Advances in Social Science, Education and Humanities Research, volume 555
Proceedings of the 1st International Conference on Education:
Current Issues and Digital Technologies (ICECIDT 2021)

Research on the Construction of Multimodal Corpus of Tibetan Teaching
Taking Gannan Tibetan Middle School as an Example

Dawa Pengcuo[1,*] Daojie Ben[1]

[1] *Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu, China*
[*]*Corresponding author. Email: 742837195@qq.com*

ABSTRACT
The multimodal corpus of Tibetan teaching uses video and audio documents as the corpus, which truly reflects the teaching activities and contents; furthermore, it has great potential in promoting Tibetan teaching and researches. Compared with the textual corpus, the multimodal corpus of Tibetan teaching has the advantage of the authenticity and the timeliness of the corpus. This article mainly discusses the construction of multimodal corpus, the corpus collection, the segmentation of annotation, the tools of database construction, and the retrieval and replacement, etc. This article analyzes the application value of multimodal corpus in Tibetan teaching and puts forward the applied countermeasures in teaching domain.

Keywords: *The multimodal corpus, Elan, The Tibetan teaching.*

1. INTRODUCTION

With the rapid development of IT, teaching methods based on multimodal corpus are also developing rapidly. Multimodal corpus refers to the integration of the audio, video, text corpus, and other information. Researchers can research corpus by way of multimodal processing, retrieval and statistic [1]. At present, various large multi-modal Corpora have been established at abroad, and they have made remarkable achievements in corpus collection, processing, labeling, standardization and analysis. Multi-modal corpus construction and related researches are in the ascendant in China as well, especially in the application and research of multi-modal corpus in foreign language teaching. Compared with the traditional text corpus, the multi-modal corpus can reflect the original appearance of teaching activities in a more comprehensive, objective and detailed way, and greatly improve the application of teaching ideas and methods. Therefore, it has broader academic value and application prospect in teaching.

Gannan Tibetan Autonomous Prefecture (GTAP) belongs to Gansu Province. It is located at the junction of Gansu, Qinghai and Sichuan provinces between the Qinghai-Tibet Plateau and the Loess Plateau. The autonomous Prefecture has jurisdiction over the seven counties of Xia-he(bla rang), Luqu(klu chu), Maqu(rma chu), Diebu(the bo), Zhouqu(brug chu), Lintan(ba tse) and Zhuoni(co ne), as well as the city Hezuo(gtsos), with a total of 99 towns and 664 villages, covering an area of 45,000 square kilometers and a population of 730,700, of which Tibetans account for 54.2%.(2011's. Ma qu County Annals) Tibetan is one of the languages widely used in GTAP, most Tibetans begin to learn Tibetan from primary school, but in the new era, the deficiencies of the traditional teaching methods are becoming more and more obvious. Therefore, we planned to harness the Elan software to build a multi-modal corpus for middle school students, who can get the needed teaching materials, thus they can use the language better, at the same time, it could improve the quality of language teaching.

*Fund: Supported by the Humanities and Social Science Fund of Ministry of Education of China (Grant No.19YJC740062).

2. CREATION OF MULTIMODAL CORPUS

Compared with the text teaching corpus, the process of the construction of the multi-mode teaching corpus is pretty boring, and the construction of the corpus requires a huge amount of manpower and material resources. Multi-modal corpora requires a lot of manpower in different aspects, such as collation, segmentation and multi-level labeling, etc., and more skilled researchers still need spending about 50-80 hours to complete the 1-hour corpus. The available public multi-mode library building tools include Annotation Transcriber, Audacity, Praat, Aboboo, Anvil, Elan, and so on. Elan has a powerful database function, and it has its own functions such as retrieval, replacement, and statistics [2]. Therefore, we prefer to create a multi-modal corpus for Tibetan teaching corpus through Elan. The process of the construction of multi-modal corpus is as follows: (1) speakers and recording equipment; (2) corpus' selection; (3) creation and import of metadata; (4) segmentation, transliteration and word segmentation of corpus; (5) playback and retrieval of multimodal corpus under Elan environment.

2.1 Speakers and Recording Equipment

We selected three native speakers here who have been living in GTAP, using fluent Tibetan with standard and clear pronunciation. They are selected for our recording. They have been accumulated enrich experience for Tibetan teaching for years, and they all got bachelor degrees. The classes they taught were the first, second and third grade respectively. There are 32 students in the first grade, 36 in the second grade and 40 in the third grade, all of whom come from GTAP.

The recording video equipment is SONY HXR-NX100 camera, the video resolution is 1920*1080, the editing software is Premiere Pro CC, and the format is saved as "*.mp4". The recording equipment is SONY ICD-UX575F recorder, and the recording software USES Adobe Audition1.5. The audio sampling rate is 22050Hz, mono channel, sampling accuracy is 16BT, and the format is saved as "*.Wav". Elan software was used to observe the recording effect and finally conduct the processing of the corpus in Elan.

2.2 Corpus' Selection and Classification

According to the contents, the objects and the environments of the Tibetan language teaching in middle schools, the selection of the small multi-modal corpus should be followed the representative principles, meanwhile consider the scene of teacher-student communication, such as the teaching class in the grade seven, eight, and nine. In addition, the dialogue between teachers and students should be close to life and fully consider the scene of language communication. According to the above principles, this corpus mainly collects two aspects: first, the audio and video materials of classroom teaching recorded by me myself; the second is the ready superior CD on Tibetan teaching downloaded on the internet, as shown in "Figure 1":

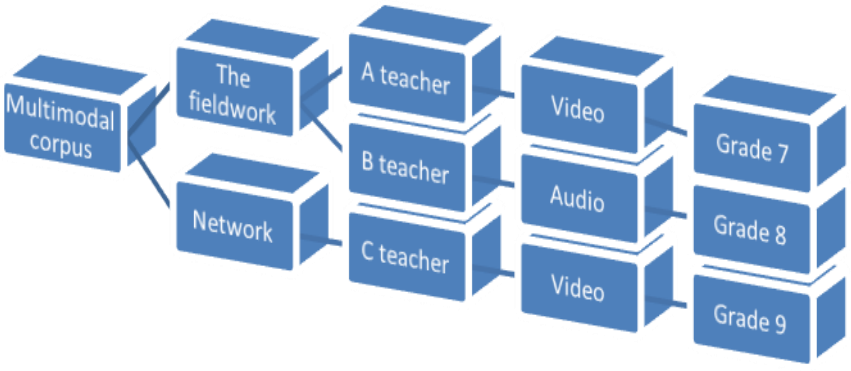

Figure 1 Corpus sources.

2.3 Creation and Import of Metadata

Elan's official supply the website https://archive.mpi.nl/tla/elan, you can download Elan software for free. It's easy to install directly after downloading it, and then going into operation interface, then finding the "new" button, importing the original multimodal Corpora. During the

operation, different tiers are divided according to different contents, and different contents are transcribed and marked by different layers, such as video layer, voice layer, part of speech layer, translation layer, cultural layer and so on. We divides it into five levels, namely, title, Tibetan, Chinese, body language and teaching content, as shown in Table No. 1 according to the framework of Tibetan teaching contents in middle schools and the actual situation of teaching activities, as shown in "Table 1":

Table 1. Annotation layer

Tier	Linguistic type	Annotation content
1	Title Tier	Title name
2	Tibetan Language Tier	Tibetan Annotation
3	Chinese Language Tier	Chinese Annotation
4	Action Language Tier	Action Annotation

2.4 *Segmentation, Marking and Transliteration of Corpus*

The layers should be set up after importing the multi-modal corpus. Firstly, it segments the multi-modal corpus into natural sentences as the smallest unit, and marks them. Then, it needs to find out the "Segmentation mode" in the drop-down menu of "Options", and then divide the corpus by the shortcut key "Enter". After that, we can transcribe the different annotations. There are three steps in the marking process: the first step is to transliterate the Tibetan language of teaching videos; the second step is to transfer the Teaching contents of Tibetan language with Chinese characters. The third step is to mark non-verbal corpus with special symbols or text messaging and methods, for instance, body language: "nod" marked as "HD", "shake" marked as "HS", "clap" marked as "HP", "smile" marked as "Hm" and so on. These special symbols have the same characteristic in search ability and consistency, which will make the later research and analysis fast and convenient. When labeling is finished, then saving the corpus, then outputting the corpus in a certain format (TXT, TextGrid, Eaf), after that we can adopt Explorer8 fieldwork language exploration software for word segmentation, and then importing them into the Elan, do the further analysis, finally the file should be saved as a file *. Eaf, it will be of great help to retrieval or analysis the corpus. The hierarchical labeling model diagram of multimodal corpus is shown in "Figure 2":

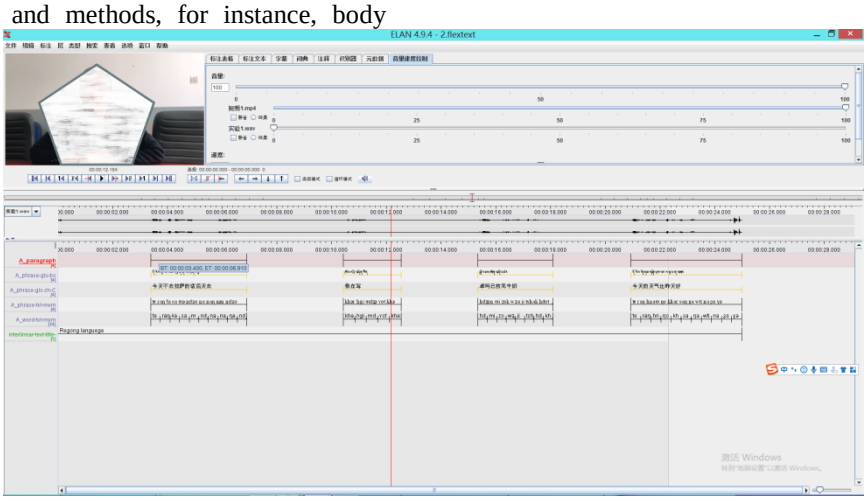

Figure 2 Hierarchical labeling model diagram of multimodal corpus.

2.5 *Retrieval and Playback of Multimodal corpus Under Elan Environment*

It is a time-consuming and tough task to annotate the transliteration of corpus, also the core link of establishing multimodal corpus. The multimodal corpus file is saved as "*.eaf" after annotation. Elan software has a powerful search function, and you can retrieve a single corpus or all of them by the "search" item. Elan is an extraordinary flexible software which can control the speed of playback, pausing, and revisiting autonomously, meanwhile it can point out the output of multiple discrete search results as well as shown in "Figure 3":

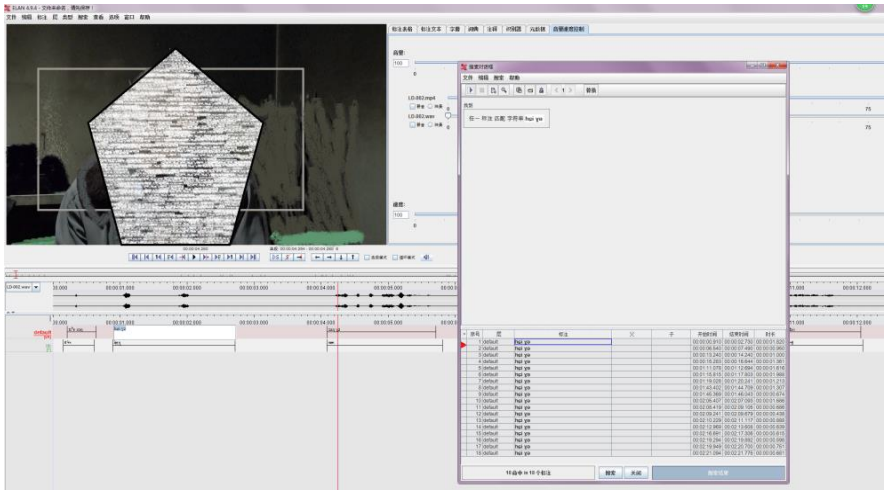

Figure 3 The text retrieval.

3. APPLICATION OF MULTIMODAL CORPUS IN TEACHING

Promoting students' knowledge of Tibetan language is the key to training native language learners. The fundamental approaches include the combination of traditional teaching methods and modern multi-modal teaching methods. Elan is used in this paper to build a small multi-modal Tibetan language teaching corpus, which can promote and inherit the excellent traditional culture of the Chinese nation and meet the current reform demand of Tibetan language teaching in Tibetan areas.

As an important teaching resource in classroom teaching, teachers and students can pick up the teaching contents from the self-built multi-modal corpus of Tibetan language teaching, which can greatly improve students' understanding of the classroom and get closer to the topic of classroom. Compared with traditional teaching resources, multi-modal corpus teaching resources have a great degree of openness and sharing. The teaching content and teaching dynamics can be added and improved at any time, and the videos that reflect the original appearance of the teaching activities in the multi-modal corpus can be incorporated into the teaching corpus, so as to effectively mobilize the learning enthusiasm and improve the self-learning ability of native language learners. Based on Elan's self-built multi-modal corpus, it can control the playback speed of the corpus itself, optimize classroom teaching and save learning time. It is helpful for practicing after class and previewing before class.

4. CONCLUSION

Multi-modal corpus is the most authentic corpus in the corpus family. The research on Tibetan teaching that based on multi-modal corpus contains rich methodology, which can effectively expand the research horizon of Tibetan language teaching, promote the rapid development of Tibetan teaching, and promote the reform and innovation of Tibetan teaching. On the basis of summarizing the basic framework of building multimodal corpus for Tibetan teaching, this paper not only discusses the use of Elan software to mark the corpus of multimodal corpus, but also discusses the application value of multimodal corpus in Tibetan teaching. In the author's opinion, the above research contents can be applied in various fields of Tibetan language teaching and provide valuable feedback information for compiling Tibetan language teaching materials and optimizing teaching process in Gannan Tibetan areas.

AUTHORS' CONTRIBUTIONS

Dawa Pengcuo wrote the manuscript, and Daojie Ben contributed to revising and editing.

REFERENCES

[1] Foster, M.E. & Oberlander, J. Corpus-based generation of head and eyebrow motion for an embodied conversational agent [J]. Language Resourcesand Evaluation, 2007 (3/4): .305.

[2] Bocklet, T.et al. Erlangen-CLP: A Large Annotated Corpus of Speech from Children with Cleft Lip and Palate [A]. In Kipp, M.et

al. (eds.). Proceedings of LREC [C]. Paris: ELRA, 2014:2671-2674.

[3] Zhu Jielin. GanNingqing history of Ethnic Education compendium [M]. Xining: Qinghai People's Publishing House, 1993.

[4] Boholm, M. & Allwood, J. Repeated head movements, their function and relation to speech [A]. In Nicoletta C.et al.(eds). The LREC 2010 Proceedings [M/CD]. Valletta: European Language Resources Association (ELRA), 2010.

[5] Tohyama, H. & Matsubara, S. The relationship between fillers and ease of listening in English Japanese simultaneous interpreting [J]. (in Japanese)Interpretation Studies, 2007(7).

[6] Zhu Kelin. Gannan Education History [A]. Gannan Cultural and Historical Materials (Vol. 8) [C].1991.

[7] Gannod GC, Bure JE, Helmick MT. Using the inverted classroom to teach software engineering [A]. Acmlieee International Conferenceon Software Engineeing [C]. IEEE, 2008.

[8] Qu er tang, Zangzu de yuyan he wenzi [The language and writing of the Tibetans]. Beijing: Chinese Tibetology publishing, 1996.

[9] Chen Qiao-hong, Wang Hao-yong. Research on ELAN based Multimodal Discourse — A case study of the first prize teacher discourse in the National College English Teaching Competition [J]. English abroad, 2018(1).