东北亚国家政治话语数据库建设和应用研究综论

傅兴尚

（大连外国语大学中国东北亚语言研究中心，大连 116044；黑龙江大学俄罗斯语言文学与文化研究中心，哈尔滨 150080）

提　要：政治话语数据库是指收集和整理政治话语的电子数据库，其中包括了政治话语的来源、时间、作者、内容、语言特点、文化背景等多个方面的信息。建立层级化政治话语数据库可以帮助研究者进行大规模语言文化分析和深入的话语批评分析，从而揭示政治话语背后的社会、文化和权力结构的信息。因此，“东北亚政治话语数据库建设及应用研究”体现了我国与东北亚合作的战略导向和现实需要。核心成果：“东北亚国家政治话语结构化数据库及管理应用平台”不仅是国家级语言资源基础设施，而且在此基础上可开发和实现政治领域的语言查询、知识查询、汉外对比与精确翻译、政治话语语言动态监测、重大政治事件舆情分析、政治倾向梳理预测、政治倾向和政治目的文化语用计算等功能。此外，通过数据共享机制，为政治话语的多维研究提供数据资源保障和语言学计算模型保障。本文从四个方面对其进行概要性综合论述：1）课题提出的背景和总体目标；2）课题的理论依据和意义；3）课题研究的主要内容；4）支撑课题研究的 NLP 能力建设。

关键词：东北亚；政治话语；数据库；NLP；数据库应用

中图分类号：H08　　**文献标识码：A**

政治话语是指与政治相关的言论、宣传、报道、演讲等，它是政治现象的重要表现形式之一。政治话语可以反映出政治文化、权力结构、意识形态和社会影响等方面的信息，因此对政治话语的研究是对政治现象及其背后的社会、文化和权力结构深入理解的重要途径。政治话语数据库是指收集和整理政治话语的电子数据库，其中包括了政治话语的来源、时间、作者、内容、语言特点、文化背景等多个方面的信息。建立政治话语数据库可以帮助研究者进行大规模语言文化分析和深入的话语批评分析，从而揭示政治话语背后的社会、文化和权力结构的信息。本文从四个方面对规划课题“东北亚国家政治话语数据库建设和应用研究”（以下简称“课题”）进行概要性综合论述：1）课题提出的背景和总体目标；2）课题的理论依据和意义；3）课题研究的主要内容；4）支撑课题研究的 NLP 能力建设。

1 课题提出的背景和总体目标

东北亚国家包括中国、俄罗斯、日本、朝鲜、韩国、蒙古。2019 年 8 月，习近平总书记在第十二届中国-东北亚博览会上指出：“东北亚是全球发展最具活力的地区之一，要汇聚共识、推进合作、扩大成果。”为助力国家战略的实施，教育部语言文字信息管理司会同有关高校、研究机构专门成立了中国东北亚语言研究中心，布局和启动了“东北亚外交外事高端人才”博士培养项目。东北亚研究中心要研究什么？东北亚外交外事高端人才要学习什么、掌握什么、怎么培养？如何把握各国政治外交立场？如何精准分析各国政策、预测外交格局

的走向？不言而喻，政治话语和应用研究是重中之重。目前，在大数据驱动和数字人文背景下，我国还没有建设东北亚国家政治话语数据库，更没有基于数据库的系统性多维应用研究。所以，课题“东北亚政治话语数据库建设及应用研究”首先体现了我国与东北亚合作的战略导向和现实需要。核心成果“东北亚国家政治话语结构化数据库及管理应用平台”不仅是国家级语言资源基础设施，而且在此基础上可开发和实现政治领域的语言查询、知识查询、汉外对比与精确翻译、政治话语语言动态监测、重大政治事件舆情分析、政治倾向梳理预测、政治倾向和政治目的文化语用计算等功能。此外，通过数据共享机制，为政治话语的多维研究提供数据资源保障和语言学计算模型保障。

当前，大数据、云计算、自然语言处理等人工智能新一轮科技革命为语言学与其他哲学社会科学以及理工科交叉融合研究提供了理论支撑和实践保障。课题秉承数字人文研究发展理念，综合大数据思维以及政治语言学、计算语言学、文化语用学、文本计算和内容分析、国别与区域学等学科的研究方法，开展创新型的新文科建设和应用实践。其中的“政治话语数据库”要建成涵盖中、俄、日、韩、朝、蒙等多语种的国家级系统、全面、统一、开放、共享的数据集成平台和管理应用平台。该平台既包含东北亚国家政治话语文本库（如首脑讲话、国情咨文、峰会致辞、竞选演讲、施政纲领），又包括知识库平台，如双语术语库（人类命运共同体、中俄蒙经济走廊、冰上丝路带、碳达峰）、缩略语库（如 ВСУ<乌克兰军队>、ЗРП<防控导弹团>）、文化习语表达库（如不吃这一套，压舱石）、专名库（如工商联、今日俄罗斯）、重大政治事件背景库（如北约东扩、俄乌冲突）等。

在功能设计上，总平台可提供语言查询、知识查询、政治话语语言动态监测、重大政治事件评价、政治倾向梳理预测、政治倾向和政治目的文化语用计算等语言学模型和智能技术的实现。换言之，课题是集数据资源、科研创新、教育创新、语言服务、知识服务、智库服务功能为一体的应用型基础设施，又属于新基建框架下的语言资源基础设施。可体现为多方面的成果，包括但不限于：1）综合平台：东北亚国家政治话语结构化数据库及管理应用平台；2）数字人文背景下的东北亚国家政治话语多维研究；3）面向政治话语内容和意向计算的语言学模型；4）基于政治话语数据库的应用：理论研究和实现技术；5）政治话语汉外语言文化属性对比研究。在 AI 时代和新文科建设背景下，以上研究不仅具有理论创新意义，而且具有重大的应用研究价值。

2 课题的理论依据和意义

顾名思义，政治话语数据库是指收集和整理政治话语（指与政治相关的言论、宣传、报道、演讲等）的电子数据库。其目的旨在研究政治话语的语言特点、政治文化、权力结构、意识形态和社会影响等方面。建立政治话语数据库的理论依据主要来自语言学、社会学、政治学和文化研究等多个学科的相关理论。其中，语言学为政治话语语言结构和语用特点的分析提供方法；社会学和政治学视角可研究政治话语背后的权力结构和政治文化的理解；文化研究则关注政治话语在文化交流和符号建构中的作用。综合这些理论，可以构建出较为全面的政治话语研究框架，从而为建立政治话语数据库提供理论支持。

建立政治话语数据库具有多重意义。首先，政治话语是政治现象的重要表现形式之一，通过建立政治话语数据库可以对政治现象进行更加系统和深入的研究。其次，政治话语是权力争夺和文化交锋的重要场所，因此政治话语数据库可以为政治和文化研究提供重要的素材。再者，政治话语是社会舆论的重要组成部分，因此政治话语数据库也可以用于研究社会舆论和公众意见的变化。具体体现为以下几个方面。

（1）政治话语是语言与政治关系研究中的一个重要概念，它是以政治为目的的言语行为，是政治意识的语言实践。语言学界开始关注和研究政治话语始于 20 世纪 50 年代的欧洲，也包括俄罗斯学者。作为一种话语类型和语言存在或呈现形式，政治话语具有普通语言

所具有的基本特征，同时它也是一种政治权利工具、一种操纵社会的手段，其典型功能是实现一定的政治目的，为政治权力而斗争。近年来我国一些学者开始注重政治语言学的研究，成果主要包括引介国外政治学有关理论，包括概念、类型、一些基本方法；就某一具体类型和主题研究政治话语的语言表达特点、政治隐喻现象、政治习语的翻译策略等展开研究。尤其值得一提的是，俄语语言学界从服务中俄战略协作的国家战略需求出发，在政治语言学的框架内开展外事外交领域的话语研究，阐释俄语外交话语的建构规律和立场、意图、话语操控等分析研究，以期为外交政策制定和话语建构提供智库服务。但总体来说，目前的研究队伍还不大、研究内容还较零散，局限于部分学者的个人兴趣点，研究内容在深度和广度上都还不够，研究方法上没有发挥大数据的作用。本课题克服上述缺陷，系统建构东北亚国家多语种政治话语数据库，既有文本数据库、也有领域知识库，这种数据平台无疑对语言研究和智能技术研发的“双轮驱动”意义重大。政治话语的多维研究将以中俄领域话语数据库为基础，利用自动学习、大数据技术、计量语言学、政治语言学的研究方法，支撑理论体系建设并具体化研究内容，研究成果回归大数据，为语言研究、语言教学、语言服务、语言智能、语言规划和管理以及知识服务、智库服务提供数据源和方法论的双重保障。

课题首次以政治话语为研究对象，从人类政治领域出发，对领域和话语的对应性、领域对话语的选择性、领域语言群体、领域活动的社会功能、领域知识等进行基于大数据的研究和阐释，构建理论体系，在学科建设上具有创新意义；研究的视角、内容和方法体现当代语言学研究趋势，具有时代性意义；课题呈现各学科的研究方法和自动学习技术、大数据技术，为汉语和其他语种的对比研究提供新的视角和例证依据，也为相关问题研究奠定方法论基础，具有示范和探索意义。研究成果涉及当今热点、难点，可应用于多个领域，具有广阔的应用前景。从“需求导向”和“服务社会”出发，课题具有重大应用价值。目前，不论是“一带一路”倡议，还是建立“人类命运共同体”，都需要语言互通，以语言的钥匙打通心扉、达到“人心相通”和“文化认同”。不论是“传递中国声音、贡献中国智慧”，还是“提升国家语言能力，建构国际话语体系，获取国际话语权”都需充分重视语言的作用。语言先行，因为语言是人类交际和思维最为重要的符号系统，是文化最为重要的组成部分，是文化最为重要的载体，也是交流、交往、交互和消除误解、化解矛盾的手段。政治话语研究可解决政治领域问题，解决政治语言需求，如政治外交领域要解决好话语建构、意图实现、立场表达等，此研究具有现实指导意义。

（2）服务于国家战略需求的数字人文研究与实践方兴未艾，文、理、工跨界与学科交叉融合研究蓬勃发展。作为研究热点，政治语言学本身就是研究语言和政治共变关系的交叉学科，如果以政治话语（文本）为研究对象，再与国别区域学、文化语用学和语言智能技术结合起来开展研究和成果转化应用，那么其前景是非常广阔的。人类的社会活动分不同的领域，政治、经济、文化、科技、教育、人文、司法等。政治领域涉及政党纲领、党派主张、治国理政、国际交往等，是重要的社会活动领域。与此相应，政治话语是政治信息的符号载体和政治交流的工具，它是以政治为目的的言语行为，是政治意识、政治主张、政治立场、政治鼓动的语言实践。换言之，政治话语是社会政治生活的语言表述，是政治活动、政治制度及政治文化在话语上的反映。政治话语具有普通语言所具有的基本特征，同时也是一种政治权利工具、一种操纵社会的手段，其典型功能是实现一定的政治目的，为政治权力服务。政治文本表征语言与政治的对话关系。政治交际内化于政治文本，政治文本显像政治交际。所以，研究政治话语语言平面的属性和特征，研究话语政治观点、政治意识、政治立场、政治目的、政治倾向、权利操控等政治平面的信息、交际意向、深层意义，以及政治话语的意向类别和意识演变的逻辑过程，研究话语特点与政治功能的对应性，研究以数据库平台为支撑实现政治话语的语言监测、政治观点和政治倾向的监测计算等都具有重要意义。

（3）人工智能革命带来社会变革和技术红利。目前，标志着第四次科技革命的人工智能（AI）时代已到来，智能技术不仅本身是重要的科学领域和产业集群，而且渗透并融合到政治、经济、文化、教育、科技和外交等各个领域，从根本上改变了人类生存、生活、生产

方式，颠覆了人类感知方式、认知方式和创新方式。AI 正重新定义世界、定义未来、定义一切。这一大背景决定了本课题要采用人工智能领域的成果，尤其是要对自动学习和大数据技术展开研究，同时把成果通过大数据平台应用到科研活动、语言智能开发、教学和人才培养等方面，服务于社会，服务于国际合作、大国外交、国家战略、国家经济社会发展需要。

（4）语言学研究呈现新的发展趋势。人类成为万物之灵是因为具备最重要的智能——语言能力。这些都决定了语言是具有多维属性的矛盾统一体。这为语言及其功能的多元化、立体性研究提供了必要和可能。纵观异彩纷呈、派别林立的语言学发展历程，语言学研究呈现以下趋势：1）更加重视不脱离具体应用场景的话语研究；2）更加重视面向自动学习和语言智能的知识数字化研究；3）更加重视基于大数据技术的语言和知识领域的研究；4）更加重视语言学与其他学科交叉融合研究；5）更加重视语言规划、语言政策、国际话语体系建构、语言服务、语言智能、语言教育等实践或应用研究。

（5）开展语言智能研究和人才培养的需要。从近几年开展语言智能研究和语言服务的实践来看，语言研究要从领域入手。政治话语是重要的语言生活领域，也是国家语言能力的重要体现。即便是表达同一个政治立场、政治意向，政治话语也呈现诸多文化差异。通过大数据分析，揭示这种属性差异，挖掘这种语言差异，对于国家间民心相通、政治互信、大国外交具有重要的意义。在语言服务、语言翻译、语言教学、人才培养化与国际话语体系建构、语言治理等方面也具有重要价值。外语教学要注重语言能力和领域知识素养并举的复合型人才培养，语言规划和语言生活管理也需要以社会领域为单元做到精准和精细，所以开展政治话语研究十分有益、十分必要。这也与近年来语言学研究热点和趋势协调一致、相得益彰。

3 课题研究的主要内容

政治话语数据库的建设涉及到政治话语的分类、收集、整理和分析等多个方面，其中最为关键的是政治话语的分类和整理。政治话语的分类可以根据政治话语的内容、来源、时间、作者、地域、语言特点、文化背景等多个方面进行分类。政治话语的整理可以采用自然语言处理技术，对政治话语进行文本清洗、关键词提取、主题分析、情感分析、网络分析等多个方面的处理，从而提取出政治话语的主要信息和特点。政治话语数据库的应用范围非常广泛，包括但不限于以下几个方面：

1）政治话语研究：政治话语数据库可以为政治话语研究提供大量的素材和数据，揭示政治话语的社会、文化和权力结构的信息。

2）社会舆论研究：政治话语数据库可以为社会舆论研究提供大量的素材和数据，研究社会舆论和公众意见的变化和演变。

3)政策制定和决策支持：政治话语数据库可以为政策制定和决策支持提供参考和依据，从而更好地制定和实施政策。

4）教育和宣传工作：政治话语数据库可以为教育和宣传工作提供参考和依据，提高政治素质和意识形态水平。

总之，政治话语数据库是政治话语研究的重要工具和平台，它为政治话语研究提供了更为广阔和深入的研究空间，也为社会和政治变革提供了更为科学和有效的支持和指导。

在研究目标和基本思路上，一方面，设计、建设好东北亚国家政治话语结构化数据库及管理应用平台，另一方面，从任务驱动、需求驱动出发，研究如何应用好该平台。不论设计、建设平台，还是应用好平台，都需从两个层面展开研究：一是理论层面，包括支撑语言智能技术的语言学模型研究、政治领域知识文化信息的研究、政治交际与话语建构的关系；二是实践层面，研发实用性智能系统，进行语言智能技术和算法实现研究。成果最终在教学与人才培养、语言智能技术、语言文化对外传播以及语言、知识和智库服务等领域服务。因而，涉及诸多研究任务和问题，构成课题的基本研究内容和总体框架，见下表。

表：课题主要研究内容一览表

主要研究内容	政治话语数据库平台的设计和建设	政治话语数据库平台的应用
理论层面（语言学模型和政治交际模型）	政治话语的边界和类别研究； 政治话语的多维属性和结构层级（注 1）； 政治话语的语言特性和政治交际特性； 政治话语中术语、缩略语、专名、特有文化习语的语言文化特征； 政治话语的交际主体、交际场景研究；	政治话语的语言内外因素研究； 政治话语的语言特征库（注 2）； 政治话语社会功能研究； 政治话语交际意图类型研究； 实现政治目的的话语建构策略； 政治话语的交际意图与语言选择； 面向政治话语内容和意向计算的语言学模型（如，鼓动性政治话语的在遣词造句上的特点）； 政治话语汉外文化属性的对比研究；
实践层面（语言技术和算法实现）	政治话语的自动识别； 术语和专名自动识别； 数据的属性化、标签化处理； 重大事件背景知识的自动提取； 以上列成果为基础，建成涵盖中、俄、日、韩、蒙等多语种的数据集成平台和管理应用平台。既包含东北亚国家政治话语文本库，又包括政治话语知识库。	利用 NLP 中的统计技术、词向量计术、深度学习技术和语义相似度计算等技术实现； 政治话语语言动态监测； 政治观点和政治倾向的动态监测； 政治话语的知识抽取和信息深度挖掘； 政治话语的双语句对库建设；
应用领域	以上还可应用于：1）机器翻译、知识问答等语言智能技术的研发；2）在语言和领域知识层面为科研和教学以及外交、外事人才培养赋能；3）提高翻译服务、知识服务和智库服务质量和效果方面；4）助力语言推广和文化对外精准传播。	

*注 1：政治话语的多维属性和结构层级是为建设结构化数据库服务的，如：语种属性（中文、俄文、韩文、朝文、蒙文）、双语对应属性（中俄、中日、中蒙、中韩、中朝）、话语交际场合属性（首脑致辞、会谈、国情咨文、政党纲领、治国理政、就职演说、外交声明）、话语主题属性（气候问题、人口问题、就业、教育、环保、知识产权、地区安全等）；语言属性（词频统计模型、词类统计模型、句子长度、关键词等）2：政治文本语言特征库包括： 1）代表整个语言空间的全域语汇库，含有词频属性、词性属性；2）政治领域语汇分库，如基础语汇库、领域常用语库、领域核心语库、领域术语库；3）句法特征集，包括句式、句长、句型等；4）语义特征集，如具体名词和抽象名词分布、多频词语义特征、情态词分布；5）修辞特征集：如否定句和反问句、排比句的使用，隐喻、反语等修辞格的使用等；6）语用特征集，如领域话语中言语行为的主要类型表达手段等。

4 支撑课题的 NLP 能力建设

NLP（Natural Language Processing, NLP）是一种计算机科学和语言学的交叉学科，涵盖了语言学、计算机科学、数学等多个学科领域，旨在让计算机能够理解、解释、生成自然语言。在本课题研究中，应用许多自然语言处理的技术和方法可以极大地提高研究效率和质量。如以下 NLP 能力建设非常重要。

1）分词：将连续的文本划分为有意义的词语序列，为后续处理和分析提供基础。分词可以使用现成的分词工具，也可以基于特定的词典和规则进行自定义。

2）词性标注：对分词后的词语进行词性标注，即确定每个词语的语法类别，如名词、动词、形容词等。词性标注可以帮助研究者深入了解政治话语的语言结构和用法。

3）词形还原和形态信息标注。

4）命名实体识别：识别文本中的人名、地名、组织机构名等命名实体，为后续的实体关系分析提供基础。政治话语中涉及的命名实体较多，命名实体识别的准确性对话语数据库的质量影响很大。

5）语义角色标注：对句子中的每个词语进行语义角色标注，即确定其在句子中的语义角色，如主体、客体等。语义角色标注可以帮助研究者更准确地理解政治话语的语义结构。

6）情感分析：识别文本中的情感极性，如正面、负面或中性，以及情感强度。政治话语中经常涉及政治态度、情感表达等，情感分析可以帮助研究者了解政治话语的情感取向和情感表达方式。

7）文本分类：将文本按照某个分类体系进行分类，如按照主题、作者、时间等分类。文本分类可以帮助研究者快速地进行大规模的文本分类和管理。

8）文本聚类：将文本按照相似度聚类，从而实现文本分类和管理的目的。文本聚类可以帮助研究者快速地对大量的文本数据进行分类和管理，也可以帮助研究者发现潜在的文本主题和话题。

9）关键词提取：自动提取文本中的关键词或短语，为后续的文本分析和可视化提供基础。关键词提取可以帮助研究者抓住政治话语中的重点和主题，进而深入分析和理解政治话语的内涵。

10）文本摘要：自动提取文本中的关键信息，生成简洁而准确的摘要，为快速了解政治话语提供帮助。文本摘要可以帮助研究者快速地了解政治话语的主要内容和意义。

11）信息抽取：自动从文本中提取结构化的信息，如事件、实体关系、时间等，为后续的信息检索和可视化提供基础。信息抽取可以帮助研究者更快速地发现和理解政治话语中的重要信息和关系。

12）文本生成：通过机器学习和自然语言生成技术，自动生成符合特定主题和语言风格的政治话语。文本生成可以帮助研究者快速地生成大量的政治话语数据，也可以为政治决策和媒体传播提供帮助。

以上这些 NLP 处理能力可以被应用到政治话语数据库的建设和分析中，不同的处理能力可以针对不同的研究问题和目的进行选择和组合。同时，这些处理能力也需要根据不同的语言环境和文化背景进行优化和调整，以提高政治话语数据库的质量和可用性。

以下是一些常用的自然语言处理软件模块和平台，它们可以用于政治话语数据库的建设和分析：

NLTK（Natural Language Toolkit）：Python 自然语言处理库，提供众多的自然语言处理工具和数据集。可用于文本清洗、分词、词性标注、命名实体识别、句法分析等。

Stanford CoreNLP：斯坦福大学开发的自然语言处理工具，提供词法分析、命名实体识别、句法分析、语义角色标注等多项功能。可以用于多种处理任务。

Gensim：Python 库，用于主题建模、文本相似度计算等任务。可以用于政治话语数据库的主题建模和文本相似度计算。

spaCy：Python 自然语言处理库，提供了分词、命名实体识别、词性标注、句法分析等功能。可以用于政治话语数据库的文本处理和分析。

OpenNLP：由 Apache 开发的自然语言处理工具，提供了词法分析、命名实体识别、句法分析等多项功能。可以用于政治话语数据库的多种处理任务。

Word2Vec：基于神经网络的语言模型，用于词向量的训练和表示。可以用于政治话语数据库的词向量表示和文本分类。

TensorFlow：开源的机器学习框架，可以用于文本分类、情感分析等任务。可以用于政治话语数据库的多种处理任务。

这些自然语言处理软件模块和平台都可在其官方网站上下载和安装。同时，也有相应的文档和使用示例，可帮助研究者快速地上手和使用。此外，还有一些在线平台和工具，如 Google Colab、Azure Notebooks 等，可以帮助研究者在云端使用自然语言处理功能。

除了上述提到的自然语言处理软件模块和平台，还有一些工具可以用于政治话语数据库的建设和分析：

GitHub：代码托管平台，可以用于政治话语数据库的代码管理和分享。

Apache Solr：基于 Lucene 的全文搜索引擎，可用于政治话语数据库的全文检索和查询。

ElasticSearch：基于 Lucene 的分布式搜索引擎，可用于政治话语数据库的全文检索和查询。

MongoDB：NoSQL 数据库，可以用于政治话语数据库的存储和管理。

这些工具都可以通过它们的官方网站或 GitHub 获取使用。其中，GitHub 和 Apache Solr 等工具是免费的开源软件，ElasticSearch 和 MongoDB 等工具则有免费和商业版可供选择。

需注意的是，这些自然语言处理软件模块、平台和工具，应根据具体任务和研究需求选择使用且需具备一定的编程技能和经验，研究者应该有一定的计算机科学基础和编程能力。

俄罗斯在自然语言处理领域也拥有丰富的研究和开发经验，以下是一些俄罗斯的 NLP 软件工具及其特点、文本分析工具以及获取方式和地址：

Yandex NLP：俄罗斯最大的搜索引擎和互联网公司之一，其自然语言处理技术领先于俄罗斯国内其他公司和机构。Yandex NLP 提供了包括分词、词性标注、命名实体识别、句法分析、情感分析等功能，支持俄语、英语、乌克兰语等多种语言。Yandex NLP 的 API 开放给开发者使用，详细信息和获取方式可在 https://yandex.com/dev/nlp/上查看。

ABBYY Lingvo：一款俄语-英语-德语等多语言在线词典和翻译软件，提供了包括词典查询、单词发音、文本翻译等功能。ABBYY Lingvo 也提供了一些自然语言处理功能，如分词、词性标注、语法分析等。ABBYY Lingvo 的详细信息和获取方式可在 https://www.lingvolive.com/ru-ru/about/lingvo_online 上查看。

SynTagRus：由俄罗斯语言学家和计算机科学家共同研发的一个俄语语料库，包含了 200 万个句子和 6800 万个词语，涵盖了多种文本类型和领域，如新闻、科技、文学等。目前已经成为俄语自然语言处理领域的重要资源，可在 https://github.com/UniversalDependencies/UD_Russian-SynTagRus 上免费下载和使用。

Texterra：一款基于人工智能和自然语言处理技术的文本分析工具，提供了包括文本分类、情感分析、关键词提取、实体识别等功能。Texterra 支持多种语言处理，包括俄语、英语、中文等。Texterra 可通过其官方网站 http://texterra.ru/免费试用，也可以选择付费购买其 API 和定制化服务。

Krasivo：一款基于俄罗斯语言的文本生成工具，可用于生成俄语口头表达和书面语，如商业信函、新闻稿、论文摘要等。Krasivo 使用神经网络技术进行自然语言生成，能够生成高质量的俄语文本。Krasivo 可在其官方网站 https://krasivo.ai/上免费试用。

OpenCorpora：免费开源的俄语语言资源，包括了俄语词典、语料库、词性标注、句法分析等资源。OpenCorpora 可以帮助开发者进行俄语自然语言处理的研究和开发。OpenCorpora 的详细信息和获取方式可在 https://opencorpora.org/上查看。

RusVectōrēs：俄语自然语言处理和机器学习研究团队，在俄语文本处理和自然语言生成方面拥有丰富的经验和技术。RusVectōrēs 提供了多种俄语文本处理工具和数据集，包括俄语词向量、词性标注、情感分析等。RusVectōrēs 的详细信息和获取方式可在其官方网站 https://rusvectores.org/en/上查看。

NLTK：虽然 NLTK（Natural Language Toolkit）是由美国的计算机科学家开发的一个自然语言处理工具包，但它也支持多种语言处理，包括俄语。NLTK 提供了分词、词性标注、

命名实体识别、句法分析等功能，同时也支持文本分类、情感分析等应用。NLTK 可以在其官方网站 https://www.nltk.org/上免费下载和使用。

以上是一些俄罗斯 NLP 软件工具和文本分析工具的介绍，这些工具都在俄罗斯自然语言处理领域拥有广泛应用和影响，有助于促进俄语自然语言处理技术的研究和应用。

Gensim：用于向量空间建模和主题建模的 Python 库，可帮助开发者进行自然语言处理任务。Gensim 支持多种语言，包括俄语，可进行文本处理、词向量建模、主题建模等任务。Gensim 还提供了多种模型，例如 Word2Vec、Doc2Vec 等，可用于文本相似度计算、主题分类等应用。Gensim 可以在其官方网站 https://radimrehurek.com/gensim/上免费下载和使用。

Pymorphy2：Python 库，用于俄语自然语言处理，包括词形变化、词性标注、单词屈折、单词转换等。Pymorphy2 支持多种俄语变体和方言，并且可以通过使用数据文件进行自定义。Pymorphy2 可以在其 GitHub 地址 https://github.com/kmike/pymorphy2 上获取。

DeepPavlov：开源对话系统框架，提供自然语言处理和机器学习的多种组件。DeepPavlov 可以用于自然语言理解、机器翻译、机器阅读理解等任务。DeepPavlov 使用 TensorFlow 作为其核心，同时支持多种语言，包括俄语。可在其官方网站 https://deeppavlov.ai/上获取。

Corpy：基于 Python 的自然语言处理工具，用于俄语文本处理和分析。Corpy 提供了词形还原、词性标注、句法分析等功能，同时还支持自定义规则和模型。Corpy 可在其 GitHub 地址 https://github.com/natasha/corpy 上获取。

以上工具和文本分析工具都是非常流行的俄语 NLP 工具，在俄罗斯自然语言处理领域有着广泛的应用。这些工具的开源性质使得开发者可以自由地使用、修改和分发这些工具，有助于促进俄语自然语言处理技术的发展。

综上所述，课题以数字人文研究和新文科建设为背景，以数字中国建设和加强国际交往语言能力建设为需求导向，建设东北亚政治话语数据库，并在数据驱动下进行多任务的理论研究和应用实践，具有重大的理论意义和广阔的应用前景。在 NLP 能力的支撑下可取得多领域、多学科、多目标的研究成果。

参考文献

[1]安　妮，伯迪克等. 数字人文[M]. 马林青等译. 北京：中国人民大学出版社，2017.

[2]白文昌. 俄语教学与研究[C]. 哈尔滨：黑龙江大学出版社，2017.

[3]陈　虹. 俄语语料库的标注[J]. 中国俄语教学，2012(2).

[4]丛亚平等. 实用商贸俄语[M]. 北京：外语教学与研究出版社，2010.

[5]冯志伟. 机器翻译与人工智能的平行发展[J]. 外国语，2018(6).

[6]傅兴尚. 现代俄语事格语法[M]. 北京：军事谊文出版社，1999.

[7]傅兴尚等. 俄罗斯计算语言学与机器翻译[M]. 北京：语文出版社，2009.

[8]何安平. 语料库语言学与英语教学[M]. 北京：外语教学与研究出版社，2004.

[9]何安平. 运用电脑语料库改革英语课堂教学[J]. 华南师范大学学报，1998(12).

[10]柯　飞. 汉语把字句特点、分布及英译研究[J]. 外语与外语教学，2003(12).

[11]李绍哲. 俄语语料库和基于语料库的语法研究[D]. 黑龙江大学博士学位论文，2012.

[12]李彦宏等. 智能革命——迎接人工智能时代的社会、经济与文化变革[M]. 北京：中信出版集团股份有限公司，2017.

[13]刘　淼，邵　青. 俄汉文学翻译语料库的创建——基于契诃夫小说平行语料库的设计与建构[J]. 外语学刊，2016(1).

[13]罗　颖. 利用语料库分析中学英语课堂提问技巧[J]. 国外外语教学，1999(4).

[14]王　臻. 俄语语料库语言学研究现状与瞻望[J]. 中国俄语教学，2007(2).

[15]王克非等. 双语对应语料库研制与应用[M]. 北京：外语教学与研究出版社，2004.
[16]肖依虎，潘翠琼. 语料库在语言测试中的应用[J]. 外语教学，2002(6).
[17]许汉成. 俄语语料库的新发展[J].中国俄语教学，2005(1).
[18]亚里士多德. 诗学[M]. 北京：商务印书馆，2010.
[19]原　伟. 俄汉新闻可比语料库的构建、评估及应用展望[J]. 解放军外国语学院学报，2017(6).
[20]张俊萍，金　婷. HSK（初中等）语法提高教程[M]. 北京：北京广播学院出版社，2004.
[21]张禄彭，张超静. 自建语料库在俄语教学研究中的应用[J]. 中国俄语教学，2012(3).
[22]朱珊珊，原　伟. 面向俄文情感分析的新闻评论语料库建设与应用[J]. 外语学刊，2020(1).

An Overview of The Construction and Application of Political Discourse Database in Northeast Asian Countries

Fu Xing-shang

(China Northeast Asian Languages Research Center, Dalian University of Foreign Languages, Dalian 116044; Center for Russian Language Literature and Culture Studies of Heilongjiang University, Harbin 150080, China)

Abstract: Political discourse database refers to an electronic database that collects and organizes political discourse, including source, time, author, content, linguistic features, cultural background and other aspects of information. With the help of a hierarchical political discourse database, researchers can conduct large-scale linguistic and cultural analysis, as well as critical discourse analysis, so as to reveal the social, cultural, and power structure information behind. In this sense, the construction and application of political discourse database in Northeast Asian countries accord with the strategy and practical needs of cooperation with Northeast Asian countries. As the core product, structured database and management platform of political discourse in Northeast Asian Countries, is not only a national-level language resource infrastructure, but can also develop functions such as language query, knowledge query, comparison and accurate translation between Chinese and foreign languages, dynamic monitoring, analysis of public opinion on major political events, prediction of political trends, and cultural-pragmatic calculation for political purposes. In addition, through the data sharing mechanism, it provides data resource and computational model for the multidimensional study of political discourse. This paper makes an overview study from four aspects: 1) the background and overall objective; 2) the theoretical basis and significance; 3) the main content; 4) the development of natural language processing capability.

Keywords: Northeast Asia; political discourse; database; NLP; database application

作者简介：傅兴尚（1966—），河北青龙人。北京语言大学语言资源高精尖创新中心特聘研究员，大连外国语大学启航学者，黑龙江大学俄罗斯语言文学与文化研究中心兼职研究员，教育部院校规划发展中心专家，教授，博士生导师。研究方向：语言学、计算语言学、外国语言学。

收稿日期：2023-03-10　　[**责任编辑：靳铭吉**]