

The Application of Multi-modal Discourse Annotation and Analysis Tool ELAN in the Construction of Cross-border Endangered Chinese Dialect-Dungan Corpus

Lian Jing^{1,a}, Kang Xiaoming^{2,b*}

¹College of Foreign Language and Literature, Northwest Normal University, Lanzhou, Gansu, China

²College of Chinese Language and Literature, Northwest Normal University, Lanzhou, Gansu, China

^alj885046039@163.com

^{b*}Kangxm1370105705@163.com

Abstract

Dungan Language is unique among the languages in the world. It is unique in the sense of linguistic typology and has been paid more and more attention by the academic circle at domestic and overseas. This paper takes ELAN, a multimodal discourse annotation and analysis tool, as the research tool, and the Dungan acoustic data as the research object, and introduces in detail how to build a Dungan voiced corpus based on ELAN, in order to provide data support and language information for Dungan research at home and abroad, and promote the level of Dungan Language research.

Keywords: ELAN, Dungan Language, voiced corpus

多模态话语标注和分析工具 ELAN 在东干语有声语料库建设中的应用

廉静^{1, a} 康小明^{2, b*}

¹西北师范大学外国语学院, 兰州, 甘肃, 中国

²西北师范大学文学院, 兰州, 甘肃, 中国

^alj885046039@163.com

^{b*}Kangxm1370105705@163.com

摘要

东干语在世界语言之林中独树一帜, 具有语言类型学意义上的独特性, 日益受到国内外学术界的重视。本文以多模态话语标注和分析工具 ELAN 为研究工具, 以东干语有声语料为研究对象, 详细介绍如何基于 ELAN 建设东干语有声语料库, 以为海内外东干语研究提供语料数据支持和语言信息服务, 促进东干语研究水平的提升。

关键词: ELAN, 东干语, 有声语料

1. 引言

东干语是中亚东干族所使用的语言, 它是以我国近代汉语西北方言基本词汇和语法结构为主体, 同时融合和吸收了中亚地区其他民族常用的外来借词和

语法结构的汉语域外变体^[1], 在语言类型学上的具有重要地价值, 有丝绸之路上的“语言奇葩”之称。

东干文是东干族学校教育、文学创作、报刊出版的官方文字形式, 对于提高东干民族的教育水平、保护和传承东干族独特的语言和文化起了重要的作用。

但在中亚国家独立以后,随着各国主体民族语言意识日渐上升和族际交际用语“俄语”的双重挤压下,东干语逐渐被边缘化,具体表现为:青少年一代中会说东干语的人口数量在急剧下降;以俄语为主的“双语人”或“多语人”数量激增;东干语在东干人的语言使用选择的地位日渐式微,正在被其他语言排挤出社会生活的各个领域,并且由社区语言向家庭语言的方向在发展,这些现象都促使东干语朝着濒危语言的方向发展。因此,保护这一宝贵的跨境汉语资源已是当务之急^[2]。

2. 语料库建设与语言资源保护

国际上有名的语料库基本上都是基于语言的标准变体,而有关非标准语言变体的语料库却相对较少,比较有名的诸如:FRED 英语方言语料库、北欧五国方言语料库(Nordic Dialect Corpus)、阿拉伯语方言平行语料库(PADIC, A Parallel Arabic Dialect Corpus)等,这些方言语料库的建设都在一定程度上提升了方言的研究水平。

近些年来,我国的方言语料库的建设也有较快的发展,基于方言调查的方法与技术,在语音、词汇、语法等方面都开发了专门的语料库^[1],但我国的方言语料库开发与建设还存在诸多方面的局限性,主要有:

(1) 方言语料库范围主要限于国内方言,尚未见到面向跨境汉语方言(如东干语)的语料库;(2) 语料库规模较小,收词量偏少,可开展历时和共时对比研究的大规模语料库不多;(3) 语料库结构比较简单,缺乏全面和规范性的标注,检索功能也不够强大;(4) 语料获取手段单一,主要依赖方言调查手段获得数据,自然情景下的真实语料较匮乏;(5) 很多方言语料库只有单机版系统,以光盘介质保存,缺乏开放访问和资源共享的网络版本。

国家语委于2008年启动了“中国语言资源有声数据库建设”,系统调查收集当代汉语方言、少数民族语言和带有地方特色的普通话的实态、有声语料,并进行科学整理、加工和有效保存,对于保护我国的语言资源起到了极其重要的作用^{[3][4]}。

通过语料库技术,以及基于语料库的语言研究、词典编撰和教科书编写等衍生功能,可以完整、全面、客观地记录和保存东干语全貌,保护东干汉语语言资源,“抢救”这一跨境濒危汉语方言,同时也有利于服务国家“一带一路”建设战略。

2021年,世界首个东干语语料库在武和平教授团队的努力下已完成,基于第四代语料库分析工具CQPweb建立了网络版东干语语料库,该语料库分为单语语料库和双语语料库,可以实现单语和双语的检索。但随之而来的问题是,当我们想对东干语的有声语料进行保护和研究时,该如何基于有声语料建设语料库呢?因此本文以ELAN为研究工具,尝试解决该问题,以期对东干语研究添砖加瓦。

3. 语料内容及多模态话语标注和分析工具 ELAN

中亚东干语有声语料库主要包括“音频”“视频”两部分,其中“音频”包括“朗读、广播”两大类,而“视频”又分为“纪录片、对话、诵读、歌曲”四小类。经过武和平教授团队的收集,目前已形成包含1976分钟(计32小时56分钟,仍在持续增长)的东干有声语料。但如果缺乏数据库管理,没有检索功能,那么该语音语料也只不过是些录音材料,无法对这些有声资源进行深度挖掘和深入研究。

目前用来对音视频语料进行转写或标注的软件比较多,主要有:Anotation Transcriber、Praat、Anvil、ELAN、EXMARaLDA、TASX等。但Praat不能分析视频文件,软件界面也无法显示较长音频的波段;Anvil仅在Windows上运行良好,在Mac和Linux上的运行情况并不理想;EXMARaLDA的一体化程度不够高,其分割、转写、标注等功能都需由独立的工具来承担,对新用户的使用造成了一定的困难^[5]。

基于此,本研究选用荷兰Planck心理语言研究所开发的多模态话语标注和分析工具ELAN对收集的东干语有声语料进行转写和标注。相较于其他软件而言,ELAN是一个免费的开源软件,支持Windows、MacOS、Linux三个操作系统,它可以使音视频、转写文本、标注内容、数据库等内容做到多位一体,ELAN可以生成已转写并标注过的.eaf文件,方便查看及检索,还可以生成.html格式的网页版文件,能同时显示音频和转写标注过的文本。此外,ELAN还自带强大的检索查询功能,可以在指定范围内实现对关键词语的检索。

4. ELAN 在跨境濒危汉语方言—东干语有声语料库建设中的应用

湖南师范大学教师李斌从4.11版起就对ELAN进行简体中文的汉化,其汉化文件已经加入ELAN官方安装程序。登录ELAN官方网站<https://archive.mpi.nl/tla/elan>,选择适用于Windows系统的最新版本6.2版进行下载。

4.1. 导入多模态语料及添加层

打开ELAN 6.2,点击“文件”,选择“新建”,选择需导入的音频,点击“确定”后完成音频文件的导入。

在ELAN中导入音频文件后,系统默认进入到“标注模式”,在这个模式下需要改变并添加所需的层名及属性。ELAN会给导入的文件自动添加一个缺省层,层名为default,需添加一个新的层类型“Dialect”,原始类型为None不变;接着再添加两个新类型,类型名称分别为“dungan”(用于东干语转写)和“Pth”(用于普通话转写),原始类型为Symbolic

Association。接着将层名由“default”修改为“方言转写”，层类型由“default-It”修改为“Dialect”。

4.2. 多模态语料分割、转写及标注

将模式切换为“分割模式”，分割方式选择“一个标记确定一个标注段（相邻的标注段）”，在该方式下，我们只需按下“Enter”键即可确定一个时间点（如图1所示）。如果分割的范围过长或者过短，都可以再次进行调整，可以在分割模式下多次进行检查是否有误，确定无误后，就可以切换到“转写模式”。

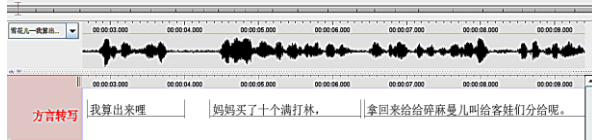


图1 在分割模式下分割段

接着再依次完成多模态语料的转写（如图2所示），转写完成之后，再切换到标注模式。我们根据语料库建设所需，对东干语词汇的语源信息进行了标注——即特定的东干词汇来源于哪一种语言，如阿拉伯语、俄语、土耳其语、吉尔吉斯语等；同时，也在语料库的元信息中对语料的语体进行了标注，以方便比较不同语体特征下东干词汇的特点。

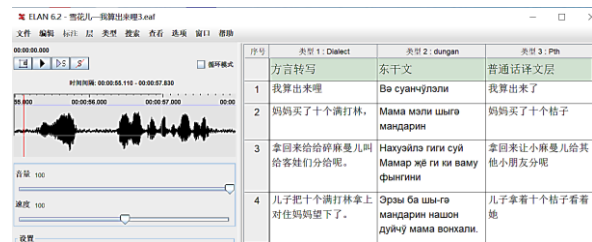


图2 有声语料的逐句转写

4.3. 转写标注内容的输出及建立语料库

ELAN 可以把转写标注过的内容输出为.txt 文本文件，也可以把转写标注的内容输出为.html 格式的网页版文件（如图3所示），上方是该文件的音频，下方按照“方言转写—东干文—普通话译文”的顺序排列，当点击方言转写一行的蓝色字体时，系统会自动播放该句的音频（需将输出的网页版文件 and 该文件的音频放在同一文件夹下，系统才能正常播放）。



图3 网页版文件的输出

将所有东干语的有声语料都在 ELAN 中逐一认真转写并标注，生成对应的.eaf 文件。用 ELAN 建立语料库就是将相关的.eaf 文件导入进来，然后指定检索范围^[6]。ELAN 指定的检索范围较为灵活，可以是单个、多个或所有的.eaf 文件，根据需要灵活设定。根据研究的需要，中亚东干语有声语料库暂设定了两类检索范围，第一类是通过网络或其他渠道收集的语料，第二类是通过远程田野调查收集的自然话语语料。

4.4. 关键词检索

例如在“网络或其他渠道”这个范围中搜索“BA”这个东干语高频词，在搜索框中输入“BA”，需把“正则表达式”勾选上，再点击“搜索”按钮，这时在这个检索范围内全部含有“BA”的句子就会被检索出来，在这个词之前以及之后的句子也会显示出来（如图4所示），可以看出“BA”在“标注”这一列中是加粗显示的，可以将该检索结果输出为.txt 格式的文件，就可以很方便看到包含“BA”这个词的所有例句。若要检索双词或三词搭配也是按照上述步骤进行操作。



图4 “BA”在 ELAN 中的检索结果

5. 结语

通过多模态话语标注和分析工具 ELAN 在建设中亚东干语有声语料库中所起到的重要作用可以看出，ELAN 对于有声语料的处理较为方便、快捷，对于音频或视频语料的转写、标注以及检索都可以在一个软件内完成，而且也易操作，最终能输出网页版的有声语料库文件，方便后期查阅。方言研究者在学会 ELAN 基本的操作后，进行几次语料转写和标注的练习后，就可以很熟练地掌握 ELAN 的基本操作，方便快捷，这对于方言研究有非常重要的意义。

本文以多模态话语标注和分析工具 ELAN 为研究工具，以东干语有声语料为研究对象，从 ELAN 的下载，到有声语料的分割、转写及标注，再到输出和关键词的检索，层层推进，详细介绍了转写和标注有声语料以及建设有声语料库时的步骤，以期海内外的东干语研究添砖加瓦，为海内外东干语研究者搭建一个数字化网络资源共享和研究平台，同时也有利于服务国家“一带一路”建设战略和东干族文化传承民心相通工程。

REFERENCES

- [1] Wu, H.P., Zhao, H.G. (2018) Promoting the

- construction of Dungan Language corpus. *Chinese Social Sciences Today*, 4:2.
- [2] Zhou, Q.S. (2020) 'The Belt and Road' catcher: language adaptation and language inheritance of the Dungan people. Nankai University Press.
- [3] Li, Y. M. (2010) On the establishment of the audio database of Chinese language resources. *Chinese Language*, 4:356-363+384.
- [4] Wang, T.K. (2010) Language planning based on the conception of language resources-An exemplification of "monitoring study of language resources" and "construction of the corpus of vocal data of Chinese language resources". *Journal of Shaanxi Normal University (Philosophy and Social Sciences Edition)*, 6:58-66.
- [5] Liu, J. (2019) A comparative study of four software functions for multimodal interpreting corpus construction. *Comparative Study of Cultural Innovation*, 3:63-64.
- [6] Li, B. (2012) Constructing multimedia dialect corpus with ELAN. *Dialect*, 2:178-190.
- [7] Li, B., Gao, G.A. (2012) Constructing a PSC standard sample database in Hu'nan with ELAN. *Applied Linguistics*, 3:126-133.
- [8] Lin, T. (2005) Dungan Writing-A successful trial of Chinese Alphabetic Writing. *Journal of The Second Northwest Institute for Ethnic Minorities*, 4:33-38.
- [9] Lin, T., Hui, J. D. (2015) The present situation and development direction of Dungan Language and culture research. *Tangut Research*, 3: 121-124.
- [10] Lv, H. H., Liu, H. (2014) The establishment and utilization of the Chinese Sign Language corpus based on the ELAN software. *Chinese Scientific Journal of Hearing and Speech Rehabilitation*, 12: 298-301.
- [11] Cao, G. Y. (2008) Language resources of China. *Journal of Zhengzhou University (Philosophy and Social Science Edition)*, 1: 7-8.
- [12] Cui, G., Sheng, Y.M. (2000) Labeling of linguistic materials in the linguistic material data-bank. *Journal of Tsinghua University (Philosophy and Social Sciences)*, 1: 89-94.
- [13] Hai, F. (2006) On the Linguistic Value of Dungan Language in Central Asia. *Journal of Xinjiang University (Philosophy, Humanities & Social Sciences)*, 3:146-150.
- [14] Hu, Z. H. (2004) About "Dungan", "Dungan Language", "Dungan Bilingual" and "Dungan Studies". *Language and Translation*, 1: 12-14.
- [15] Wang, L.F., Wen, Y. (2008) Using multi-modal approach in applied linguistics research. *Technology Enhanced Foreign Language Education*, 3:8-12.
- [16] Xiao, Z.H. (2017) On the former digital dialect data processing by EXMARaLDA. *Dialect*, 39:496-502.
- [17] Zhang, Z.H., He, M., Han, Z. (2014) Construction and application of college English multimodal-corpus. *Shandong Foreign Language Teaching Journal*, 35:50-55.