广西科学院学报
Journal of Guangxi Academy of Sciences
2018,34(1):18～26
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-03-20 DOI:10.13657/j.cnki.gxkxyxb.20180320.002
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180320.1021.004.html

壮、蒙古、维、哈、柯、朝语信息处理研究进展*

Research Progress of Information Processing in Zhuang, Mongolian, Uighur, Kazak, Kirgiz and Korean

刘连芳[1,2],海银花[3],那顺乌日图[3],黄家裕[1,2],吐尔根·依布拉音[4],玄龙云[5]
LIU Lianfang[1,2], HAI Yinhua[3], Nasun-urt[3], HUANG Jiayu[1,2],
Tuergen·Yibulayin[4], XUAN Longyun[5]

(1.广西达译商务服务有限责任公司,广西南宁 530007;2.南宁市平方软件新技术有限责任公司,广西南宁 530007;3.内蒙古大学蒙古学学院,内蒙古呼和浩特 010021;4.新疆大学信息科学与工程学院,新疆乌鲁木齐 830046;5.新疆多语种信息技术实验室,新疆乌鲁木齐 830046;6.中国朝鲜语信息处理学会,吉林延吉 133002)
(1. Guangxi Daring E－commerce Co., Ltd., Nanning, Guangxi, 530007, China; 2. Nanning Pingsoft New Software Technology Co., Ltd., Nanning, Guangxi, 53007, China; 3. School of Mongolian Studies, Inner Mongolia University, Hohhot, Inner Mongolia, 010021, China; 4. School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, 830046, China; 5. Xinjiang Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang, 830046, China; 6. Chinese Korean Information Processing Society, Yanji, Jilin, 133002, China)

摘要:少数民族语言文字处理是中国语言文字信息处理的重要组成部分。自20世纪80年代以来,少数民族语言文字处理在各民族科研、产业工作者的共同努力下,在操作系统、输入输出、编辑排版、标准化、语言资源建设、机器翻译、软件平台、人才培养等各个方面取得了长足的进展。本文综述壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜6个少数民族的语言文字信息处理历史、现状及存在问题,并对其未来发展方向进行展望。

关键词:壮文 蒙古文 维哈柯文 朝鲜文 信息处理
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1002-7378(2018)01-0018-09

Abstract: Minority language and word processing is an important part of language information processing in China. Since the 1980s, with the joint efforts of scientific research and industrial workers of various nationalities, great progress has been made in the areas of operating system, input/output, editing and layout, standardization, construction of language resources, machine translation, software platform and personnel training. This article summarizes the history, status quo, and existing problems of language information processing in Zhuang, Mongolian, Uyghur, Kazak, Kirgiz and Korean, and forecasts the future development direction.

Key words: Zhuang, Mongolian, Uighur Kazak and Kirgiz, Korean, information processing

0 引言

从20世纪80年代开始,我国的中文信息处理

收稿日期:2018-02-10
作者简介:刘连芳(1946—),女,研究员,主要从事自然语言处理、语言资源建设研究,E-mail:lianfangl@yeah.net。
*国家社会科学基金重点项目(10AYY006),国家自然科学基金项目(61063026)和国家电子基金项目(工信部财[2009]453号,工信部财建[2008]329号,工信部运[2008]97号)资助。

进入了快速发展阶段,中文信息处理效率得到极大提高。中国有56个民族,使用着数十种文字和近百种语言。因此,中文信息处理所涉及的语言文字不仅包括简体汉字、繁体汉字,也包括藏文、蒙古文、维吾尔文、壮文、朝鲜文、彝文等大量民族语言文字,少数民族语言文字处理是中国语言文字信息处理的重要组成部分。随着信息化时代的到来,少数民族语言文字信息处理技术将成为我国少数民族文化传承的一种重要手段,对少数民族语言文字信息处理的现状有一个清晰的认识才能够更好地把握其未来发展方向。

1985年10月中国中文信息学会成立了民族语言文字信息专业委员会,着重开展我国少数民族语言文字信息处理及其有关工作。自20世纪80年代以来,少数民族语言文字处理在各民族科研、产业工作者的共同努力下,在操作系统、输入输出、编辑排版、标准化、语言资源建设、机器翻译、软件平台、人才培养等各个方面取得了长足的进展。本文对壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜6个少数民族的语言文字信息处理历史、现状及存在问题进行总结概述,并对其未来发展方向进行展望,以期为我国少数民族语言文字的信息处理提供参考和借鉴。

1 少数民族语言文字信息处理历史与现状

1.1 壮文信息处理

壮语是汉藏语系壮侗语族壮傣语支的一种语言。目前存世的壮族文字有古壮文和现代壮文。对古壮文的信息处理研究与开发的主要目的是抢救民族文化遗产,而对现代壮文信息处理的研究与开发主要在于现实应用,二者均具有重要的研究价值和现实意义。壮文有悠久的历史,壮文作为一种文化符号对民族文化的传承起到了积极的作用。

壮文信息处理技术和汉文信息处理技术存在许多方面的差别。壮文信息处理技术包含两个方面的内容:一是对壮文字的处理,二是对壮语言的处理。壮文字的处理涉及到所使用的操作系统、文字的输入输出和编辑等方面,而语言的处理涉及壮语语音识别、信息检索、壮文翻译等方面。语言和文字之间是紧密相联的,壮文信息处理技术离不开对文字和语言的依托。

1.1.1 壮文字信息处理技术

近年来,壮文字信息处理技术研究领域取得了一些成果。从磁盘操作系统(Disk Operating System,DOS)时代开始,古壮文的信息处理技术研发就存在困难,这一时期的主要研究成果是广西计算中心研发的"DOS古壮文编辑排版系统",该系统初步解决了壮族古籍数字化出版难题[1]。随着计算机技术的发展,壮文字信息处理技术水平得到进一步提高,主要成果有南宁市平方软件新技术有限责任公司研制的"Windows下的古壮文处理系统",该系统持可视化的编辑排版[2];广西骆越文化研究会会员覃志强开发的"古壮文输入法"和"新壮文输入法"简化了复杂的输入过程,降低了录入的差错率。在互联网广泛应用的今天,壮文处理技术与互联网融合发展,南宁市平方软件新技术有限责任公司开发的"古壮字收录及字典管理",可通过互联网面向社会收录新发现的古壮字,并对其进行快速查重和字频统计。此外,一些个人开发的在线壮汉词典也在网络上流行。壮文信息处理技术较以往带来更大的社会价值。

1.1.2 壮语言信息处理技术

壮语言信息处理技术在字典和翻译方面也取得了一些成果。由于古壮字没有形成统一规范,异体字多,笔划过繁,没有被行政公文和正规教育所采用,当今能识读古壮字的人越来越少。字典软件和翻译软件满足了人们解读古文献,传唱山歌的需要。

1.1.2.1 古壮字的电子字典

"古壮字释义电子字典"由南宁市平方软件新技术有限责任公司根据广西壮族自治区少数民族古籍整理出版规划领导小组办公室编撰的《古壮字字典》研发而成,可检索古壮字的发音、释义、例句等,并可通过古壮字检索出其汉语释义。

1.1.2.2 现代壮文的电子词典

"壮文电子词典及辅助翻译软件"研发由中国民族语文翻译中心科研处和壮语文室合作完成[3],该软件具有5个重要功能:①壮文联想输入功能,该功能不仅能实现壮文单词联想,还能实现壮文词组联想和句子联想;②壮文标点符号输入功能,使用者不必在中文状态或英文状态下转换,即可输入需要的标点符号;③单词查找功能;④词组翻译功能;⑤长句的简化录入功能。

"在线双向汉壮词典"是中央民族大学壮侗学研究所、广西壮学学会和广西骆越文化研究会支持的"壮族在线"提供的在线双向汉壮词典,共收录词条25 986条,这些词条基本来源于stoneman、honghlaj等贝依制作的Sawloih CuenghGun电子版和一些新加入的方言词汇,目前该词典已在互联网上使用[4]。该词典除了壮汉双向翻译外,还提供壮语声

母表、壮语韵母表、壮语音节表及壮语拼写规则等，方便广大爱好者学习壮文。

“壮汉英电子词典”由南宁市平方软件新技术有限责任公司研发，可检索壮语词的发音、解释、汉语对应词、英文对应词、例句等，支持壮汉、汉壮、壮英、英壮4种双向翻译，支持单机版及网络在线版。

1.1.2.3 壮文语料库及机器翻译

中国民族语文翻译中心研发的“壮文电子词典及辅助翻译软件”，将《壮汉词汇》和《汉壮词汇》作为语料库中基本词汇的来源[3]。南宁市平方软件新技术有限责任公司研发的“基于短语的汉壮统计机器翻译”，采用广西少数民族语言工作委员会编撰的《壮汉英词典》作为语料库来源。

1.2 蒙古文信息处理

蒙古语属于阿尔泰语系，是一种粘着性的、词汇形态变化特别丰富的语言，这是蒙古语最大的特点。文字方面，蒙古语是纯粹的拼音文字，从左到右竖写。蒙古文信息处理最大的难点是由蒙古语丰富的形态变化形式和竖写形式造成的。另一方面，蒙古语语言资源非常丰富，从时间跨度上讲，蒙古语贯穿13世纪至今的800余年的历史；从地域上讲，蒙古语横跨欧亚大陆，几乎遍布整个地球，目前世界上有近800万蒙古族人口，主要聚居在中国、蒙古、俄罗斯等国家，并散居在世界各地。世界上现行的蒙古文有5种(传统蒙古文、托忒蒙古文、蒙古国西里尔文、布里亚特西里尔文和卡尔梅克西里尔文)，加上历史上曾经用过的文字，蒙古民族曾经使用过的文字已超过十余种之多。蒙古语作为中国、蒙古国、俄罗斯等国家的官方语言，蒙古语言资源的保护、开发和利用在当今“一带一路”战略中是不可忽视的重要环节。

1.2.1 蒙古文信息处理成果

从研究开发历史视角来讲，从20世纪80年代起我国蒙古文信息处理工作在数据资源的采集、知识资源的挖掘和技术资源的开发方面取得的成绩包括：

(1)成功制定蒙古文国际编码标准(ISO-2000)和国家标准等诸多标准规范。

(2)成功研发MongxeGal输入法(蒙科立公司)和Microsoft蒙古文输入法等输入法。

(3)成功构建“蒙古秘史”(汉蒙对照版)、回鹘体蒙古文、托忒文、八思巴文等文献语料库，100万、500万、1 000万词蒙古语单语语料库和汉蒙、蒙汉等各类双语平行语料库以及口语语料库等资源库。

(4)成功构建蒙古语语法、语义信息词典、多义词词典、类语词典、熟语等知识库。

(5)成功研发汉-蒙机器翻译系统、汉蒙电子词典(V.2.0)、方正蒙古文排版系统、传统蒙古文-西里尔蒙古文转换系统、面向蒙古语语音特点的参数自动采集软件、蒙古语语音合成软件、蒙古文农业专家系统、“蒙古语双文少儿词典”“汉蒙英日大词典在线服务平台”“数字农业专家系统在线服务平台”“内蒙古省地县三级政府蒙古文网站群管理系统”“天地图蒙古文在线地图”“汉蒙俄英多文种标准化共享服务平台”等Web式系统和好乐宝、草原雄鹰、蒙古文化网、Monghegal、Ulaaci、Ehshig等100多家蒙古文网站。这一系列工程的研发，一直占领这一领域的制高点。

1.2.2 整体发展现状

(1)创新团队：内蒙古大学蒙古学学院、内蒙古大学计算机学院等若干机构科研团队为省部级创新团队(内蒙古自治区“草原英才”工程高校创新团队)，正在争取提升为国家级创新团队。

(2)科研项目：近5年以来，蒙古文信息处理科研团队承担的省部级以上科研项目的数量不断增加，其中国家级项目居多，甚至还有过国家社科重大项目、教育部人文社科重大项目。

(3)学术论著：据初步统计，蒙古文信息处理研究领域已出版《蒙古文编码》《蒙古文信息处理理论与实践》《面向信息处理的动词短语结构规则研究》和《面向信息处理的蒙古名词语义研究》等30多部专著。

(4)奖项：蒙古文信息处理研究领域学术成果和教学成果曾获得内蒙古自治区教学成果一等奖、北京市科技进步二等奖、钱伟长中文信息处理技术奖一、三等奖，内蒙古自治区哲学社会科学优秀成果政府奖二、三等奖、内蒙古自治区优秀博士论文奖等诸多奖项。

(5)科研平台建设：近年来蒙古文信息处理研究领域连续产生内蒙古自治区协同创新中心、“机器翻译联合实验室”“民族语言资源产业化基地”“计算语言学联合实验室”、应用语言学实验室等科研平台，实现由政府、高校、企业相联合的软件研发基地，并同步实现了人才联合培养模式。

(6)人才培养：民族地区科研人员先后到蒙古国、日本、美国、芬兰、韩国、俄罗斯、新加坡、台湾、捷克、爱尔兰等国家和地区进行学术交流，拓展了学术视野；以内蒙古大学蒙古学学院蒙古文信息处理研

究方向为例,在过去的 10 年里已培养约 40 名博士研究生和 140 多名硕士研究生。

1.3 维哈柯文信息处理

维吾尔语、哈萨克语、柯尔克孜语属阿尔泰语系突厥语族,在形态结构上属黏着语类型。随着信息技术的不断发展,人类的语言文字只有信息化才有存续、发展的生命力,因此,维哈柯文信息处理工作直接关系着维吾尔文、哈萨克文、柯尔克孜文的命运。

30 多年来,维哈柯文信息处理在操作系统、信息技术标准、语言信息处理及综合应用等方面取得了不少成绩[5]。

1.3.1 操作系统

1984 年,新疆大学刘诚信、袁保社、吐尔根·依布拉音等开发了支持维、哈文的 UHDOS1.1 操作系统。1985 年 5 月新疆大学吴宗尧、吾守尔·斯拉木等相继研发成功维吾尔文、哈萨克文、柯尔克孜文微机操作系统 UHKDOS 3.0、UHKDOS 4.0、UHKDOS 5.0、UHKDOS 6.0 及 UHKDOS 7.0,实现了维、汉、英文混合编辑。1992 年,新疆大学吾守尔·斯拉木、吐尔根·依布拉音等开始进行支持维哈柯文的 Windows 操作系统的开发,相继开发出支持维哈柯文的 Windows 3.1、Windows 95、Windows 98 操作系统。2001 年开始,新疆大学开发出外挂维哈柯文的 Windows 2000 及 Windows XP 操作系统。2003 年,新疆大学首次开发出维哈柯多语种 Linux 操作系统。2005 年国家 863 重大专项"民族语言版本 Linux 操作系统及办公套件研发"项目取得成功,维哈柯文 Linux 操作系统达到了汉、英文同等的技术水平。同年起,新疆大学还先后开发了基于 QT 的维哈柯多文种嵌入式操作系统、基于 Linux 的嵌入式设备用维哈柯文操作系统、支持维哈柯文的 Windows CE 以及支持维哈柯文的 Android 嵌入式操作系统。2008 年,新疆大学等单位研发了基于 Android 的维吾尔文输入法。2010 年,新疆大学等单位进行 Windows 7 维哈柯文化研究与开发。

1.3.2 信息处理标准化研究

吾守尔·斯拉木等起草制定了首个信息处理交换用维文、哈文三项国家标准 GB/T 12510—1990(代码标准、点阵字型数据标准、键盘布局标准)并发布实施。随着信息技术的发展,同时也为了与国际标准接轨,吾守尔·斯拉木等对《信息技术用维、哈、柯文编码字符集基本集》进行了的修订,形成国家标准 GB 21669—2008。之后,新疆维吾尔族自治区又先后制定了《古维文编码字符集》等国际标准,以及《信息交换维哈柯文编码字符集》《信息交换用维哈柯文(曲线)字型白体黑体》《信息交换用维哈柯文点阵字型》《信息交换用维吾尔文、哈萨克文、柯尔克孜文字体字形》及《信息技术 维吾尔文常用术语》等国家标准。

1.3.3 自然语言处理技术研究

1.3.3.1 语言资源建设

新疆师范大学玉素甫等于 2002 年构建了 800 万词次的维吾尔文语料库。新疆大学吐尔根等自 2002 年起开展维哈柯文语料库建设工作,最终建成 123 万词次的维吾尔语词法标注的语料库和 3 000 句的句法标注的语料库,建成 30 万维汉句对、15 万哈汉句对及 10 万柯汉句对的双语语料库。新疆大学古丽拉·阿东别克等构建了现代哈萨克语词级标注语料库[6]。

1.3.3.2 词法分析与句法分析

1997 年新疆师范大学玉素甫等对维吾尔语词干和词性标注、句法分析等开展初步研究。2004 年中央民族大学力提甫·托乎提对计算机词干提取过程中遇到的元音和辅音的弱化、增音、脱落等进行系统地描述。之后,多位学者先后对维吾尔文(语)开展了下述研究:基于大规模语料库的字母统计,字母的熵计算,音节自提取算法,词根库建设,名词形态结构研究及规则总结,基于词典的词性标注方法,基于词性标注的文字校对方法,基于 N 元语法的词性标注模型,词频统计,基于最小编辑距离的候选词产生算法,基于规则的元音弱化处理算法,基于规则的句子边界识别算法,新疆师范大学信息处理用维语词汇标注标记集的确定,基于规则的对偶词识别,汉维翻译中的人名、维吾尔语缩写词识别算法,基于隐马尔科夫(HMM)模型的词性标注模型。2009 年新疆大学艾山·吾买尔对维吾尔文从生文本至严格按照规范标注的语料库建设、词法分析、浅层句法的各个环节展开深入的研究。

2006 年以来,新疆大学古丽拉、达吾勒等对哈萨克语开展了如下研究:词频统计,文本分类,基本名词短语识别,词性自动标注及标注规范制定,哈萨克语人名识别词法分类,哈萨克阿拉伯文与哈萨克斯拉夫文文本转换等。

1.3.3.3 框架语义知识库研究

2007 年以来新疆大学阿里甫·库尔班等对维吾尔语框架语义知识库工程开展研究,探索了词一

级的知识库的构建方法及技术路线。目前已就维吾尔语名词、形容词、动词、量词和副词等4 252个词元构建了405个框架，并制定了以框架为单位的分类描述规则、词语分类体系和相应标记集。

1.3.3.4 语言动态监测与研究

2009年中央民族大学与新疆师范大学联合共建“国家语言资源监测与研究中心少数民族语言分中心维吾尔语文研究基地”，2010中央民族大学与新疆大学又联合共建“国家语言资源监测与研究中心少数民族语言分中心哈萨克和柯尔克孜语研究基地”，上述两个基地对维吾尔语、哈萨克语、柯尔克孜语的主要媒体进行动态监测与研究。

1.3.4 综合应用研究

1988年新疆大学袁保社等研制了四通2400，2401系列维哈柯文电子打字机。1989年新疆大学等单位开发了维吾尔文、哈萨克文、柯尔克孜文与汉英文全兼容的“博格达书报排版系统”。1990年中国计算机软件与技术服务总公司等单位推出了能排版蒙藏维哈柯文的北大方正多文种文书报版系统。之后，新疆大学协助北大方正、潍坊华光开发了维哈柯文方正排版系统(1991)、潍坊华光排版系统(1992)、三立书版排版系统(1994)、锡伯文、满文文字处理和轻印刷系统(1996)、“新疆2000”多文种图文排版系统(2000)等。

新疆理化所协助永中软件公司开发了维哈柯文永中Office办公套件；新疆大学开发了维哈柯文OpenOffice办公套件，协助上海中标公司开发了维哈柯文中标Office办公套件。

新疆大学吐尔根·依布拉音等自2003年起研发“基于Unicode的多语种-多向-多媒体大型电子词典资源开发系统(3MLDMDRP)”及“基于Unicode的碧黎库特英汉维电子词典软件(ECU Dictionary)”，乌鲁木齐市安卡维文软件开发有限公司研发了“维软大词典”系列软件，乌鲁木齐市一帆电子有限公司研发了“汉-维哈柯文一帆掌上电子词典”。

1996年新疆大学王世杰提等开始开展基于规则的汉维机器翻译研究，2005年起新疆大学哈力木拉提等开展了基于词典的计算机辅助翻译系统的研究，2009年新疆大学吐尔根等与新疆信息产业有限公司开展了汉维哈柯计算机辅助翻译软件的研发。2010年中国科学院计算技术研究所刘群等与新疆大学吐尔根等合作推出基于统计的维汉机器翻译系统。新疆理化所周俊林等自2009年以来开展基于短语的汉维/维汉统计机器翻译研究[7]。

2004年，新疆大学哈力木拉提和清华大学丁晓青完成了首款支持维吾尔文、哈萨克文、柯尔克孜文以及阿拉伯文的印刷文档识别系统的研发。新疆师范大学的玉苏甫等及新疆大学的哈力木拉提等对维哈柯文文字手写识别以及联机手写进行了探索性研究[8]。

20世纪90年代初，新疆大学吾守尔·斯拉木成功研制了联想式维吾尔语音识别系统。20世纪90年代后期，新疆师范大学王昆仑等开展了基于音节的非特定人语音识别研究。2000年后，新疆大学的吾守尔等与中国社会科学院民族学与人类学研究所鲍怀翘等构建了“维吾尔语语音声学参数库”，并成功研发了维吾尔语音合成软件[9-10]。2017年科大讯飞股份有限公司发布了维汉语音翻译终端设备。

1.4 朝鲜文信息处理

中国朝鲜文是随着朝鲜民族移居中国大地时起发展起来的少数民族语言，是源自朝鲜与韩国，又与中国当地的多重文化相融合而形成的较为独特的语言，与朝鲜、韩国语言既有相同又有区别，是中国少数民族语言文化中非常有特点的语种之一。

1.4.1 标准的制定

20世纪70年代，根据国务院的决定东三省成立了朝鲜语文工作协调小组(简称“三协”)，以统一管理中国朝鲜语文工作。在“三协”的指导下，朝鲜语规范委员会先后制定了朝鲜文信息处理领域相关的规范原则和朝鲜语规范统一方案，完成了国家标准《信息交换用朝鲜文字编码字符集》(GB 12052—1989)的制定，组建了中国朝鲜语信息处理学会，并组织延边电子信息中心、延边大学等单位和大专院校的学者、专家完成了多个朝鲜语信息处理系统的研发。1996年，中国朝鲜语术语标准化工作委员会成立，并完成了《朝鲜语术语数据库的一般原则与方法》的编写工作，制定了《朝鲜语术语标准化工作原则与方法》，研制开发出朝鲜文电脑激光排版印刷系统。

全国信息技术标准化技术委员会从2004年开始先后成立了蒙维藏彝傣壮朝文信息技术工作组，大力推进民文信息化建设。

朝鲜文是朝鲜、韩国和中国三国通用语言，除了共同使用ISO 10646韩文(朝鲜文)字符集以外，并没有其他统一的国际标准。2013年朝鲜文信息技术国家标准工作组成立，2015年该工作组完成了2项国家标准的制定，即《信息技术 朝鲜文通用键盘

字母数字区的布局》和《信息技术 基于数字键盘的朝鲜文字母布局》,并于 2017 年 11 月正式发布。2017 年该工作组又完成了《朝鲜文信息技术术语和定义》和《朝鲜文编码字符 24 点阵字型》2 项吉林省地方标准[11]。此外,该工作组还带领技术团队研发了基于 Windows、Linux、Android、IOS 平台的 4 种朝鲜文输入法和 10 种朝鲜文字型。

1.4.2 朝鲜语言文字信息处理

为加速我国朝鲜语言文字规范化、标准化、信息化进程,进一步促进朝鲜语信息技术国际标准的制定,国家民族事务委员会于 2014 年 4 月在延边大学正式成立"中国朝鲜语言文字信息化基地"(简称"朝鲜文基地")。近年来,该基地在朝鲜语言文字信息处理方面开展了多项研究,取得了一些可喜的成果。

1.4.2.1 朝鲜语言文字规范化建设

2016 年,朝鲜文基地全面调查中国朝鲜族新闻、广播、出版等媒体和中小学朝鲜语使用和教学情况,协同中国朝鲜语规范委员会、中国朝鲜语学会修订了 2016 年版《中国朝鲜语规范集》,重新审定及发布了朝鲜语新名词术语及中小学教材中的朝鲜语术语,继续深化研究了中国朝鲜语罗马字标记法原则与细则。

1.4.2.2 朝鲜语言文字字符集及其平台建设

在吉林省科技发展计划项目(20140101186JC)、国家语委 2015 年度科研项目(教语信司函[2015]21 号)的支持下,朝鲜文基地研究多语种文本图像中的文字语种辨识方法[12],针对汉字、朝鲜文字和英文单词混合的文本图像,提出了基于主成分分析技术以文字为单位进行文种辨识的方法。该方法在没有分割错误的情况下,能获得 99.78%的识别准确率,有效地解决了在汉、朝、英 3 种文字混合构成的文档图像中的文种辨识问题。

在吉林省科技厅自然科学基金项目(20140101225JC)的支持下,朝鲜文基地提出了一种基于基音频率特征的中国朝鲜族语言、韩国朝鲜语和朝鲜朝鲜语方言的自动辨识方法[13]。研究结果表明,该方言辨识方法比传统的移位差分倒谱系数特征方法识别率高,可以有效解决中国朝鲜族语言、韩国朝鲜语和朝鲜朝鲜语的方言辨识问题。

朝鲜文基地应用基于图像与音频的朝鲜语自动辨识方法,开发了中韩科技信息加工综合平台。此外,通过对中韩科技文献信息采集与智能处理的研究,朝鲜文基地不仅开发了科技文献采集系统,同时还构建了丰富的科技术语语料库。

1.4.2.3 朝鲜语文本资源库建设

中国朝鲜语文本资源库包括:中国朝鲜语文本语料库、词性标注文本语料库、朝(韩)汉对译语料库、朝鲜语(韩国语)病句语料库等。

2016 年朝鲜文基地建设了近 1.5 亿字中国朝鲜语文本语料库,该语料库主要分为文本语料库(1 亿 2 千万字)、双语(多语)对译语料库(2 千万字)、朝鲜语(韩国语)病句语料库(820 万字);研究重点放在对已建语料库进行分类,整合,扩充,加工方面;此外,新录入并重新整理和分类文学类杂志,共计 1 193万字。

朝鲜文基地加工了朝鲜文文本语料库,包括词语切分、词类划分、句法、语义属性标注等,并应用所研发的朝鲜语词性切分软件,对中小学教科书和延边日报文本进行了词性标注,共计 141 万字;对韩汉对译小说和朝汉法律对译文本进行段落对齐,建设共计 580 万字的朝(韩)汉对译语料库。

近 10 年间,朝鲜文基地从朝鲜语专业 700 多名学生教学考核中收集到的作文语料中,共搜集朝鲜语病句语料 800 多万字,并按年级、学期、姓名、考试时间、作文题目、作文体裁等对该语料进行了详细的信息方面的整理与分类,建成了朝鲜语(韩国语)病句语料库。

2015 年 3 月至 2016 年 8 月朝鲜文基地携手朝鲜金日成综合大学研究人员共同研发了集《우리말(试用版)》词性自动标注和语料统计分析于一身的中国朝鲜语综合分析软件。《우리말(试用版)》具有语法错误分析、词性切分、语义查找、语料检索统计等功能。其中语料检索统计功能基本实现了①音素频度;②字节频度;③词汇频度;④单字上下文(以逗号或句号为边界);⑤单词上下文(以句子或段落为边界);⑥按词类大类、小类提取总清单,统计分类总数和分类频度;⑦其他信息等检索统计功能。

2017 年朝鲜文基地开发了"中国朝鲜文文本自动校对软件"(测试版),并将其搭载在 2016 年开发的《우리말(试用版)》词性自动标注和语料统计分析系统中,进而初步实现了开发文本校对和文本分析为一体的综合型应用软件的目标。

1.4.2.4 中国朝鲜语口语语料库建设

2016 年朝鲜文基地完成了 100 h 的标准口语音频数据和 100 h 的标准语双频数据的收集和 140 万字正字法转写库、140 万字语言转写语料库、2.7 万句对平行语料库的构建。此外,朝鲜文基地还把韩国的 33 个实词分类体系扩展到 1 763 个小类,这

一分类体系在国内外均为首创，不仅适用于韩国语本体论研究和韩国语教学研究，也对今后提高韩国语词素分析器的准确度，开发韩国语句式分析器和韩-汉口语计算机辅助翻译工具等具有重要的意义。

在国家社会科学基金一般项目“面向智能信息处理的韩国语口语词汇研究”(16BYY176)和朝鲜半岛研究协同创新中心2016年度拔尖创新人才培育基金项目“基于计量语言学的韩国语口语研究与韩国语教育中的应用”的支持下，朝鲜文基地全面、系统地完成了朝鲜语口语中出现的全部实词的研究，即实词词类研究、体词研究、谓词研究、修饰词研究。体词的研究分为普通名词、依存名词、代词、数词的研究；谓词的研究分为动词、形容词、补助谓词的研究；修饰词的研究分为冠形词和副词的研究。

为了分析韩国语教学中的“语言纯正”问题，卢星华等[14]基于韩国语准口语语料库，运用统计学研究韩国语口语的特征，并对今后编写韩国语听力教材中的对话例文提出了建议。

2017年，朝鲜文基地在已有的66万语节的朝鲜语准口语语料库的基础上，继续按照实际发音进行语音转写，并制作成以语节为单位的“实际发音训练语料库”，该语料库的规模目前已达到90万语节。此外，还研发了“朝鲜语发音软件”，其主要工作：1)逐步完善朝鲜语口语实际发音规则库，主要通过对朝鲜语准口语语音转写语料库中前一个音节的终声和后一个音节的初声之间发生的实际语流音变进行研究；2)完成字库的研制，目前已制作了11 172个朝鲜语字库，这对语音合成与语音生成具有重要意义；3)完成字素库的研制，即对字库中的每一个字进行字素分析，并提取出每个字的初声、中声、终声。

1.4.2.5 其他研究成果

2017年朝鲜文基地针对朝鲜语语言文字结构与识别、朝鲜语与蒙古语语音对比分析、朝鲜语与汉语跨语种的信息检索，以及朝鲜语语言文字和语音语料库等方面持续进行研究开发和建设工作，并在初步开发“中-朝-日-英生物学术语对应软件”的基础上，收录了“中-朝-日-英生物学术语”628项对应词库和图片。

2 存在问题

30多年来，壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字信息处理技术领域取得了不少研究成果，产生了积极的社会效益和经济效益，并有力地推动了相应少数民族语言文字信息技术的发展，但与中文信息处理技术相比，目前还存在技术研发难度大、人才缺乏、政策关注度低、研究成果不多等较多问题，这些问题制约了壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字信息处理技术的进一步向前发展。

2.1 有些语言缺乏统一规范化的标准

标准化是壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字信息处理技术上需要进一步解决的问题。如古壮字的编码体系的标准化，包括古壮字编码字符集标准、古壮字的输入码和古壮字字形标准等。标准化更有益于壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族文化的研究、交流和传播。

2.2 技术应用范围狭窄

当前，壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字的推广力度还不够，应用范围也不够广泛，且其信息处理技术也多用于学术研究，这虽然能够在很大程度上支持壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字的研究和发展，但不应该忽视其潜在的更大的社会价值。因此，壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字信息处理技术不仅在语言和文化推广上，而且在民族关系上都应该起到更积极的促进作用。

3 展望

民族文化的信息化发展是文化大发展大繁荣的必然道路。壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族文化的传承和发展，迫切需要相应少数民族语言文字信息处理技术的支撑。在互联网、移动互联网、智能终端和云计算技术应用日益广泛的今天，少数民族语言文字的信息处理技术应该更加丰富，更多层次化、实用化，以满足学术研究、古籍出版、民族教育、日常生活等多方面需求。因此，壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字信息处理工作的主要发展方向以后将集中在语言资源开发和信息处理技术应用两大方向。前者坚持大规模、深度加工和多维度研究的发展方向；后者以工程开发和产品化相结合为主，把语料、软件的开发与应用从以往的单机版变成网络版模式，并将其推广给众多用户，从而将过去单一的、只面向科学和教育教学的研究转化成面向服务、面向应用的趋势，以达到高效利用、共享资源的效果，全面发挥其应用性和社会价值。未来可能的工作：

(1)建设互联网上共享的壮、蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族文化资源库。中国有56个民族,有数以万计的群众在日常中使用其民族语言进行交流,现有的少数民族语言电视、广播等宣传媒体数量少、内容简单、覆盖面小,无法满足广大少数民族群众对现代资讯的需求,影响城乡均衡发展。因此,建立数字化的少数民族文化资源库,通过互联网、智能终端面向全社会开放共享,将是当前及未来行之有效的形式。

(2)研究开发古壮文和现代壮文之间的翻译系统、古壮文和汉文翻译系统以及蒙古、维吾尔、哈萨克、柯尔克孜、朝鲜等少数民族语言文字和汉文翻译系统的工作迫在眉睫。通过技术手段加速古籍的收集、整理、出版工作,使少数民族的宝贵文化财富能够保留和传承下去,意义重大。

(3)建设面向社会的实用化的少数民族语言文字信息处理平台和手机APP,开发少数民族语言文字多媒体教学系统、语音识别系统,研发云服务方式的现代少数民族语言文字学习和在线翻译软件,以更加形象和方便快捷的方式普及少数民族语言文字,推动民族教育,传承民族文化。

(4)从政府层面加以积极引导,加强少数民族语言文字标准化研究工作,进一步扩大制定标准的领域与行业,以提高少数民族语言文字信息处理技术的创新能力,促进其融合当前先进信息技术,提供丰富多彩的技术应用产品,从而使少数民族文化的传播范围更为广泛。

参考文献:

[1] 刘连芳,顾林,廖宏.古壮文操作系统和编辑排版系统[J].计算机应用研究,1993(6):32-34.
LIU L F, GU L, LIAO H. Ancient Zhuang operating system and edit typesetting system[J]. Application Research of Computers, 1993(6):32-34.

[2] 刘连芳,顾林,黄家裕,等.壮文与壮文信息处理[J].中文信息学报,2011,25(6):175-182.
LIU L F, GU L, HUANG J Y, et al. Zhuang language and its information processing[J]. Journal of Chinese Information Processing, 2011, 25(6):175-182.

[3] 覃忠群.《壮文电子词典及辅助翻译软件》语料库建设的经验[J].民族翻译,2013(2):74-78.
QIN Z Q. "Zhuang electronic dictionary and auxiliary translation software" corpus construction experience [J]. Minority Translators Journal, 2013(2):74-78.

[4] 僚人导航网[EB/OL].[2017-11-12]. http://www.jiu60.com/hoiz/.
Liao navigation network[EB/OL]. [2017-11-12]. http://www.jiu60.com/hoiz/.

[5] 吐尔根·依布拉音,袁保社.新疆少数民族语言文字信息处理研究与应用[J].中文信息学报,2011,25(6):149-156.
TURGUN·IBRAHIM, YUAN B S. A survey on minority language information processing research and application in Xinjiang[J]. Journal of Chinese Information Processing, 2011, 25(6):149-156.

[6] 古丽拉·阿东别克,达吾勒·阿布都哈依尔,木合亚提·尼亚孜别克,等.现代哈萨克语词级标注语料库的构建研究[J].新疆大学学报:自然科学版,2009,26(4):394-401.
GULILA ALTENBEK, DAWEL ABILHAYER, MUHEYAT NIYAZBEK, et al. A study of word tagging corpus for the modern Kazakh language[J]. Journal of Xinjiang University: Natural Science Edition, 2009, 26(4):394-401.

[7] 董兴华,周俊林,郭树盛,等.基于短语的汉维/维汉统计机器翻译[J].计算机工程,2011,37(9):16-18,21.
DONG X H, ZHOU J L, GUO S S, et al. Phrase-based Chinese-Uyghur/Uyghur-Chinese statistical machine translation[J]. Computer Engineering, 2011, 37(9):16-18,21.

[8] 达吾勒·阿布都哈依尔,古丽拉·阿东别克.基于ANN的哈萨克文手写文字识别系统的研究[J].计算机工程与应用,2008,44(1):225-228.
DAWEL ABILHAYER, GULILA ALTENBEK. Hand-written Kazakh character recognition system using artificial network[J]. Computer Engineering and Applications, 2008, 44(1):225-228.

[9] 姑丽加玛丽·麦麦提艾力,艾斯卡尔·艾木都拉.基于音素及其特征参数的维吾尔语音合成技术[J].中文信息学报,2008,22(4):100-104.
GULIJIAMALI MAIMAITIAILI, AISIKAER AIMUDULA. The phoneme feature based Uyghur speech synthesis[J]. Journal of Chinese Information Processing, 2008, 22(4):100-104.

[10] 孜丽卡木·哈斯木,那斯尔江·吐尔逊,吾守尔·斯拉木.维吾尔语词首音节元音声学分析[J].中文信息学报,2009,23(5):114-118.
ZILIKAM KASIM, NASIRJAN TURSUN, WUSHOUR SILAMU. Acoustic analysis of the initial syllabic vowels in Uyghur language[J]. Journal of Chinese Information Processing, 2009, 23(5):114-118.

[11] 玄龙云,崔荣一.关于朝鲜文信息技术标准化[J].中文信息学报,2016,30(3):85-89.
XUAN L Y, CUI R Y. Towards Korean information

technology standardization[J]. Journal of Chinese Information Processing,2016,30(3):85-89.

[12] 朴明姬,崔荣一. 多语种文本图像中的文字语种辨识方法的研究[J]. 中文信息学报,2017,31(2):220-225.
PIAO M J,CUI R Y. An approach to script identification in image with multi-lingual texts[J]. Journal of Chinese Information Processing,2017,31(2):220-225.

[13] 刘双君,金小峰,崔荣一. 基于基频的朝鲜语方言辨识方法的研究[J]. 中文信息学报,2017,31(2): 55-60,70.
LIU S J,JIN X F,CUI R Y. Research on Korean dialect identification based on pitch feature[J]. Journal of Chinese Information Processing,2017,31(2):55-60,70.

[14] 卢星华. 韩国语听力教材研究——基于韩国语准口语语料库中的代词计量特征分析[J]. 东疆学刊,2017,34(3):62-67,112.
LU X H. Research on Korean listening textbooks—The analysis of the characteristics of pronouns based on the speech corpus of Korean language [J]. Dongjiang Journal,2017,34(3):62-67,112.

(责任编辑:陆 雁)

(上接第 17 页 Continue from page 17)

[36] 钟卿. 基于 SVM 的联机手写新傣文字符识别[D]. 昆明:云南大学,2016.
ZHONG Q. Online New Tai Lue handwritten character recognition based on the SVM[D]. Kunming:Yunnan University,2016.

[37] 陈瑞新. 基于随机森林的联机手写新傣文字符识别技术研究与实现[D]. 昆明:云南大学,2016.
CHEN R X. Online handwritten character recognition of New Tai Lue based on random forest[D]. Kunming:Yunnan University,2016.

[38] 李慧. 词典与统计相结合的傣文分词方法与实现[D]. 昆明:云南大学,2016.
LI H. Dai language segmentation based on dictionarv and statistics [D]. Kunming: Yunnan University, 2016.

[39] 胡刚,王嘉梅,李炳泽,等. 傣泐文-汉文互译有声电子词典[J]. 计算机系统应用,2016,25(7):8-16.
HU G,WANG J M,LI B Z,et al. Daile Wen-Chinese translation audible electronic dictionary[J]. Computer Systems & Applications,2016,25(7):8-16.

[40] 殷建民,玉康龙,岩香,等. 西双版纳傣文电子词典及辅助翻译技术研究[C]. 第十五届全国少数民族语言文字信息处理学术研讨会. 延边,2015.
YIN J M,YU K L,YAN X,et al. Research on electronic dictionary and auxiliary translation technology of Tai Lue[C]. The Fifteenth Session of the Academic Symposium on the Information Processing of the Language and Character of the National Minorities. Yanbian,2015.

(责任编辑:陆 雁)