[X], и[Y] мы дальше пойдём.	并列从句	сент-соч	独立分句	IC
Будем[X] настраивать[Y] всю систему.	复合谓语	аналит	语态结构	MT
она будет[X] использована[Y].	被动语态	пасс-анал		
Если[X], то [Y].	关联结构	Соотнос	关联结构	CS

Сбережение[Y] народа России – наш высший национальный приоритет[X].	复指	Пролепт
друг[Y] на друга[X]	补足	вспом

3. 实际的句法划分人工校对时发现 python 智能分析可能会发生计算偏差，需人工更正。

原因 1 位置范围

俄语语句，即使其语序较为自由，但每一个词的存在位置并非随意，而是有固定的范围（диапазон позиций），位置范围的划分首先受说话人所要表达句意的影响，词的位置如果超出该范围就会改变语义，随之带动句法结构的改变。咨文中有这样一句话：

例句[1]：Нужно начинать это делать эффективно и быстро.（要开始迅速有效地做这件事情）。

例句[2]：Нужно эффективно и быстро начинать это делать.（要迅速有效地开始做这件事情）。

эффективно и быстро 两个副词是 делать 的从属词，因为它们邻近 делать。因位置改变，由第一句中形容“做”到第二句形容“开始”，位置变化改变了句意，依存关系和依存距离随之改变。эффективно и быстро 的位置范围在 делать 之后时，划分依存关系时将副词划入 делать 的从属关系。第二句移动其位置改为 начинать 之前时，可认为 быстро 修饰限定 начинать，而并非是 делать 的状语（эффективно начинать 搭配不是很恰当，这里先忽略语义搭配问题）。所以依存关系的划分首先要考虑其位置范围，进而决定从属关系。

原因 2 语义因素

初步界定位置范围，断定从属关系后，还要考虑到语义因素的影响。STR 树库中将“依存关系”称为“句法关系”，但其实质一致。依存关系的研究属于句法范畴，需进一步明确依存距离的定义。刘海涛将其定义为支配词和从属词之间的线性距离，要进一步将其限定为支配词[X]与在句法阶上为 1 的从属词[Y]之间的线性距离，而非语义阶为 1 的从属词。但依存关系既是一种语法功能关系，又是语义角色关系。依存语法不可避免地受到语义层的影响，甚至需要检查依存关系的划分是否涵盖了句中单词所要表达的全部语义，并以此来判断依存关系划分的准确性，由此可见依存关系并非孤立存在。例如咨文结尾处中有这样一句话：

例句[3]：За всей текущей работой мы, безусловно, не должны забыть целей нашего стратегические развития, национальные цели развития и совершенствовать механизмы в достижении этих целей.（对于目前的所有工作，我们当然不应该忘记我们的战略发展目标，国家发展目标，也不能忘记完善实现这些目标的机制）。

句中 совершенствовать（完善）完全可以与 забыть 并列做 не должны 的从属词，译为“我们不应该去完善实现这些目标的机制”，但语义很明显与上下文不符，因此应该将 совершенствовать 视为 забыть 的补足语关系。依存语法的研究是以配阶理论为基础的，研究词与词结合的能力，语义关系的研究也是研究词间关系，他们之间存在着密切的关系，甚至依存关系很容易转换为语义关系。

原因 3 使用价值

послать 的常用配阶从属词是 на，отзыв 亦是 на，因依存语法的划分有个重要原则：一个节点不能同时从属于两个或以上的支配词，虽然可以有多个从属者，但只能有一个支配者，据此需明确 на 的从属关系只能有一个。这里除了对位置范围、语义、语法的考量外，还应考虑到日常表述中该搭配出现的频率，писать отзыв на книгу, отзыв на книгу, на статью 等表述中，на 是 отзыв 的固定前置词搭配，而 послать 可以是 куда, на, в, 所以在切分时将 на 处理为 отзыв 的从属词。例如：

例句[4]：Такой подход должен стать нормой в работе всех уровней власти（这种做法应

该成为各级政府工作的常态）.

例句[5]：Такой подход должен стать нормой в работе власти всех уровней（这种做法应该成为各级政府工作的常态）.

句 5 为改写句。句 4 中，根据 власти 的位置范围、语义以及语法可以断定 власти 修饰 уровень，уровень 修饰 работа，句 5 的改写也完全符合表达习惯，也可以认为 власти 修饰 работа，всех уровней 限定 власти。此时则需要人工干预校对。我们以“使用价值”为原则，取应用频率最高的表达方式进行划分，句 5 的语序更常见，所以在句 4 的划分时认为 власти 修饰 работа，всех уровней 限定 власти，没有受到 власти 位置的影响。

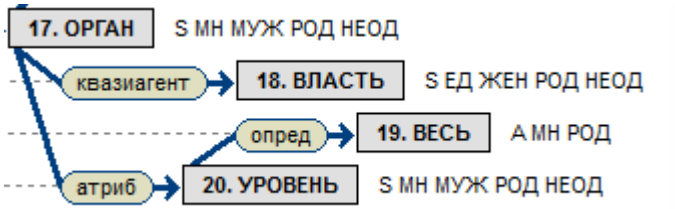

图 2. 依存语法划分人工干预校对

原因 4 语法限制

同时，依存关系的不孤立性还体现在它与语法关系的紧密联系上。在处理俄语中特殊的语法现象，比如形容词名词化时，也需要人工校对。如咨文中 Чтобы уже с этого года больничный по уходу за ребёнком в возрасте до 7 лет включительно оплачивался в размере 100 процентов от заработка.（从今年开始，7 岁以下儿童的医疗账单按父母收入的 100%返还。）如果按照配阶模式来分析，больничный 是形容词修饰名词，那么就将其划分为 год 的定语，但是这样不符合语法规则，больничный 应该按照一致原则与 год 同为二格，而且句子缺失了主语。这时，简单地依靠词类配阶模式进行分析是不可靠的，所以综合语法、语义，以及表达习惯，得出 больничный 是形容词名词化，译为医疗账单，作为句子中的主语。

四、统计数据说明

表 4. 1DD%为相邻词间的依存关系占比，即 MDD=1/-1 在全文的比例

语种	总句数（句）	总词数（个）	平均句长	依存关系数量	支配词置后占比（正值）	支配词居前占比（负值）	MDD	NDD	1DD %
汉语 2021 年	500	7803	15.5	8449	58%	42%	2.86	1.16	44.2%
俄语 2021 年	515	7694	14.9	7179	55%	45%	2.1	1.41	54.4%
汉语 2020 年	455	8193	18.0	7738	61%	39%	2.8	1.17	45%

数据解读如下：

1）数据统计后分析发现，国情咨文俄汉版本的总句数和总词数差别不大，因两者存在翻译关系，所以总句数基本一致。

2）平均句长的词数量对比显示，汉语并不短于俄语句子，2020 年汉译咨文句中含词量甚至达到 18。但在笔者所做的二语习得者问卷中，90%的被问卷人认为俄语句子明显比汉语长，5%的认为俄语句子稍稍长一点，5%没有注意。或许因俄语单词较长，而且同样一句话中，所用的标点少于汉语版本。俄语说话习惯也是一个长句子一气呵成，这就造成俄语习得觉得俄语句子长的错觉。

3）依存关系的数量汉语文本为 8449，俄语文本为 7179，汉语远远高于俄语。推断原因在于汉语作为意合语言缺乏与俄语类似的语法形态变化，如词的性、数、格等等，语法又恰恰可以限定并牢固从属词与支配词之间的关系，正是通过语法规则，我们可以断定从属词和哪个支配词相连接。语法变化的缺失使得汉语句法分析约束特征减少，直接导致句法分析结果多于俄语。

4）汉语是一种支配词置后略占优势的混合型语言，俄语同样支配词置后占比略占优。此类数据的计算有助于从统计学上认识俄汉语语序结构特点，换言之，当不知如何安排句法结构时，按照统计数据所得，支配词后置于从属词的情况居多。汉语中负值关系包括：右附加、动宾、动补、介宾、并列。俄语中负值包括：主谓、动宾动补、前置词、并列、并列从句、定语从句、比较关系、关联结构、复合谓语、补足关系、插入语、同位语。由上可看出俄语版本中支配词在前的比例远高于汉语，多在从句的几种关系和主谓关系上。俄语连接主句与从句的关联词分为两个部分，主句中的关联词为支配词，从句中为从属词，类似“因为，所以”“如果，那么”，汉语中“因为，所以”并没有划分单独的依存关系，而是认为句中动词产生了“关联结构（CS）”的关系。在主谓关系上，俄语的主语位置相对灵活，可置于动词之后，但汉语不允许，会因此改变句意。有趣的是，汉语宾语的位置相比较主语而言，比较自由，可以在“被”字句或者在可以省略“被”的被字句中前置，例如在咨文汉译本中：目前俄罗斯在科学技术方面研发的潜力也在这场抗击疫情的战役中激发出来。“潜力”被分析为激发的前置宾语。

5）汉语文本依存距离均值为 2.86，俄语文本为 2.1。可以看出，俄汉文本支配词及其从属节点的依存距离大致相当，汉语略高于俄语，数据与刘海涛相关分析计算结果相吻合，刘海涛计算出汉语 MDD 为 2.84（刘海涛 2009：258）雷蕾定量计算的俄语 MDD 为 2.03（雷蕾，Jockers 2018：8），同样是汉语略高。这里推断汉语版本依存距离大于俄语的主要原因可能在于：第一，汉语的依存关系数量多于俄语。第二，汉语含有“的”字结构、“们”字结构的 RAD 右附加模式频繁出现（俄语没有对应关系，故在关系列表中未列出该关系），这使得该名词与支配动词距离大于 1 或者间隔更远。俄语定语通过形容词的词尾变化来实现功能，“们”也是通过名词的数进行词尾变化，它们与支配动词之间距离不会因此增大。第三，俄语动词的时态根据动词词尾的不同来区分，尤其是过去时，如：Видел это по информации из наших регионов. （从各地的信息中看到了这个现象）Видел -это 的依存距离为 1，而汉语为 3，两个词之间有“了”表示时态、“这个”表示数量。第四，汉译本是 yandex 智能翻译软件翻译的结果，是一一对应的机械翻译，其句式以俄语表述为基础，汉语译文与俄语句法复杂程度不相上下。为了排除翻译软件的翻译影响，在此选取观察网 2020 年俄罗斯国情咨文汉译版进行量化对比，结果为：2020 年汉译文文 MDD 为 2.8，人工校对后的 2020 年咨文汉译文本 MDD 数据基本等同于 2021 年咨文 yandex 智能翻译版，可见翻译水平可影响 MDD 值，但影响不大。

6）俄语 NDD 为 1.41，大于汉译文本 2020 年的 1.17，也大于 2021 年汉译文本的 1.16。俄语的 MDD 小于汉译文本，反而 NDD 大于汉语。这点与之前举例“交管处没有收到有汽车被困暴风雪的消息”得出的数据结论一样。与 MDD 相比，NDD 在计算中加入了动词位置值和句长。笔者认为在研究词序比较灵活的屈折语时，NDD 更能准确地对句子的认知消耗进行数字描述，NDD 更加贴合地体现句子理解的难度。NDD 值大于汉语，这个数据也从一个侧面证明了俄语版咨文中的动词位置比汉语版本更偏后，更多居于句子后端。

7）依存距离为 1 或-1 的俄语占比 54.4%，汉语占比 44.2%。俄汉文本中从属词与支配词的依存距离并不算大，间隔为 1 的比例在 50%左右，最远的间隔出现在汉语的 COO 并列关系中，为 56，也即第一个出现的并列成分与最后一个并列成分间隔 56 个词。并列结构用并列连接词“和、与、同、跟、以及和顿号”引入，该结构中各个并列成分在句法层次上是同级别的，但是为使依存理论可以处理这种现象，我们使用 COO 代表并列关系。这种方式下的支配节点，并非是实际的支配者，所谓的实际支配者需要参考语义层面进行处理。很显

然，对并列关系的不同处理方式对句法分析的精确度是有影响的，试设想，我们将并列结构划分为与邻近词的关系，而不是以第一个出现的并列成分为基础节点，数据会有较大差异。所以并列结构的处理需要特别声明：这里是以第一个出现的词为支配词，其它位置在后的认定为从属词，COO 的值都是负值。

五、结论与新问题

依存距离是句中有句法联系的词汇的线性距离。依存距离越长，句子复杂度越高，其平均值是句法复杂度的指标。同时使用定量的方式对词的配阶能力进行具体数字化描述，可以更好地解释词的支配和被支配关系，更准确地处理句中词与词的关系，即句中所有词之间的关系，也使我们有可能更准确的研究熟语和固定搭配。本文主要对 2021 年俄罗斯国情咨文俄汉文本进行 MDD 和 NDD 的计算，并加以对比分析。得出以下结论：

1）MDD 与 NDD 都可以作为对句法复杂度的指标，MDD 与 NDD 距离值越大表明支配词和从属词之间的距离就越大，句子也就越难理解。通过计算发现 MDD 与 NDD 的比对结果是矛盾的，俄语文本的 MDD 小于汉译文本，而 NDD 又大于汉译文本，其中最大的影响因素是动词的位置，动词的位置越偏后，NDD 的值就越大，而动词位置对 MDD 并未产生直接影响。所以笔者认为 NDD 的曲线变化走向更加贴合动词的移动轨迹。

2）同汉语一样，俄语也是支配词置后占比略占优的混合型语言，但俄语版咨文中支配词在前的比例要高于汉语。

3）俄语的词序相对于汉语比较自由，但具相对性。经过计算发现俄语相邻词依存关系占比超过 50%，高于汉语。也就是说，在使用俄语的交流过程中，大脑更习惯于进行最简单、便捷的处理，将支配关系近的词进行直接组合。

4）依存关系可分为三种：形式依存，句法依存和语义依存，形式依存是指通过语法形态和词序的变化来体现元素之间的依存关系。（刘海涛 2009：99）由此可见，句法关系并非是孤立存在的，其配阶的实现受到几个方面的影响，其中包括：语义、语法、语言实践中的配阶使用价值。这几个层面相互影响，但又并非一一完全对应。俄语的语法弥补了词类配阶的不足。而汉语缺乏词类的形态变化，这给汉语的依存关系划分带来很大的问题。因为约束越多，分析出来的关系就越少，反之，仅仅是意合的粘结关系，语法约束少，词与词之间产生的句法关系就增多。

5）句法分析，需要控制分析语料的体裁，体裁越正式，句法越规则，准确率越高。计算机对于短句的分析准确率高于对长句的分析。因为越是长句，越是在配阶的过程中受到语义、语法等方面的影响，在这点上计算机很难做到基于实际交流经验的综合考量分析。

一个新的问题在于，是否可以根据汉语的平均 MDD、NDD 值进行翻译文本的比较，当译文 MDD 值高于汉语平均 MDD 值 2.84 时，从某程度上可以说翻译过于忠实俄语源语而忽略了汉语的语序习惯，当译文 MDD 值低于 2.84 时，则认为语句过于零散，简单。当然需要确切范围，高出或低于平均值多少可以认定为不当翻译？有待于进一步的计算论证。

参考文献

[1]雷　蕾，Matthew L. Jockers. Normalized Dependency Distance: Proposing a New Measure[J]. Journal of Quantitative Linguis，2018 (1).

[2]刘丙丽，牛雅娴，刘海涛. 基于依存句法标注树库的汉语语体差异研究[J]. 语言文字应用，2012(4).

[3]刘海涛. Probability Distribution of Dependency Distance[J]. Glottometrics, 2007(15).

[4]刘海涛. Dependency distance as a mentic of language comprehension difficulty[J]. Journal of Cogitive Science, 2008 (2).

[5]刘海涛. 依存语法的理论与实践[M]. 北京：科学出版社，2009.

[6]王巧林，李雯雯. 基于依存树库的语言计量特征对比分析—以乔治华盛顿和唐纳德特朗普就职演说为例[J]. 安徽理工大学学报（社会科学版），2019 (2).
[7]王　永，刘海涛. 俄语名词的计量特征研究[J]. 浙江大学学报（社会科学版），2013(6).
[8]观察网 https://www.guancha.cn/f-putin/2020_01_19_532168.shtml
[9]STR https://ruscorpora.ru/new/instruction-syntax.html

Comparative Study of Chinese and Russian Based on Dependent Grammatical Annotation Tree Library

— Taking the 2021 Russian State of the Union Address as an Example

Zhang Rong

(Liaoning University, Shenyang 110036, China)

Abstract: This paper uses quantitative linguistics as the research method and dependency grammar as the theoretical framework. The Russian and Chinese versions of the 2021 Russian State of the Union Address are calculated for MDD and NDD, and the proportion of positive and negative dependencies is counted, as well as the proportion of dependencies between adjacent words. It is concluded that, first, the data verifies that the trends of MDD and NDD are not exactly the same; second, using numbers can clarify the language types of Russian and Chinese, and third, Russian word order is relatively free and limited by the range of positions. Discoveries are that grammatical relationship does not exist in isolation, and its realization is affected by several aspects, including: semantics, grammar, and the use value of order in language practice.

Keywords: dependency grammar; dependency distance; State of the Union address; MDD; NDD

作者简介：张嵘(1982—)，辽宁沈阳人，辽宁大学外国语学院教师，博士，研究方向：传媒语言学、对比语言学。

收稿日期: 2023-01-02　　[**责任编辑：叶其松**]