广西科学院学报
Journal of Guangxi Academy of Sciences

2018,34(1):27～31
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-02-22 DOI:10.13657/j.cnki.gxkxyxb.20180222.001
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180222.1158.002.html

东南亚语言及信息处理研究进展*

The Progress of Studies on Southeast Asian Languages and Information Processing Thereof

黄家裕[1],刘连芳[1],邓姿娴[2],温家凯[2]
HUANG Jiayu[1], LIU Lianfang[1], DENG Zixian[2], WEN Jiakai[2]

(1.南宁市平方软件新技术有限责任公司,广西南宁 530007;2.广西达译商务服务有限责任公司,广西南宁 530007)
(1. Nanning Pingsoft New Software Technology Co., Ltd., Nanning, Guangxi, 530007, China; 2. Guangxi Daring E-commerce Services Co., Ltd., Nanning, Guangxi, 530007, China)

摘要:本文首先介绍各东南亚语言的特点,并重点介绍越南语与中国壮语的关联以及它们在信息处理上的相近性;然后介绍国内外越南语信息处理的工作现状,并分析汉越机器翻译的发展趋势;最后展望东南亚语言及壮语信息处理的下一步工作。

关键词:东南亚语言 越南语 壮语 信息处理
中图分类号:TP391 **文献标识码**:A **文章编号**:1002-7378(2018)01-0027-05

Abstract: This article first introduces the characteristics of each Southeast Asian language and focuses on the relevancy between Vietnamese and the Chinese Zhuang language as well as their similarity in information processing. Then it introduces the status quo of Vietnamese information processing at home and abroad, and analyzes the development tendency about Chinese-Vietnamese machine translation. Finally, it looks forward to the further work of information processing in Southeast Asian languages and the Chinese Zhuang language.

Key words: Southeast Asian languages, Vietnamese, Zhuang language, information processing

0 引言

东南亚国家位于“一带一路”的海上丝绸之路沿线,是“一带一路”合作的重要组成部分。随着中国-东盟(东南亚国家联盟)自贸区升级版的打造,以及“一带一路”倡议的推进,中国和东盟国家之间的经贸往来、社会交往及文化交流更加密切,语言信息服务需求迅猛增长。虽然英语是国际交流中使用最广泛的语言,但是大多数亚洲国家的公民还不能使用或不能流利地使用英语交流。要真正做到“一带一路”沿线国家沟通顺畅,必须用本国语言进行交流。因此本地语言服务是海上丝绸之路建设的最基本需求,东南亚语言的信息处理研究势在必行。

广西与东南亚海陆相连,是历史上海上丝绸之路的起点之一。2004 年,一年一度的中国-东盟博览会永久落户广西首府南宁市;2008 年,国家批准实施《广西北部湾经济区发展规划》,广西北部湾经济区正逐步打造成中国-东盟开放合作的物流基地、商贸基地、加工制造基地和信息交流中心;2015 年 3 月,习近平主席明确对广西的“三大定位”,即构建面向东盟的国际大通道,打造西南中南地区开放发展新的战略支点,形成 21 世纪海上丝绸之路与丝绸之路经济带有机衔接的重要门户。在国家各项政策的支持和帮助下,以及各族人民的辛勤努力下,广西正

收稿日期:2017-11-01
作者简介:黄家裕(1975－),男,高级工程师,主要从事中文信息处理研究,E-mail:huangjy98@163.com。
*广西壮族自治区工业和信息化发展专项资金(信息服务)项目(桂工信电软〔2017〕128 号)资助。

在朝着成为中国连接东南亚的枢纽的方向努力。

另一方面，广西的许多民族与周边东南亚国家的民族同宗同源，如，壮族和越南的岱依族、依族是跨境民族，而广西京族和越南京族则是同族，他们语言相通，文化相通；又如壮语和泰语都属于汉藏语系壮侗语族壮傣语支，其语言非常相似。因此，广西在与东盟各国的交往中具有相对的人文优势和语言优势。在中国-东盟博览会的推动下，广西各大高校的东南亚语言专业不断扩招，并新增开设其他东南亚语言文学专业[1]，目前已形成较为全面的专业设置；同时广西加大与东南亚国家的留学生互派，目前互派留学生人数居全国首位[2]。因此在广西开展东南亚语言的信息处理研究工作具有良好的基础和较大的优势。

本文首先介绍各东南亚语言的特点，并重点介绍越南语与中国壮语的关联，进而阐述借助壮语信息处理基础从越南语着手东南亚语言信息处理研究的合理性；然后介绍国内外越南语信息处理的工作现状，并分析汉越机器翻译的发展趋势；最后展望东南亚语言及壮语信息处理的下一步工作。

1 东南亚语言特点

除英语之外，东南亚主要官方语言有越南语、泰语、老挝语、缅甸语、高棉语（柬埔寨语）、马来语、印尼语、菲律宾语（他加禄语）。这些语言的特点如下：

（1）所属语系

越南语系属存在争议，一般认为属于南亚语系；泰语、老挝语和缅甸语属于汉藏语系；印尼语、马来语和菲律宾语属于南岛语系；高棉语属于南亚语系。

（2）语言形态

越南语、高棉语、泰语、缅甸语及老挝语都是分析语，即孤立语，这点跟汉语相同；其名词没有性、数、格的变化；复合词多，派生词少；在组句时主要靠词序和虚词来表达语法关系。

印尼语、马来语及菲律宾语属于黏着语；其词形变化丰富，主要的构词手段是词根附加成分和词根重叠；组句时通过词的形式变化来表达语法关系。

（3）语音声调

越南语、泰语、老挝语及缅甸语都是声调语，像汉语一样，同一个语音，不同长短、高低的声调可表达不同的语义；其中越南语和老挝语的声调最丰富，多达6个声调。

印尼语、马来语、菲律宾语及高棉语则是非声调语，像英语一样，语音声调的不同长短只表示语气，不影响语义。

（4）句法

越南语、高棉语、印尼语、马来语、泰语及老挝语都跟汉语一样使用“主-谓-宾”的基本语序，但与汉语不同的是，其修饰语都放在被修饰的中心词之后。

缅甸语使用“主-宾-谓”语序。

菲律宾语的语序比较特殊，它使用“谓-宾-主”或“谓-主-宾”语序。

（5）借词

东南亚语言都存在大量的借词（外来词），不同的历史发展也使得他们的借词来源各不相同。

其中，高棉语、马来语、泰语、老挝语、缅甸语受佛教传播影响最深，很多借词都来源于巴利语和梵语，多数属多音节词，仍保留其源语的性和数的特征。

越南语受汉语、法语、英语影响较深，汉语借词占比非常大，来自法语和英语的借词主要集中在科技词汇上。

印尼语受荷兰语体系影响较大，其词汇含大量爪哇语、荷兰语借词。马来语的借词则主要来源于梵语和阿拉伯语。马来语和印尼语非常相近，使用这两种语言的人基本上可以彼此沟通，而它们的差异主要来自印尼语的爪哇语和荷兰语借词。

菲律宾语词汇则受西班牙语影响较深。

（6）书写系统

越南语、印尼语、马来语、菲律宾语都采用拉丁字母表音文字。其中越南语拼写相对复杂，其包括7个拉丁字母变体（越南语专用字母）和6个声调符号。

高棉语、泰语、老挝语及缅甸语采用各自独特但渊源相近的专用字母表音文字。

广西是壮族自治区，毗邻东南亚，壮语与东南亚语言有不少相通、相近之处。特别是广西与越南海陆相连，壮语与越南语源远流长。

2 越南语与中国壮语的关联

2.1 越南语的发展历程

越南语受汉语影响非常大，据统计，越南语中的汉语借词约占词汇总数的60%[3]，在科学、政治、法律等领域的借词占比甚至达到70%～80%，而且越是正式场合，汉借词就越受青睐。在文字方面，越南从接受汉文化开始，就长期使用汉字。由于越南语与汉语的发音不同，汉字难以准确记录越南语，约公元13世纪前，越南文人开始创造一种方块文字[4]，

即“喃字”(或叫“字喃”)。这种文字有的直接取用汉字,有的借用汉字的偏旁,仿照中国“六书”中的假借、会意、形声创造而成,这类似于古壮文的造字方法。公元 13 世纪左右,“喃字”趋于系统化并被广泛推广,但汉字仍然一直是官方文字。17 世纪上半叶,在越南传教的西方传教士创制出拼音越南文;1945 年后,这种拼音文字正式成为越南的官方文字使用至今,称国语字(图 1)。

Gần đây, Hội chợ thương mại máy móc công trình quốc tế
Trung Quốc (Từ Châu) lần thứ 3 (gọi tắt lạ "triển lãm
Từ Châu") đã khai mạc tại Khu vận tải hàng hoá Tân Trường
Giang Từ Châu, với thời gian là 3 ngày, có gần nửa số c
ác doanh nghiệp nổi tiếng thuộc top 50 trong bảng xếp hạng
thế giới về lĩnh vực này gồm Từ Công, Jimmy Carter Ho
a Kỳ, Doosan Hàn Quốc đến tham dự triển lãm.

图 1 越南国语字

Fig. 1 The writing of the national language of Vietnam

越南国语字共有 29 个字母,其中 12 个是元音字母,17 个是辅音字母;书写以音节为单位,每个音节由声母、韵母和声调 3 部分组成,音节与音节之间以空格或标点间隔;多数的音节本身是一个词,也有一些音节必须跟其它音节结合才构成词,因此越南语在多数使用场合下需要分词处理。

2.2 越南语与壮语的对比

广西壮族与越南主体民族京族世代相邻而居,长期交往,语言接触频繁。另外壮语与越南语有着相似的发展历史:两个民族在同一时期受到汉文化影响,语言中都存在大量的汉借词;都借用汉字创造本民族文字(壮族的土俗字,或称古壮字;越南的喃字),这些文字都长期在民间使用;而近代都先后在外部的帮助下创造拼音文字(壮族的新壮文和越南的国语字),成为各自目前使用的文字(图 2)。

Siengzlungz Siujyoz dwg Dungz- lingj Cin aen
gyauyozdenj ndeu, caeuq cunghsinhyau
doxgek yaek caet cienmij. Gangj aen hagdangz neix
“iq”, dwg aenvih gizneix hagseng iq、saeseng
noix. Aen hagdangz de cijdwg aen gyauyozdenj ndeu,
cij banh 1 daengz 3 nienzgaep, cijmiz 3 aen
ban, 96 boux siujhagseng, lauxsae hix ngamq miz 6 vunz.
Aen gyauyozdenj neix daj 2013 nienz 9
nyied hwnj ngamq hainduj guh Cuengh Gun song saw
vah sonhag.

图 2 新壮文

Fig. 2 The writing of new Zhuang language

虽然越南语的系属问题尚存争议,一般认为,越南语属于南亚语系孟高棉语族越芒语支,而壮语属于汉藏语系壮侗语族壮傣语支,但研究表明,越南语和壮语在词法和句法上都非常相似,很多语法事实对应一致[5]。

例:

(1)名词的修饰词后置

汉:牛肉

壮:noh(肉)vaiz(牛)

越:thit(肉)bò(牛)

(2)状语后置

汉:我先走

壮:gou(我)byaij(走)gonq(先)

越:tôi(我)đi(走)trước(先/前)

另外,从图 1 和图 2 中可以看到,新壮文和越南国语字书写形式相似,都是拼音拉丁文,都以空格间隔音节,因此在信息处理上可以采用相近的方法。

综上,越南语和壮语天然相近,文字、语法相似,在信息处理研究上完全可以互相促进。由于对壮文信息处理的研究已经有 20 多年的历史,因此,在从事东南亚语言信息处理中,可以从最为熟悉、较为容易的越南语信息处理切入。

3 越南语信息处理

3.1 越南国内的越南语信息处理

21 世纪第一个 10 年,越南国内设立了 2006—2010 年度 5 年期国家重点科技攻关项目:Vietnamese Language and Speech Processing(VLSP)[6],该项目有 11 个研究组参加,目的是解决之前存在的诸多问题,如底层工作偏少、各自独自工作、没有标准、没有继承、没有共享、没有合作、没有资源、没有工具等等。项目设立了两大目标:构建 VLSP 基础设施,特别是不可或缺的资源和工具;开发几款典型的 VLSP 公众终端产品。

该项目的实施目标产品包括面向应用的语音识别与合成系统、具有较大词汇量的语音识别系统、英越翻译系统、互联网应用支持系统、越南语拼写检查系统、语音识别语料库、语音合成语料库、特殊词汇语料库、越南语树库、英越词典、英越句对库、越南语词典、越南语分词器、越南语词性标注器、越南语短语识别器、越南语句法分析器等。

该项目的实施为越南国内的信息处理打下较好的基础,项目成果正在获得使用并被持续改进中。

在机器翻译方面,越南国内关注重点是英越机器翻译的研究,目前采用的技术仍以 Statistical Machine Translation(SMT)为主。

3.2 中国在越南语信息处理领域的研发

因地理、文化、人才上的优势，广西和云南是国内较早开展东南亚语言信息处理研究的省区。随着“一带一路”倡议的推进，国内众多研究机构、企业单位都纷纷加入东南亚语言的研究，如讯飞的越南语语音合成，百度、搜狗、阿里的越南语机器翻译等等。

南宁市平方软件新技术有限责任公司和广西达译商务服务有限责任公司是国内较早开展东南亚语言信息处理的企业，这两个公司重点从事汉越双语平行语料库建设、汉越机器翻译及辅助翻译研究。

3.2.1 语料库建设

语料库的构建是一个长期艰苦的过程，需要科学的规划和良好的组织，合理、有效地推进。根据汉越/越汉语料现状，制定了包括采集、加工、维护以及组织4大过程的语料建设流程及各环节的规范（图3）。

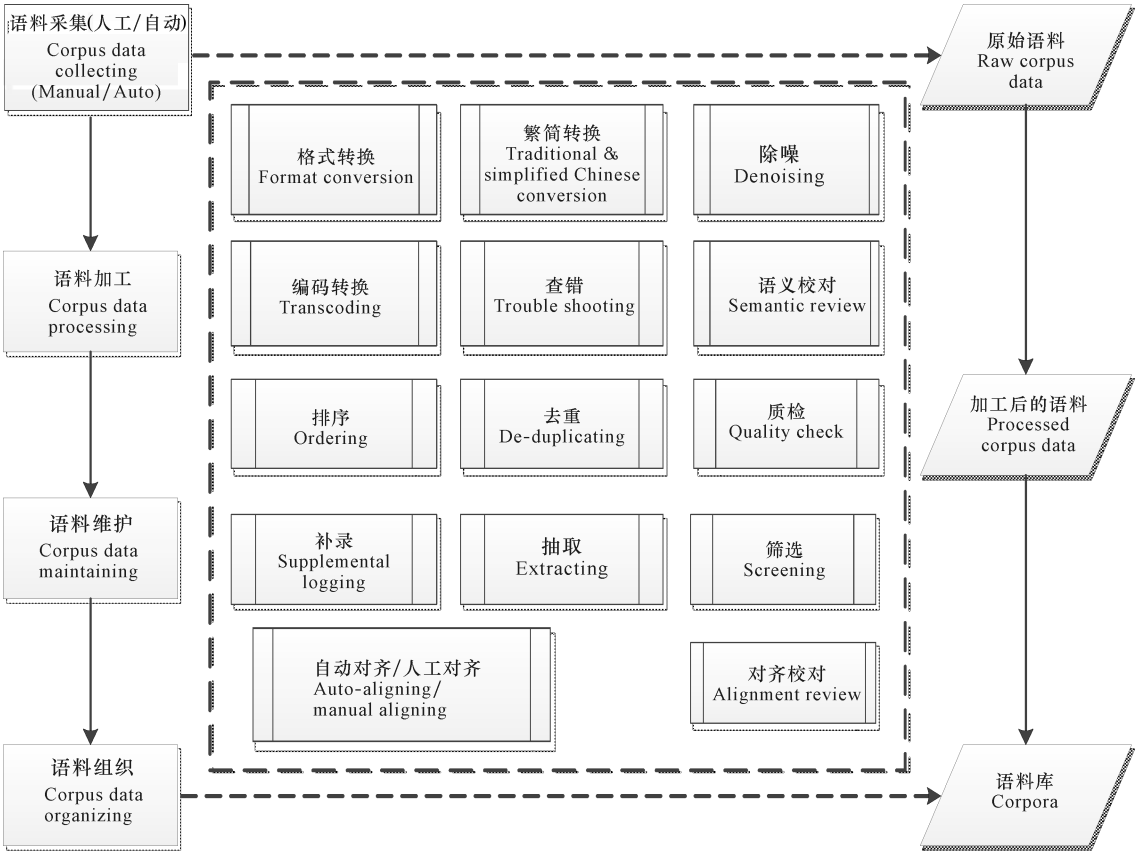

图3 语料库建设

Fig. 3 The construction process of corpora

根据越南语的语料特点，研发了大量的工具以辅助这些过程，包括互联网语料自动发现、自动采集、格式转换、编码转换、除噪、去重、过滤、篇章对齐、句子对齐及错误检查等等，大量减轻人工处理的负担。

目前，语料的数量和质量是制约汉越机器翻译效果的主要因素，汉越平行语料远远少于英汉、英越的平行语料，获取难度也远远大于英汉、英越平行语料的获取。在这种情况下，除了加大常规语料库建设的工作力度外，目前正在采取以下技术手段增加汉越平行语料：

(1)借助较为丰富的汉英、越英语料，挖掘汉越语料；

(2)采集丰富的单语语料，运用现有的汉越机器翻译系统“生产”汉越双语语料，经过人工处理，产生有质量保证的汉越双语语料。

3.2.2 汉越机器翻译

在汉越双语平行语料基础上，采用SMT技术开发汉越机器翻译系统。在该系统中，针对越南语的文字特点进行特殊处理及改进，如越南语数字的小数点和千位符处理、调序规则改进、专有名词音译等等。内部人工评测显示，系统总体效果与目前几大在线机器翻译系统效果相当。在汉越人名音译方面，准确率达到97.41%[7]。

谷歌的Google's Neural Machine Translation System(GNMT)取得巨大成功后，Neural Machine

Translation(NMT)已经成为机器翻译的主流方向。可以预计,国内外的汉越机器翻译技术研究也将以NMT为主。

许多NMT系统倾向于抛开传统的自然语言处理方法,但近期多项研究表明,传统自然语言处理技术对NMT依然有较大帮助。如微软将句法知识引入到神经网络编码和解码之中,得到了更佳的翻译效果[8];又如,腾讯人工智能实验室在神经机器翻译中进行源句法建模,实验结果显示翻译效果获得显著提高[9]。因此开展越南语自然语言处理的基础研究仍然很有必要。

4 展望

在"一带一路"倡议的带动下,将会有更多的企业、单位加入东南亚语言信息处理行列,这也将会促进东南亚语言信息处理水平的提升。在众多工作中,语料库建设仍然是最重要的工作之一,不管Computer Aided Translation(CAT)、SMT还是NMT,都离不开大规模双语平行语料。2017年4月份在南宁召开的"亚洲语言资源国际研讨会"上,来自中国西藏、新疆、广西、香港等省区,以及意大利、泰国、越南、老挝、日本、韩国等国家的与会专家一致认为,语言资源是"一带一路"助推器,应该开展广泛的语言资源建设与共享的国际合作。我们将加强国际合作,共建共享大规模、高质量平行语料库,包括汉越平行语料库,为提升SMT、NMT汉越双向机器翻译的准确率打下良好的数据基础。

鉴于东南亚语言信息处理与壮文信息处理的相互关联,在进行东南亚语言信息处理的同时,继续推进壮文信息处理是我们的工作重点之一。

参考文献:

[1] 陈慧.中国东南亚语言专业现状及发展趋势[J].东南亚纵横,2007(3):72-75.
CHEN H. The situation and developing trend of language department of South-east Asia in China[J]. Around Southeast Asia,2007(3):72-75.

[2] 蒋雪林,王雪.广西与东南亚国家互派留学生规模居中国第一[J].海外华文教育动态,2013(6):34-35.
JIANG X L,WANG X. Guangxi takes the first place in China at the scale of two-way overseas students with Southeast Asian countries[J]. Overseas Chinese Education News,2013(6):34-35.

[3] 谭志词.汉语汉字对越南语言文字影响至深的原因初探[J].东南亚,1998(2):47-50.
TAN Z C. A primary analysis on reasons for which written Vietnamese is profoundly influenced by Chinese characters[J]. Southeast Asian,1998(2):47-50.

[4] 林明华.越南文字浅谈[J].现代外语,1983(3):55-59.
LIN M H. A brief discussion on written Vietnamese [J]. Modern Foreign Languges,1983(3):55-59.

[5] 黄巧丽.越南语和壮语词的词与词组的结构对比研究[D].南宁:广西民族大学,2013.
HUANG Q L. Comparative study of words and phrases' structure between Vietnamese and the Zhuang language[D]. Nanning:Guangxi University for Nationalities,2013.

[6] NGO Q H,WINIWARTER W,WLOKA B. EVBCorpus—A multi-layer English-Vietnamese bilingual corpus for studying tasks in comparative linguistics[EB/OL].[2017-08-10]. https://www.researchgate.net/publication/259163984_EVBCorpus_-_A_Multi-Layer_English-Vietnamese_Bilingual_Corpus_for_Studying_Tasks_in_Comparative_Linguistics.

[7] 申文明,刘连芳,黄家裕,等.基于概率模型的汉语和越南语的人名音译方法[J].广西科学院学报,2010,26(4):439-442.
SHEN W M,LIU L F,HUANG J Y,et al. The approach of Chinese-Vietnamese name transliteration based on probabilistic model[J]. Journal of Guangxi Academy of Sciences,2010,26(4):439-442.

[8] 微软亚研院副院长周明:口语机器翻译在未来肯定会完全普及[EB/OL].[2017-08-10]. http://www.geekpark.net/topics/219714.
ZHOU M Vice President of Microsoft Research Asia: Machine interpretation will surely get completely universal in the future[EB/OL].[2017-08-10]. http://www.geekpark.net/topics/219714.

[9] LI J H,XIONG D Y,TU Z P,et al. Modeling source syntax for neural machine translation[C]. Proceeding of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver,Canada,2017:688-697.

(责任编辑:米慧芝)