doi:10.3969/j.issn.2096-2991.2020.05.012

基于马克思主义中国化话语语料库的句子研究

邓伯军[1],谭培文[2]

(1.南京航空航天大学 马克思主义学院,江苏 南京 211106;
2.广西师范大学 马克思主义学院,广西 桂林 541004)

[摘 要] 句子研究属于定性研究,语料库研究属于定量研究,句子相对稳定的句法结构为基于语料库的句子的统计学频率研究提供了可能性。基于语料库的句子研究就是要搜集句子结构中的频率证据,为建立句子系统的概率模式提供方法论基础。基于马克思主义中国化话语语料库的句子共现频率分析,在共现频率模式的变化范围中把握句子意义的创造、句子意义的转化、句子意义的分解、句子意义的迁变。基于马克思主义中国化话语语料库的句子语法结构分析,在语法结构的概率中折射出句子意义的转换和句子意义的生成。基于马克思主义中国化话语语料库的句子语义成分分析,在句子配置的频率模式中,来发现句子准确的统计学意义。基于马克思主义中国化话语语料库的句子的统计机器学习分析,决策树的逻辑推理、神经网络的模型建构、聚类的语距测度共同成就了句子的知识发现系统。

[关键词] 马克思主义中国化;语料库;句子;研究

[中图分类号] B0-0;D61 [文献标识码] A [文章编号] 2096-2991(2020)05-0080-07

语言系统具有内在的概率性,概率的意义不在于预测个案,而是预测一般模式,语料库建立了语言系统的概率模式,概率模式以量化形式来解释句子系统的意义问题。基于语料库对句子的统计学分析,主要是依据句子结构的性质和形式,对句子结构的发生频率进行统计学分析。不同的语篇采用的句子结构类型是不同的,也就是说,通过对句子结构的统计学分析,就能发现语篇的叙事手法、体裁风格甚至意识形态取向,进而达到社会批判之目的,为语言学和社会学研究架起沟通的桥梁。"通过对语料库的分析,我们可以看出语言分布的频率与语篇体裁的类型亦有关系;通过分析语料库,我们可以考察从一个给定系统中进行的选择在多大程度上受到以往选择的影响,这种概率有多大;我们可以通过分析语料库来考察语言在人类历史的发展过程(phylogenesis),人类个体对语言习得进程(ontogenesis),个体语篇的动态发展过程中(logogenesis),语言结构从简到繁的概率分布;对语料库的考察还可以帮助我们发现并存的系统之间的关联程度如何。"[1]52从系统论思维来看,句子是一套逻辑化的语言符号系统,基于

[收稿日期]2020-07-30

[基金项目]2014年国家社会科学基金一般项目(14BKS017);2013年广西哲学社会科学规划课题重点项目(13AKS002);2015年广西高等教育本科教学改革工程重点项目(2015JGZ115)

[作者简介]邓伯军(1967-),男,河北保定人,南京航空航天大学马克思主义学院教授,博士,博士生导师,研究方向:马克思主义理论及其当代价值;谭培文(1948-),男,湖南衡山人,广西师范大学马克思主义学院教授,博士,博士生导师,研究方向:马克思主义理论及其当代价值。

语料库对句子的描述性研究,就是要以量化形式衍生出句子构件的性质和结构形式,也就是句子成分的逻辑系统。通过对句子结构的统计学频率分析,来解释句子结构所要表达的意义和功能。因此,语料库语言学就是要搜集句子结构中的频率证据,建立起智能化的语言理解系统,为建立句子系统的概率模式提供方法论基础。运用概率来解释句子系统,句子系统就是一种选择,因为概率就是选择一种结构而不选择另一种结构的概率。在"系统/结构"框架下,系统概念已经将句子结构描写为选择,选择语态:主动或者被动;选择极性:肯定或者否定;选择句型:单句或者复句;选择时态:过去时、现在时或者将来时;选择句类:陈述句、疑问句、祈使句或者感叹句。简而言之,"每一种选择都有一种概率值"[1]42。基于马克思主义中国化话语语料库的句子共现频率分析,建立起马克思主义中国化话语体系的句子结构的频率模式,在共现频率模式的变化范围中把握句子意义的创造、句子意义的转化、句子意义的分解、句子意义的迁变。基于马克思主义中国化话语语料库的句子语法结构分析,从句子的逻辑系统中生成表达语法结构的聚合体,聚合思路使语法结构模型的或然性具有了统计学意义,在语法结构的概率中折射出马克思主义中国化话语体系的句子意义的转换和句子意义的生成。基于马克思主义中国化话语语料库的句子语义成分分析,对句子的主体语义成分、客体语义成分、情景要素语义成分的结构配置角色进行频率统计,从马克思主义中国化话语体系句子配置的频率模式中,来发现马克思主义中国化话语体系句子成分的统计学意义。基于马克思主义中国化话语语料库的句子统计机器学习分析,决策树技术、神经网络技术、聚类技术等交织在一起形成基于马克思主义中国化话语语料库的统计机器学习技术,基于语料库的统计机器学习系统就成为马克思主义中国化话语体系的知识发现系统。

一、句子的共现频率分析

语料库语言学是计算机对自然语言量化处理的重要技术手段。目前,语料库语言学主要探讨计算机可读自然语言文本的采集、标注、存储、检索、索引、频数、平均数、标准差、偏态值、散布系数等定量分析,同时还可以利用计量方法与模型对句子结构的依存关系、制约关系、运作机制展开更为深刻的解释。随着语料库语言学的成熟,特别是机器学习技术的发展,可以量化分析的意义单位的边界不断扩展,把意义单位从词汇组织扩大到句子结构,基于语料库的分析可以识别和解析复杂的句子结构,能够对句子结构进行语法、语义、语用等层面的语言学分析,能够对句子结构进行概念、人际、语篇等功能解释分析。也就是说,基于语料库的分析可以将语言学特征的分析和非语言学特征的分析联接起来。基于语料库的句子结构的功能分析颠覆了转换生成学派以词为基元的分析生成机制,而是以算式语法和词法的形式来生成语言的意义单位。也就是说,"这样的符号单位(多词结构)也是型式与意义的组合,符号单位越稳固地进入语言使用者的语言系统,越被更广泛地使用,反之亦然。换句话说,母语者是机械地整体调动多词结构,而非分析生成;而句子学的发展则真正与语料库语言学的发展休戚相关。毫无疑问,语料库是观察句子共现率和复现率的最佳载体"[2]78。由于句子单位意义的非合成性、结构形式的固定性、搭配序列的词汇化等特征,句子结构中的频率信息可以建立起基于语料库的句子分析的概率模型,无论是极性:肯定/否定,还是语气:直陈语气/疑问语气/祈使语气,抑或及物性:物质过程/心理过程/关系过程,都是带有某种频率的意义组合的倾向性,通过语料库技术的共现频率分析才能观察到语言系统变化的倾向性,也就能够使我们在对频率的历时变化中发现语言意义的变化。

建设马克思主义中国化话语语料库,借助计算机技术对马克思主义中国化话语体系的句子进行大规模的共现频率模式研究。句子作为模块化组织,其意义单位是以算式语法和词法的形式来生成的而不是约定俗成的。"大量语料库和计算机工具使得研究说话者使用语法的模式成为可能——即调查各种语言结构的频率分布,调查语法结构与其他语言和非语言因素之间的关系,以及影响选择不同结构的因素。"[3]36这也决定了句子在语言系统中都带有一定程度的概率

值。借助语料库技术可以识别和解析句子的基本算法和词法，能够对马克思主义中国化话语体系的句子进行语法、语义、语用等层面的语言计量学分析。针对以结构特征为标准的句型系统建立标准句型库，通过将马克思主义中国化话语体系的句子结构模式与句型库中的标准结构模式进行匹配，获得实验模型的句型匹配表，进而以量化形式展现出马克思主义中国化话语体系句子的逻辑特征，为深化马克思主义中国化话语体系的句法、语义、语用关系研究提供数据支持。在对句子定量分析基础上利用语料库技术还可以展开对句子的概念、人际、语篇等功能解释分析。以语料库技术识解句子成分对句子的概念功能作出解释，以语料库技术识别句子的述谓关系对句子的人际功能作出阐释，以语料库技术辨识句子的语言特征对句子的语篇功能作出诠释。实际上，基于语料库的马克思主义中国化话语体系句子研究的中心问题是共现频率。在“系统/结构”的框架下，马克思主义中国化话语体系句子的极性：肯定/否定，语气：直陈语气/疑问语气/祈使语气，及物性：物质过程/心理过程/关系过程等都体现了某种频率的意义组合的倾向性，对马克思主义中国化话语体系句子共现频率模式的研究，在共现频率模式的变化范围中把握句子意义的创造、句子意义的转化、句子意义的分解、句子意义的迁变。

二、句子的语法结构分析

语法是一种抽象化的符号结构，语法本质上是个或然系统，也就是说语法本身就具有内在的概率性，符号结构是可以进行统计学测量的，这样，通过对语法结构的描述性研究来发现句子的意义问题。语料库语法是基于语法学的描写性研究，语料库所揭示出来的语法结构的数量模式，也就是各个语法结构在语篇中出现的相对频率，这不仅展现了语法结构的基本特征，同时也呈现出其概念意义、人际意义、语篇意义。“语法结构就是各种成分在适当位置（place）上有秩序的排列。一个结构中的不同位置只能通过顺序来区分：如结构XXX有三个位置。另一方面，不同的成分也可以由某种关系而不是顺序来区分，如：XYZ这个结构包括三个成分，它们是（要构成一个结构就必须是）按照位置排序的，尽管可以用清单（X,Y,Z）的形式来列举构成特定结构的成分。结构总是某个给定单位的结构。”[4]35-36语法结构被表述为符号的线性排列，每个语言符号会因为成分和位置的不同而形成不同的意义，因此，引入语言计量学的精密度理论来测量语法结构所表现出来的形式意义和语境意义成为可能。“随着描写越来越精密，语法关系网络也变得越来越复杂。不同标准之间的交互作用使得范畴与范畴之间，以及范畴与说明项之间越来越成为一种‘多/少’而不是‘是/非’关系。因此，有必要根据概率来权衡各种标准并且做出表述。”[4]38-39汉语句子的语法结构主要有并列结构、主谓结构、偏正结构、动宾结构、动补结构、介宾结构、复指结构、连动结构、固定结构等。随着机器学习技术的发展，基于语料库的语法结构的自动分析和频率分析呈现出繁荣态势。通过语料库能够建立起语法系统的或然性轮廓，揭示出语法结构出现的相对频率，“通过使用通用的语法解析器，语料库使检测所体现的陈述成为可能：可能更为有效的是，通过为一些具体的语法系统设计模式匹配程序，可以把从语料库中选取的样本的分析与结果进行匹配。某种形式的解析或数量匹配对定量研究必不可少，因为要计算的数目远远超出我们所希望人工能加工的范围”[4]387。基于语料库的语法结构研究本质上属于描写性研究，其目的在于最大限度地建构起以聚合结构为标准的语法结构系统，最大限度地发挥其作为描写性研究的信息源的价值。基于语料库的语法结构研究的主要依据是句子构件及其结构形式，句子的词序、句子成分的多少、句子成分的词性，等等，都是决定语法结构的因素。因此就要以建构起基于语料库的句子成分分析和语法结构匹配的计算模式，从而实现语法结构分布统计的顺利完成。但是，由于语法逻辑并不直接等同于数学逻辑，语法逻辑遵循自然语言的演化逻辑，因而语法结构的计量学研究并不能全面地保证语法结构测量的绝对准确性，但仍然可以以准确性来实现对语法结构的描写性研究。“语法结构本质上是一种或然系统，任何特征的意义的一个重要部分就是相互定义的特征在该系统中出现的频率”[4]380。

建立马克思主义中国化话语语料库,利用语料库技术对马克思主义中国化话语体系的句子的语法结构进行计量学分析。汉语句子的语法结构主要有并列结构、主谓结构、偏正结构、动宾结构、动补结构、介宾结构、复指结构、连动结构、固定结构等。在不同的马克思主义中国化话语体系的语篇中有不同的语法结构聚居群,学术类语篇的语法结构不同于政论类语篇的语法结构,政论类语篇的语法结构不同于叙事类语篇的语法结构,叙事类语篇的语法结构不同于艺术类语篇的语法结构。基于语料库对马克思主义中国化话语体系的句子的语法结构的统计学研究,就能揭示出语篇不同的主旨和不同的意义,既能通过语法生成语义,又能通过语法结构还原语义结构。从这个意义上讲,马克思主义中国化话语体系本质上是语法—语义系统。建设马克思主义中国化话语语料库,就要引入语法结构的量化模型,语法结构在这个数量模型中纵向地表征为既定选项组成的系统网络,从系统网络中生成表达的聚合体,聚合体表征将语法从结构主义框架中解放出来,聚合思路使语法结构模型的或然性具有了意义,而正是或然性提供了系统网络中特定语法结构出现的相对可能性。"尽管系统网络不是一个神经过程模型,这种复杂程度的语法的存在并非不可能;也就是说,只要选择点的数目不算太多,每个选择点本身极其简单,作为这些选择点相互交叉的产物,可以建立一个模型。"[4]375在这个系统网络中,语义和语法能够对语言资源进行自由编排,在拓扑空间和类型空间中形成其意义的潜势,而语言的意义正是在语法结构中作为整体的意义潜势而被语境化,因此要通过语言自身的不确定性来处理语法结构的不确定性。也就是说,通过对语法结构的数量模型的建构,就能初步实现对马克思主义中国化话语体系句子的语法结构的概率分析,在语法结构的概率中折射出意义的转换和意义的生成。正如韩礼德所言,"意义实践的背后是各种各样的意义系统。事实上,我们常常用'系统'这个术语来涵盖系统和过程这两个概念,它既是潜势也是例示;这样,意义系统就是一个意义潜势加上表意行为的例示。现在,有一种特殊的意义系统,它包含了语法;这个系统在两个阶段'有意义',有一个专门的措词阶段作为意义建构的基础;换句话说,这个系统的'内容层'除了包含语义外也包含语法,我们将这个特殊的意义系统看作语法—语义系统。正是语法的存在赋予了这样一个系统独特的创造(而不仅仅是反映)意义的潜势"[4]368。

三、句子的语义成分分析

句子乃是一个语义单位,每个语义单位都有特定的树形结构,每个部分都在有机整体中拥有各自的功能。通常来讲,概念成分、人际成分、语篇成分是语义单位的组成部分,每个组成部分都各自承担着相应的功能,并且一起构成句子、小句、短语或词组。"句子有一个概念成分,基于及物性系统之上,过程、参与者和环境因子构成了真实世界中的语义;还包括将各种可以命名的成分进行分类的名称系统;它有一个人际成分,包括语气、情态、人称、调式、以及作为意义选择项的所有态度标识;它有一个语篇成分,'功能句子观'(主位和信息给予系统)和作为衔接手段的指称、省略和连接。"[4]210也就是说,每一个语义成分都对句子的整体意义建构起作用,句子的意义是概念功能、人际功能、语篇功能这三种成分在语义过程中共同作用的结果。每一种语义成分都有其各自不同的配置,也就是有各自独特的结构,而每一个部分在有机整体中拥有各自不同的功用。"句子的概念成分生成'动作者—动作—目标'类结构:由过程、中介、施动者、受益者、范围、程度、地点、方式、原因等组成的配置。"[4]210句子的成分配置是相互联系、相互制约、相互转化的,构成句子的各个配置成分不是孤立存在的个体,而是结构性配置中的角色,句子呈现为结构配置的角色链。语料库是句子概念意义的理论来源,借助语料库技术对句子的结构配置角色进行统计学研究,从句子配置的频率模式中来发现句子准确的统计学意义。"句子的人际成分生成许多所谓'情态—命题'结构:由主语、定式成分、情态、谓语、补语和状语组成的配置。"[4]210句子的人际意义体现的是其语旨问题,任何句子都有其自身的语旨,句子作为情感系统的原材料以隐喻的形式形成一种强有力的语义运动,句子的人际特征是对语篇的社会活动本质的一种隐喻,是对人

与人之间的社会符号结构框架的一种探究。句子体现了人际功能的元语言功能，凭借语料库技术对句子的认知成分、情态成分和行为成分的统计学分析，来完成对人际功能中的身份、地位、态度、动机的确认，实现对人际功能的推断、判断和评价等。“句子的语篇成分生成主位结构和信息结构，是由主位和述位的、已知信息和新信息构成的配置，此外，还有非配置关系的衔接成分。”[4]210句子在语篇的主位结构和信息系统中位置的选择，从主位到述位，从已知信息到新信息，来确定句子的指称在空间中的位置，就此把握句子的社会意义，句子的语式也就以累积的形式展现出来。句子的语法建构既体现概念意义也体现人际意义，依靠语料库技术对句子概念意义和人际意义的统计学分析，在概念意义和人际意义的交融中展现出句子的语篇意义。因此，“我们发现语义系统具有一个内在的组织，语言的各种社会功能都在这个组织中得到清楚的体现。”[5]212不同类型的语法结构与语义功能成分之间有着系统性的联系，在对语法系统的诠释中就能发现特定的语义功能成分，在对语言功能成分的解读中也能发现特定的语法结构。“每一种意义都倾向于由一种特定的结构体现。因此在对语篇的编码中，每一个意义成分都会影响着语法结构。但每种意义成分都有其特定意义模式的标记”[5]214。

与普通句子一致，马克思主义中国化话语体系的句子结构也有概念成分、人际成分、语篇成分等三个组成部分。马克思主义中国化话语体系句子结构的概念成分有主体语义成分，如施事、当事、领事等；有客体语义成分，如受事、结果、系事、客事、对象、分事、共事；有情景要素语义成分，如时间、处所、范围、方式、数量等。正是在主体语义成分、客体语义成分、情景要素语义成分的结构性配置中形成了马克思主义中国化话语体系的句子结构。通过建设马克思主义中国化话语语料库，对句子的主体语义成分、客体语义成分、情景要素语义成分的结构配置角色进行频率分析，从马克思主义中国化话语体系句子配置的频率模式中来发现马克思主义中国化话语体系句子准确的统计学意义。马克思主义中国化话语体系句子结构的人际成分主要有认知成分、情感成分、行为成分。认知成分是对客观事物的基本看法和基本观点，构成了马克思主义中国化话语体系对人际关系基本态度的逻辑基础。情感成分是对客观事物的情感体验，构成了马克思主义中国化话语体系对人际关系基本态度的核心要旨。行为成分是对客观事物的反应倾向，构成了马克思主义中国化话语体系对人际关系的意向性。建设马克思主义中国化话语语料库，对句子结构人际成分的认知成分、情感成分、行为成分进行计量学分析，以句子的成分频率模型来完成对马克思主义中国化话语体系人际关系中的身份、地位、态度、动机的确认，以句子频率统计形式实现对马克思主义中国化话语体系的人际关系的推断、判断和评价等。马克思主义中国化话语体系句子结构的语篇成分主要有主位结构、信息结构和衔接关系等。句子的主位结构乃是功能语法中语篇功能的重要组成部分，“一个句子可以分为主位、述位和连位三个部分。主位是话语的出发点；述位是话语的核心内容；连位是把主位和述位连接起来的过渡成分”[6]136。句子的信息结构乃是指根据已知信息和未知信息进行的句子组织，句子主语和话题往往是已知信息，句子谓语和述题往往是未知信息，信息结构能够传递出信息交流的基本内涵。句子的衔接关系乃是指句子以相互交织的形式进入到语篇之中，形成从对象到话题到语序到语义再到语境有规律的链条功能关系。建设马克思主义中国化话语语料库，对句子的主位结构、信息结构、衔接关系等进行统计分析，能够准确理解句子的话语信息，能够准确推断出段落的主旨，乃至整篇文章的主题等具有重要的意义。

四、句子的统计机器学习分析

机器学习的核心功能是知识发现，基于机器学习的知识发现系统是通过学习算法来实现的，也就是说从存储在数据库中的经验数据中发现知识，数据库建设成为基于机器学习的知识发现系统的数据基础。机器学习研究的热点和前沿问题主要是“决策树”“神经网络”“聚类”“增强学习”“贝尔斯网络”“数据挖掘”“支持向量机”“独立组件分析”等。目前机器学习形成了以“分类技术”为核心，向“分类算法的演进”和“计算学习

理论”两个方面演化,理论与实践的动态平衡构成机器学习演进的内生动力。基于语料库的统计机器学习逐渐成为语言理解和语言生成的自动化系统。“当前的自然语言处理研究提倡建立语料库,使用统计机器学习的方法,让计算机自动地从浩如烟海的语料库中获取准确的语言知识。”[7]10以软计算实现数学逻辑和语法逻辑的无缝对接,建构起智能计算的语言模型的统计机器学习系统,计算过程的每一步都要置放在语言模型的统计机器学习系统中来进行符码转换,而符码转换的关键是要解决词法歧义、句法歧义、语义歧义以及回指歧义等问题。从句子学角度来看,“任何单词都不具备独立含义,而只有在考察其搭配、类联接选择、语义趋向和语义韵之后才可以确定其含义和使用情况。也就是说,理解语义的最普遍单位实际上并非是单词而是多词,而词法消歧又会影响到句法消歧和语义消歧。因此,多词单位的准确理解就直接影响自然语言处理的效果和准确性”[2]83-84。在统计机器学习的框架下,“复杂现象的科学规律大多数是统计学意义上的规律,或者说概然性规律。如果要想证实或证伪一个假说,就需要足够的经验事实依据,在统计学意义上,去确认、说明和解释假说”[8]91-92。对多词单位也就是对句子的准确理解,以句子为中心对其上下文进行编码,计算出上下文与中心句子所形成的语态、时态、极性、句型、句类等的共现频率,从而发现句子的指示系统、逻辑系统、关系系统、功能系统等。“在某种条件下,说话者只能从小部分可能性的封闭集合中进行一种选择,因此谈论选择一种或者另一种可能性的概率便有了意义。我希望不仅把每个系统建构成‘选择a或b或c的形式’,而且把它们建构成‘选择a或b或c,每种选择都带有一定的概率值’的形式。换言之,我断言任何语言系统的一个内在特点便是术语的相对概率。”[1]124基于语料库的统计机器学习系统将根据聚类规则自动地对数据进行选择和调整,直至网络结构能够合理地反映出句子结构样本的统计分析结果,因此,句子的选择语域频率被整体频率模式所控制的程度便有了意义。

基于机器学习的知识发现系统为基于语料库的句子研究提供新的路径,从数据总结,到概念描述,到分类,到聚类、相关性分析,到偏差分析,再到建模,基于机器学习的知识发现系统成就了从经验积累过程,到知识生成过程,再到知识运用过程的逻辑序列。将语料库语言学引入到马克思主义中国化话语体系研究之中,使用统计机器学习的方式,建构智能化的马克思主义中国化话语语料库,将聚类技术作为马克思主义中国化话语语料库研究数据挖掘的基础方法,根据语义距离的测度作出描写来形成有意义的知识发现。利用决策树分类技术作为马克思主义中国化话语语料库研究的归纳推理算法,利用训练经验数据来实现对话语数据的聚类分析。将神经网络作为马克思主义中国化话语语料库研究样本数据的结构函数模式,以数学函数关系实现对话语数据的聚类分析。这样,决策树技术和神经网络技术等交织在一起形成统计机器学习的聚类技术。统计机器学习技术将深刻地改变马克思主义中国化话语体系的句子的研究范式。对马克思主义中国化话语体系的句子的理解,以句子为中心对其上下文进行编码,着手句子结构的共现分析,计算出上下文与中心句子所形成的核心关键词的共现频率,绘制出句子关键词的知识图谱,知识图谱的语义链接,使得检索引擎能够用基于实体的检索来代替基于字符串的检索,实现对马克思主义中国化话语体系的句子检索时的歧义消除,通过聚类分析在图谱中展现出马克思主义中国化话语体系研究领域的知识群,在对知识群的统计分析中发现马克思主义中国化话语体系句子的指示系统、逻辑系统、关系系统、功能系统等,揭示出马克思主义中国化话语体系的语态、时态、极性、句型、句类等句子特征,从而展现出马克思主义中国化话语体系的核心结构、发展历史、前沿领域、知识架构。基于机器学习的马克思主义中国化话语体系的知识发现是一个准周期的过程,从对毛泽东思想的研究,到邓小平理论研究,到“三个代表”重要思想研究,到科学发展观研究,再到习近平新时代中国特色社会主义思想研究,都是受到时代变化和环境变迁等多种因素通过迭代、化合、组合而形成的复杂过程,每次新理念、新思想、新战略的提出,都是

建立在前面类似概念或者思想的基础之上的，可能重新提出某一概念或思想，但内涵会有新的变化，也可能创新某一概念或思想，但必然有其历史根据。

［参考文献］

[1]韩礼德.语言的可计算性与可量化研究[M].北京：北京大学出版社，2015.

[2]宋丽珏.人工智能时代语料库句子学考察[J].学习与探索，2017(12)：78-85.

[3]道格拉斯·比伯，苏珊·康拉德，兰迪·瑞潘.语料库语言学[M].刘颖，胡海涛，译.北京：清华大学出版社，2012.

[4]韩礼德.论语法[M].北京：北京大学出版社，2015.

[5]韩礼德.作为社会符号的语言：语言与意义的社会诠释[M].北京：北京大学出版社，2015.

[6]胡壮麟，朱永生，张德禄.系统功能语法概论[M].长沙：湖南教育出版社，1989.

[7]冯志伟.自然语言处理简明教程[M].上海：上海外语教育出版社，2012.

[8]赵玉鹏.机器学习的哲学探索[M].北京：中央编译出版社，2013.

【责任编辑　张宝君】

Sentence Research Based on Chinese Marxism Discourse Corpus

DENG Bojun[1], TAN Peiwen[2]

(1. College of Marxism, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu 211106, China; 2. College of Marxism, Guangxi Normal University, Guilin, Guangxi 541004, China)

[Abstract] The sentence research based on corpus is to collect frequency evidence in the structure of sentence, and to provide a methodological basis for establishing probabilistic patterns in sentence system. The co-occurrence frequency analysis based on Chinese Marxism Discourse Corpus, the creation of sentence meaning, the transformation of sentence meaning, the decomposition of sentence meaning, and the change of sentence meaning are grasped in the variation range of the co-occurrence frequency pattern. The analysis of sentence grammatical structure based on Chinese Marxism Discourse Corpus reflects the translation of sentence meaning and the formation of sentence meaning in the probability of grammatical structure. The sentence semantics analysis based on Chinese Marxism Discourse Corpus, the exact statistical significance of sentences is found in the frequency pattern of sentence allocation. The statistical machine learning exploration of sentence based on Chinese Marxism Discourse Corpus, decision tree technology and neural network technology are intertwined to form the clustering technology of statistical machine learning. The probabilistic principle in statistical machine learning has enabled a corpus-based knowledge discovery system.

[Key works] Chinese Marxism; corpus; sentence; research