汉英-泰互译有声语料的数据库研究①

胡 刚[1,2], 王嘉梅[1,2], 李炳泽[1], 林 睿[1,2], 林碧彤[1,3]

[1](云南民族大学 云南省高校少数民族语言文字信息化处理工程研究中心, 昆明 650500)
[2](云南民族大学 电气信息工程学院, 昆明 650500)
[3](云南民族大学 国际教育学院, 昆明 650500)

摘 要: “汉英-泰互译有声语料库”的开发, 在泰文舆情分析领域, 解决了词典分词算法中训练语料缺乏的问题. 本文采用计算机化信息处理技术, 对大量的收集来的泰文语料进行整理、规范、加工与存储, 统计出泰文词汇 8000 多个. 然后利用词典翻译和人工校对其进行语料对齐. 最后, 结合泰文语言语法特征以及句法的语义特点, 分类归纳和规范标注泰语语料, 构建了 5 万条左右的的汉英-泰语料数据库.

关键词: 汉英-泰; 语料库; 互译有声; 舆情; 泰文分词

Research on the Database of English Chinese-Thai Translation Audible Corpus

HU Gang[1,2], WANG Jia-Mei[1,2], LI Bing-Ze[1], LIN Rui[1,2], LIN Bi-Tong[1,3]

[1](Yunnan Province for Minority Language Information Processing Engineering Research Center, Yunnan Minzu University, Kunming 650500, China)
[2](School of Electrical & Information Engineering, Yunnan Minzu University, Kunming 650500, China)
[3](School of International Education, Yunnan Minzu University, Kunming 650500, China)

Abstract: The development of “English Chinese - Thai Translation Audible Corpus”, in the field of Thai public opinion analysis, it solves the problem of the lack of training corpus in dictionary segmentation algorithm. In this paper, the computerized information processing technology is used to organize , standardize , process and store large amounts of collected Thai corpus, and then more than 8000 of the Thai vocabularies are finished. And then it uses the dictionary translation and manual calibration to align corpus. Finally, Thai corpus are classified and marked normally, combined with syntax characteristics and syntactic semantic features of Thai language, and then the database of English Chinese - Thai corpus is constructed, which contains about 50 thousand of the query terms.

Key words: English Chinese - Thai; corpus; translation audible; public opinion; Thai segmentation

近年来, 随着自然语言处理的不断发展和完善, 大规模语料文本处理已经成为了计算机语言学界的一个热门话题. 语料库资源对于自然语言处理研究的巨大价值已经得到越来越多的学者认可, 一个重要的原因是任何文本的加工、处理都是以语料库作为作基础的. 特别是双语语料库(Bilingual Corpus,包含两种语言互译文本的语料库)已经成为机器翻译、数据挖掘领域中能够不可或缺的重要资源, 特别是在少数民族语言舆情研究领域尤为重要[1].

一方面, 是来自泰文网络舆情的预测的需要, 工作开展依次经过网页信息获取、文本预处理、文本切割分词、文本分类/聚类、话题跟踪检测、倾向性分析来进行的, 而其中泰文分词环节是基于词典算法实现的, 必须由语料库-词表作为分析参考. 计算机在分词的时候都是需要一个词表的, 词表的作用是在分词的时候根据词表确定字符串是否属于一个分词单位, 然后按照词的含义、语法意义依次对文本进行词汇分割[1]. 具体采用的是词典基于字符串匹配的泰文分词方法,

①基金项目:国家自然科学基金(61363085);国家语委重大科研项目(WT125-61);云南省教育厅科学研究基金重大专项(ZD2013013);云南民族大学高水平民族大学建设科研项目(ZZZC1501-JF12002);云南民族大学研究生创新基金重点项目(2015YJCXZ17)
收稿时间:2015-11-17;收到修改稿时间:2015-12-25 [doi: 10.15888/j.cnki.csa.005242]

又叫机械分词方法, 是用获取网页文本中的待选词去跟语料库分词词表进行匹配来判断分词单位是否.

一方面, 是来自泰文网络舆情的控制的需要, 结果识别要通过泰汉辅助翻译技术实现中文呈现来进行的. 泰文舆情分析平台的开发可以有泰文专家的协助, 但对于用户而言, 大部分都不懂泰语, 无法理解和判断民族语舆情分析的结果, 跟谈不上舆情处理. 语料库的建立是解决监控分析结果的可理解性问题的关键. 舆情分析结果中文呈报主要包括两个层级的实现, 第一层为词语级的翻译呈报, 利用泰汉双语敏感词词典, 将敏感类词汇信息采用藏汉双语形式呈现, 使用户能够直观判断疑似文本所涉及的主要内容; 第二层为句子级的辅助翻译, 利用泰汉统计机器翻译模型将分析后的泰文文本翻译成中文, 使用户能够理解舆情分析结果和分析对象的具体涵义. 无论是词语级, 还是句子级的辅助翻译, 都是基于大规模泰-汉英语料库中的部分或全部实现的, 借助泰文语法和语义特点, 建立双语翻译词典和机器翻译模型[2].

因此, 从跨境网络舆情研究预测和控制的双面需求出发, 建设“汉英-泰互译有声语料库“是开展泰文跨境网络舆情分析的一个重要组成部分.

1 研究现状

近些年, 随着民族信息化的发展, 一些高校及民族研究所也相继开发了汉-英、汉-蒙、汉-维以及泰-汉语料库, 但目前这些泰-汉语料库一个是出自产权保护无法共享使用, 二是应用主要还局限在机器翻译、词典编撰、语言对比研究领域, 并不能满足舆情分析的需要[3]. 在中文舆情应用领域上, 语料库往往是需要对词库增加词性标注和平行对齐操作的, 而这对民族语同样也不例外. 如此, 针对性的用于舆情领域的语料库可以说是一个空白, 还有待深入研究. 因此,构建一个大规模、多领域、高性能的泰-汉-英互译语料库是当前研究的重点问题之一. 在民族网络舆情的工作开展中, 课题组主辅研究兼并, 从文本分词为主, 机器翻译为辅的角度出发, 提出了泰-汉双语语料库构建的组织方案. 在完成基于词典匹配算法的泰文分词处理中, 并从中获取双语词典及其翻译模式,改进传统的机器翻译方法,对于双语词典编纂、跨语言的对比研究也是具有一定价值的.

2 应用价值

(1) 可作高校或科研院所语言研究的参考文本. 如教学个案, 应用语言学接触个案, 傣泰语支音系研究, 语言比较研究、语义综合研究、词典编撰等; 如研究个案, 计算语言学分析个案, 分词方式研究、双语互译研究、文本抽取等.

(2) 可用于泰语分词的词表参考, 泰文自动分词是泰文信息处理中一项不可缺少的基础性工作. 基于词典的自动分词技术在文本信息处理中处于基础地位, 因而在自动分词的诸多问题中,词表所包含词的覆盖范围对切分的精确度有重要的影响. 大规模泰文语料包含诸多泰文词汇和语句, 为今后开展泰文分词工作提供分词表筛选来源.

(3) 可辅助泰-汉电子词典的开发, 语料库是作为词典最底层的查询数据库支撑. 后续只要对语料库做一些预处理(数据表结构制定、数据库方式选择、数据存储形式、编程开发)便可完成词典制作. 其中, 查询内容是根据各数据表内元素来确定的, 若要词典支持有声翻译查询, 需要将泰-汉英有声语料库分成词汇数据表和语音数据表, 且两个不同类型的数据表通过关键词联系起来.

(4) 可作德宏州、西双版纳州等傣泰民族地区语言文化资源, 语料库包含了少数民族(衣、食、住、行等生活习俗)的文本和语音表示, 从文字、语法结构上的不同反映了民族文化的差异. 相关语料的搜集、整理是民族文化抢救保护工作的重要议项, 在传统的纸质媒介中生疏文字容易被丢失或遗忘, 采用电子化媒介能够在文档中快速检索特定信息, 而且语料以统一文本形式集成便于检查和比较.

3 依托平台设计

3.1 系统概述

目前, 相关研究已经完成了对中英网络信息安全(如: 网页信息获取、文本分类/聚类及语言翻译、话题跟踪检测等)相关理论及方法做了一些初步研究, 本课题拟在面向中英文新理论、新方法和新技术的基础上改进和创新, 结合计算机语言学, 挖掘东盟跨境语言网络舆情信息. 泰文网络是本课题研究的网页载体, 对网络舆情信息相关挖掘技术积累是开展课题的关键, 因而泰文舆情分析云系统平台的设计与开发是基于泰文跨境网络智能舆情研究相关课题展开的.

考虑到公安机关、安全部门和有关政府部门无法实时监控印刷品、存储介质和互联网中泰文舆情动态的难题, 系统运用云计算和舆情分析技术结合泰汉辅助机器翻译的方法, 实现多载体的泰文舆情分析云系统平台, 泰文监控结果由中文方式呈报, 能够使完全不懂泰文的用户实时监控泰文舆情态势信息. 通过研究网页泰文获取、文本构建预处理、智能倾向性分析、泰汉辅助机器翻译, 应用泰文识别、泰文编码转换、泰文自动分词等技术开发泰文舆情云分析平台.

3.2 平台简介

泰文舆情云分析系统平台由网络信息监测系统、网络信息管控系统组成, 其中网络信息监测系统的技术架构, 采用三层架构的形式, 即: 网络信息获取层、网络信息处理层、网络信息挖掘层、网络信息预警层. 如下图 1 所示, 仅给出了泰文舆情云分析平台前部分流程图.

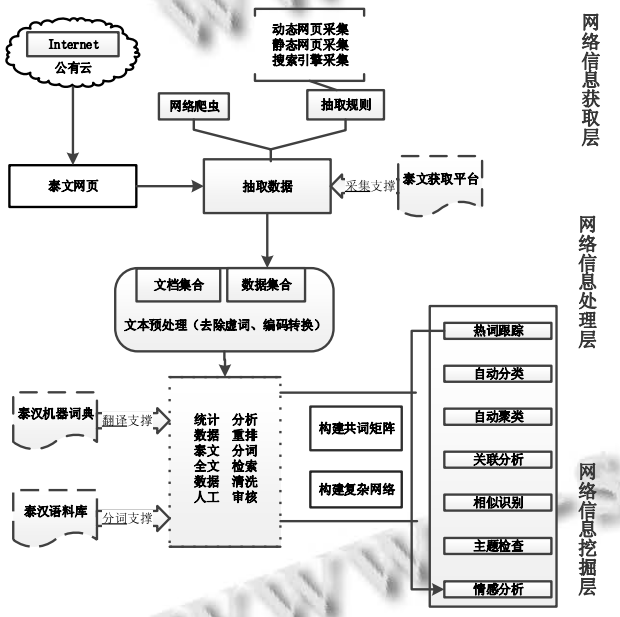

图 1 泰文舆情云分析平台部分流程图

其中网络信息处理层, 由泰汉机器词典-软件成果和泰汉语料库-数据成果作为重要技术支撑. 主要负责对文本做预处理, 基于这些元数据进行包括页面解析、去噪排重、内容提取、自动分词、翻译呈报等一系列数据处理工作, 完成舆情数据的初次加工.

从整体技术流程图可看出, 在网络信息获取层和网络信息处理层需要三个技术模块支持, 即: 泰文信息获取平台、泰汉机器词典、泰汉语料库, 分别对应泰文的采集、翻译、分词支撑. 因此本研究“泰-汉英互译有声语料库”是作为泰文网络舆情分析研究的成果之一.

4 文本库建设

4.1 泰语简介

泰文是拼音文字, 属汉藏语系壮侗语族壮傣语支. 单词由辅音、元音和声调组成. 泰语的辅音根据声调规律的不同分为中辅音、高辅音和低辅音; 元音分为单元音、复合元音和特殊元音. 其中单元音和复合元音中又根据发音时间的长短分为短元音和长元音, 长短元音在泰语中能区分词的意义. 特殊元音一般发音较短.

4.2 构建模型

语料的预处理的过程是将格式不一样的生语料进行加工, 形成统一格式(excel、xml、txt 等文本格式)语料库的过程[4]. 目前, 根据整理泰语课本收集, 在泰语专家和学者的指导下, 对汉英-泰互译语料的收集、整理及加工, 定义了一定的规范, 实现语料库的流程及模型构建如图 2 所示.

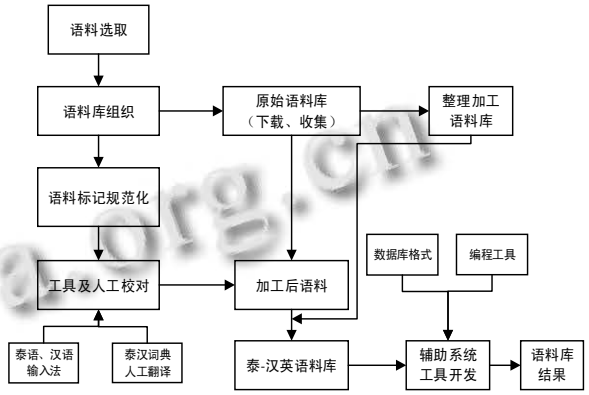

图 2 汉英-泰互译语料库构建模型

4.3 采集整理

语料库作为自然语言运用的样本,为计算语言学的研究提供了可靠的研究依据.由于语料库的巨大应用价值,各国都投入了大量的人力和物力建设语料库. 随着对蒙古语计算语言学的研究不断深入,对语言多种特性的定量研究需求越来越大,因此,我们提出研究和建设泰语语料库的课题, 由于民族语言文字的特殊性, 相关词汇词典的录入、标注是一项相当庞大、复杂, 并且时间周期长的工作. 鉴于语料库的建设是一

个浩大的工程,我们所讨论的语料库是基于大学教材的语料库.

鉴于国家目前提出, 对应于相关科学研究, 要培养“协同创新, 开放科学, 学科交叉”的研究体系. 本课题拟在工程中心机构的支持下, 结合我校东南亚语言学科优势, 发挥少数民族语言文字、自然语言处理、电子信息、计算机等文理交叉、多学科交融综合研究的特色, 开展项目合作. 本语料的建设过程中, 我校东南亚南亚语言文化学院和民族文化学院给与了极大帮助, 提供相关电子文稿, 并对语料的纠正、翻译、标注提供支持和帮助.

4.3.1 语料收集

语料库的原始语料的构成和取样一般来说按照明确的语言学原则并采取随机抽样方法收集的,而不是简单地堆积语料.在收集语料时,不仅要考虑到某一文类、体裁、语域、主题类型等的抽样比例,还要考虑到建立语料库的研究目的和具体用途.我们的语料库建设的主要目的是为语音处理、文字识别、文字分词研究提供支持.考虑到上述因素以及财力和物力所限,目前原始语料中收集了高等学校泰语专业系列教材之一包括:《大学泰语综合教程 1》(重庆大学出版社)、《大学泰语综合教程 2》(重庆大学出版社)、《大学泰语综合教程 1》(重庆大学出版社). 本课题的语料源主要来源于上述三本泰语教材, 尽管是比较权威的教材用书, 但也难免有疏漏的情况. 为了最大限度地提高基础词汇的覆盖率和可靠性, 增补部分主要来源于网上下载与整理, 筛选后综合收集约合 10 万条泰文(词汇、短语、句子)语料.

4.3.2 语料录入

原始语料是未经数字化的纸质文本.纸质语料的录入是建设语料库的第一项工作. 由于一部分语料是由教程 2 本册主编陆生(傣族, 精通西双版纳傣语, 泰国泰语等东南亚跨境民族语言, 并对泰傣语进行了多年研究, 目前在东南亚南亚语言文化学院承担泰语等科研教学工作)老师提供电子书稿, 电子语料的录入只需经过复制粘贴、归纳分类、汉英翻译等操作, 这极大的提高了语料录入效率. 还有部分语料是教程 1 纸质文稿形式, 纸质语料的录入效率耗时比较长, 操作全程需由泰语输入法和泰汉电子词典支持, 尤其对泰文字符的一一打印会比较繁杂, 因而纸质化文稿到电子化文本的转化是泰文语料库建设的核心工作.

泰文输入法是解决计算机中泰文输入、显示、存储的关键, 目前适用于泰文打印的输入法有多种, 如微软输入法(切换语言)、触宝输入法(国际版)等. 由于泰文组成及结构的特殊性, 泰文在国内仅支持直接输入, 不像彝文输入法创新型的可采用形似编码输入及中文输入法可采用五笔快捷输入, 因而输入工作需要懂泰文的专业人员(国际交流学院留学生 民族文化学院研究生)协助完成. 考虑到语料库的建设目标和泰文编码的问题,采用微软输入法来输入原始语料, 最终保存为后缀格式为.Word 的语料库.

4.3.3 词类划分

对语言采录所获取的原始文本数据, 进行信息化处理. 课题组运用输入法及电子词典相关工具和平台, 依次对词汇进行汉英翻译等、语义划分、词类划分、词性标注等操作. 分类是研究事物的基础.为了研究词汇,也必须对词汇进行分类. 如何对这 10 万多泰文词语进行分类, 实在是一门学问.可以从不同的角度进行分类. 比如, 可以从语义的角度进行分类, 把表示“物”的词语归为一类, 在“物”的大概念类之下可以划分为“生物”和“非生物”, “生物” 还可以进一步细分成“动物“和“植物”, 如此等等.

为便于语音录制分工和词汇查询分析, 开始对收集和整理来的 Word 电子文本做词汇分类, 如分成天文、地理类; 时间、方位类; 动物类; 植物类; 身体、部位类; 房屋、建筑类; 服饰、织物类; 工具、用品类; 日用品类; 文化、娱乐类; 饮食类; 动作、行为类; 性质、状态类; 人物、称谓类, 数量类等 14 大类.

4.3.4 词性划分

但泰文文本切分词典的研究领域是语法而不是语义,仅仅是语义分类是不够的, 还需要在语法的范畴内进行词类的划分. 关于词性划分, 泰国学者划分的规则不同, 划分的类别也不同. 主要有如下几种方式:

(1) 仅从词义出发, 泰国学者鸟巴吉·辛拉巴汕, 将词类划分 7 类, 分别为:名词、动词、代词、修饰词、介词、连词和叹词[5];

(2) 从功能和词义出发, 泰国学者娜瓦婉, 潘图美拉, 提出按意义来划分词类是比较科学准确的, 但问题如果按意义划分过于繁琐, 有可能会划出几十种类别, 同时有些词词义特殊难以归类[5]. 可根据词汇功能先分 6 大类, 感叹词、根词、代词、修饰词、关系词、补充词, 后又按其意义在大类下面又划分小类.

(3) 从功能和词义出发, 相关泰语学者提出, 针对大规模词汇库, 二级划分会增加词类划分复杂度. 因而泰国教育部支持编写教材(基础泰语系列丛书)出台新标准[5], 不再区分大小类, 可分为 12 类, 分别为:名词、代词、动词、助动词、修饰词、数量词、指示词、介词、关联词、结尾词、叹词、否定词.

考虑到整个语料库的词类划分工作庞大, 而这又是舆情工作(文本分词)的需要. 为便于文字识别和理解, 课题组还是以泰文翻译后的中文为词类划分对象, 这会极大的方便非泰语专业人员, 提高科研工作效率. 因此我们参照北京大学的“现代汉语语法信息词典”中的功能分类思想, 对泰文词从语法角度分为以下 12 个大的基本词类(后面括号中的字母是各个词类的代码): 名词(n)、形容词(a)、动词(v)、数词(m)、量词(q)、副词(d), 代词(r)、介词(p)、连词(c)、助词(u)、语气词(i)、叹词(e).

我们除了对基础词汇进行分类外,还对特殊词汇进行了分类和标注. 特殊词汇分为成语(v)、习用语(!)、简缩语(j)和符号(b)等,同样对这类词汇进一步的分类和编码, 部分泰文词汇分类表如表 1 所示.

表 1 部分泰文词汇分类表

6、人物、称谓类				
序号	**词条**	**词性**	**英语**	**汉语**
6-00001	มองโกล	n	Mongolian	蒙古族
6-00002	มองโกเลีย	n	Mongolia	蒙古人;蒙古
6-00003	มองโกเลียน	n	Mongolian	蒙古人
6-00004	บรรดาศักดิ์	n	the conferred title of rank of the nobility;	贵族的等级被赠予的名称;
6-00005	ประยูร	n	family; clan; household; house; kin;	家族
6-00006	ผู้ดี	n	gentleman;lady;a respected;refined or genteel person;nobleman;aristocrat	绅士;淑女;尊敬;精炼的或有教养的人;贵族;贵族
6-00007	พงศ์	n	[pertaining to] history	后裔;家族
6-00008	มอญ	n	Mon;Pegu;Peguan;an ethnic group of Burma and Thailand	孟族
6-00009	มเซอ	n	Musser (Lahu);a tribe of hill people;	慕瑟;山上的居民种族;
6-00010	ลอร์ด	n	noble; nobility; aristocrat; nobleman; baronage;	贵族
6-00011	ลื้อ	n	a Thai ethnic minority peoples from South China and North Thailand ; used when talking to a male friend	你;泰仂族;来自中国南部和泰国北部的一个泰国的人种少数的民族;熟悉的代名词;;
6-00012	ศากย	n	Sakya;an ancient tribe of India	一个印度的远古种族
6-00013	สกุล	n	family;lineage;clan	姓氏;宗族;贵族;体系

5 语音库构建

5.1 泰语特点

泰语为东亚语系之一, 是傣泰民族使用的语言. 泰语(ภาษาไทย), 旧称暹罗语(Siamese), 泰国的官方语言, 属汉藏语系壮侗语族壮傣语支. 使用人口约 5000 万. 有中部、北部、东北部和南部等 4 个方言区. 曼谷话是泰语的标准语. 泰语是一种分析型、孤立型语言, 基本词汇以单音节词居多, 不同的声调有区分词汇和语法的作用, 构词中广泛使用合成和重叠等手段[6].

词是音义的结合体, 语音是语言的外在形式, 也是词的外在形式. 研究泰语的构词体系, 同样也离不开词的语音特点. 课题组研究的泰语, 但其音系分支相当复杂(中辅音 9 个、高辅音 10 个, 低辅音 23 个), 发音较为复杂容易错音甚至偏音. 因而泰文语料库的录音过程都需要专业录音软件、录音系统、录音环境. 录音人员的支持, 都需要按照拟定的录音计划.

5.2 录音软件

我们采用泰语文字和标注序号作为泰语录入的参考依据, 录音系统采用前期工作开发的傣语原音录音

软件一套, 只是录音对象不同而已, 只需调节相关录制参数. 软件界面如图 3 所示

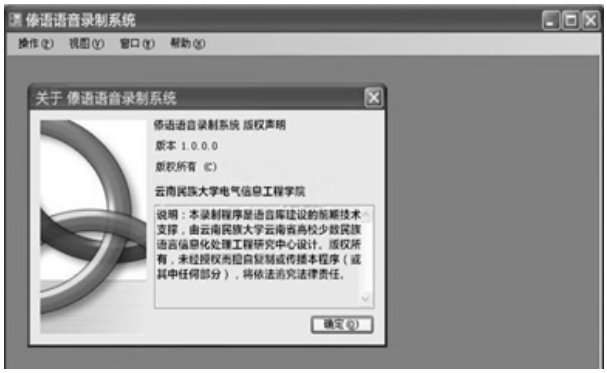

图 3　傣语语音录音系统界面

5.3 录音系统

(1)环境配置

课题组在工作室进行录音, 并在三面布置尖劈, 录音时的混响问题和噪声干扰对录音的影响均得到了很好的解决. 录音场地被划分为录音室、监听室. 播音员在录间室中录音, 监听室里有一至两名监听员现场监听.

(2)录音器材

在录音室中放置有: 显示器、鼠标、键盘、话筒、耳机, 声门阻抗仪(如需获取声门信号时, 需要使用该仪器). 在监听室中放置有: 两台电脑、放大器、Kay、小型调音台、话筒、耳机, 其中录音室中的显示器和监听室的主控电脑同屏.

5.4 录音人员

录音, 首先考虑的应该是准确性与可靠性的问题, 其次是通用性的问题. 在课题组邀请到汉语国际教育专业的泰国籍留学生(ธนินี　ลิ่มอรุณ)来录音. 该学生的汉语水平考试(HSK)已过三级, 录音工作沟通比较容易, 其家乡所在地为泰国首都曼谷, 方言(曼谷话)属于泰语的标准语, 能够保证泰语录制的原汁原味. 目前所有泰语的原音录入均由她完成的, 我们的录音与所推广的发音达到一致性, 在语音库的建立过程中融入地道、纯正、浓厚的泰语语音和语言文化.

5.5 后期处理

在安静的实验室里完成了整个原音录制过程, 最后通过对话质的筛选, 留下了完整的傣语语音文件, 将其存储为后缀为.WAV 的格式, 目的在于 WAV 是被支持得最好的音频格式, 所有音频软件都能完美支持. 且本身也可以达到较高的音质的要求, WAV 也是音频录制的首选格式, 适合保存原音素材, 同时还适合不同设备使用的音频方式转换[7]. 后期由专业人员对原始语音文件进行检查并帅选, 对有声音干扰(如咳嗽声)的音频应将冗余部分删除或补录, 对有声音断续的部分可采用合并操作, 并最终按序号打包成泰文语音库. 至此, 课题组建立了泰语语音库, 使泰文不仅能看得到而且听得到, 丰富语料库的有声功能.

5.6 成果展示

为便于录音分工和查询分析, 我们参考泰文词汇表将泰语录音库分天文、地理类; 时间、方位类; 动物类; 植物类; 身体、部位类; 房屋、建筑类; 服饰、织物类; 工具、用品类; 日用品类; 文化、娱乐类; 饮食类; 动作、行为类; 性质、状态类; 人物、称谓类, 数量类等 14 大类[8], 共约 5 万个音频词汇及短语. , 泰语音频录制后的成果如图 4 所示.

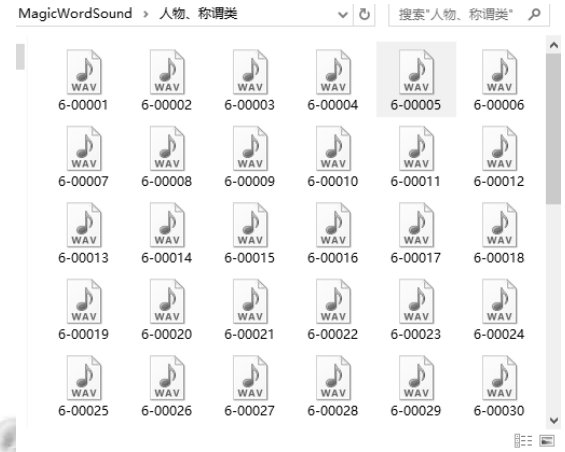

图 4　泰语音频录制成果展示

6　数据库建设

6.1 方式选择

现在流行的关系型数据库有 IBM DB2、Oracle、SQL Server、SyBase、Informix、MySQL、Access 等. 本语料库表结构是用 Microsoft Access 2003 数据库实现的[9]. Access 提供一种建立表与表之间“关系”的方法, 词表、音频的数据之间都存在这种“关系”, 这种关系将数据库里各张表中的每条数据记录都和数据库中唯一的主题相联系, 只需通过一个主题就可以调出相应的数据库来使用[10].

VisualC++6.0 提供了访问 ODBC 的接口, 不用考虑数据库的交互细节, 并对 Access 数据库提供了更好

的支持[11]. 用 VisualC++6.0 通过 ODBC 访问 Access 数据库提高建设语料库的效率, 为大量的统计工作带来很大的方便. 用 VisualC++6.0 的数据库编程和 Access 数据库管理功能来管理语料库可以兼顾自动和手工的工作, 让自动和手工工作有机地结合起来, 使语料库的管理变得简单、易用、开放.

6.2 结构制定

即确定数据表中的关键词,这些关键词大致包括: 词条序号、泰文词条、泰文词性、英语翻译、汉语翻译, 在数据表中分别用英文简称代表 wod(Word)、sph(Speech)、eng(Enghlish)、chi(Chinese). 文本数据表的查询也是通过关键词来实现的, 关键词的选择语料库的查询起到关键性的作用[12]. 本课题的语料库的研究对象是泰语文本, 故课题组将“wod”关键词作为数据表的主键.

6.3 存储形式

根据不同的元素(文本、语音、视频、图像等)属性, 课题组采用了两个数据库来记录泰-汉英有声语料库, 分别对应了文本数据库和语音数据库, 两个数据表是通过关键词“num”来查询联系的, 而”num”的唯一性决定了文本数据表中泰文词条对应的语音数据表中的泰语发音. 文本数据库, 收录了语料库中的词汇、短语、词性、汉语、英语 5 项文本元素, 语音数据库, 收录了语料中的发音 1 项音频元素.

Tai-Corpus

num	wod	s	eng	chi
6-00019	ตา	n	eye(s);mater	眼睛;外祖父;
6-00020	ทวด	n	great-grandp	曾祖父母;外
6-00021	ทวดน้อย	n	great-grandu	大师-伯叔祖;
6-00004	บรรดาศักดิ์	n	the conferre	贵族的等级被
6-00022	บรรพบุรุษ	n	eleg;ancesto	祖宗;祖先
6-00023	บุพการี	n	parent;ances	父母;祖先
6-00005	ประยูร	n	family; cla	家族
6-00024	ปู่	n	grandfather;	祖父
6-00006	ผู้ดี	n	gentleman;la	绅士;淑女;一
6-00007	พงศ์	n	[pertaining	后裔;家族
6-00002	มองโกเลีย	n	Mongolia	蒙古人;蒙古
6-00003	มองโกเลียน	n	Mongolian	蒙古人
6-00001	มองโกล	n	Mongolian	蒙古族
6-00008	มอญ	n	Mon;Pegu;Peg	孟族
6-00009	มูเซอ	n	N;Musser (La	慕瑟;一个山_
6-00010	ลอร์ด	n	noble; nobi	贵族
6-00011	ลื้อ	n	Lue;a Thai e	你;泰仂族;来
6-00012	ศากย	n	Sakya;an anc	一个印度的远

图 5 泰-汉英数据库部分展示

7 总结

本成果采用人工标注和机器翻译相结合的方法来建立的词表数据库, 设计实现了泰英-汉对照互译、泰语配套发音、泰语词性查询等常用功能, 同时支持对语料库的词类划分操作. 汉英-泰语料库的建立, 其成果特点在于, 一方面可从中筛选部分(如政治、军事、色情、暴怖)等词汇, 建立泰文敏感词汇库, 供后续舆情预警中对敏感词的查询、匹配、预警和过滤模块使用; 一方面可作为泰-汉电子词典制作的数据库来源, 完成对泰语的在线本地翻译, 方便舆情分析以中文形式呈现.

由于民族语言文字工作和研究的特殊性和复杂性, 大规模语料库的建设是一项周期长且复杂的工作. 在采集、收录、翻译的过程中难免会出现纰漏, 本语料库数据表还未授权网页下载或制作成检索平台, 因而还有较多改进和完善的地方.

参考文献

1 林政.Web 双语平行语料自动获取及其在统计机器翻译中的应用[硕士学位论文].天津:天津师范大学,2010.

2 江涛,江静,戴玉刚,李艾林.藏文舆情云分析系统平台研究.信息网络安全,2014,9:92–94.

3 才藏太,华却才让.藏语语料库加工和处理用的藏文切分词典的建立与设计.中国中文信息学会、中国科学院软件研究所、青海师范大学、五省区藏族教育协作领导小组办公室.第十届全国少数民族语言文字信息处理学术研讨会论文集.中国中文信息学会、中国科学院软件研究所、青海师范大学、五省区藏族教育协作领导小组办公室.2005,6.

4 李绍哲.俄语语料库和基于语料库的语法研究[硕士学位论文].哈尔滨:黑龙江大学,2012.

5 何冬梅.泰语构词研究[硕士学位论文].上海:上海师范大学,2012.

6 韩金玲.汉泰名词性短语语序对比研究[硕士学位论文].南宁:广西大学,2014.

7 蔡莲红,赵世霞.汉语语音合成语料库的研究与建立.语言文字应用,2013,S1:175–180.

8 王成平.信息处理用彝、汉、英三语平行语料库的建设与语料对齐技术研究.科技通报,2012,2:131–133.

9 常宝宝,詹卫东,张华瑞.面向汉英机器翻译的双语语料库的建设及其管理.术语标准化与信息技术,2003,1:28–31.

10 才让加.面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究.中文信息学报,2011,6:157–161.

11 姚树杰.面向统计机器翻译的语料处理与评价技术研究[硕士学位论文].沈阳:东北大学,2011.

12 孔迎春.纳西汉语双语语料构建及智能输入法研究[硕士学位论文].昆明:昆明理工大学,2013.