Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
May-June 2022 Page No. 1097-1104
Article History: Received: 24 February 2022, Revised: 05 April 2022, Accepted: 08 May 2022, Publication: 30 June 2022

A New Annotation Method for Phonetic-Grammatical Study of Text Resources in Chinese Minority Languages

Xiaomei Yang
Guangxi Vocational Normal University, Nanning, Guangxi, 530000, China

Abstract:

An in-depth study of minority languages in China is not only beneficial to reveal the developmental characteristics of languages at a deeper level, but also to the development of related disciplines. Text resource annotation is a basic work and a work with relatively high manual involvement. How to ensure the quality of processing? How to improve the efficiency of processing? A range of issues such as these are the main reasons for the slow development of text resources. In the face of the study of minority languages in China, the corpus is being paid more and more attention, and thus the text annotation work is extremely important. This paper designs a set of effective solutions and methods to improve text annotation, which provides linguists with new methods for efficient speech transcription annotation and grammatical analysis annotation in text resource research, and provides a series of solutions for building large corpus, multilingual speech dictionaries, "interlinear cross-referenced" processed corpus, dynamic word lists, etc., with the dynamic and interactive nature between speech, text, dictionaries and word lists. Based on the annotation methods in this paper, hundreds of texts in Tibetan, Yi, Yao and other minority languages have been annotated and processed by several researchers, obtaining relatively satisfactory processing quality and efficiency, achieving efficiency and practicality.

Keywords*: Chinese minority languages, Text, Annotation, Phonology, Grammar, Dictionary, Word list.*

I. INTRODUCTION

With the deepening of language research, especially in the face of minority and endangered languages, more and more linguists realize the importance of authentic textual corpus. Those traditional research methods based on word lists and example sentence surveys have certain limitations[1,2], and the more prominent problem is the lack of flexibility in language analysis and comprehensiveness in data mining.

The construction of language resources is a job with relatively high manual involvement, especially for text resources. Data collection, voice annotation, and grammatical analysis face a series of problems, and a large number of heavy manual operations make the research work slow.

At present, there are many annotation platforms or annotation methods both at home and abroad. Abroad, many linguists pay special attention to the study of corpus, and thus have also developed many texts processing platforms for language research in phonology, dictionary, and grammar in minority languages, and published very valuable corpus bibliographies. Some of the most used technology platforms are ELAN, a tool for speech transcription, ITE technology from France, and Toolbox for grammar annotation, among others. In China, there are mainly "Longmao Data" for ASR speech transcription, "iFlytek Development Platform" for speech recognition, "BasicFinder" for data collection and data annotation, and "crowdsourcing" approach adopted by Amazon Robotics, etc. According to the survey, Elan and Toolbox are the main tools used by linguists in China, such as the Chinese Academy of Social Sciences, Minzu University of China, Yunnan Minzu University, Guizhou Minzu University, Guangxi University for Nationalities, etc. However, more linguists point out that there are some critical problems to be solved in the annotation work.

The core function of Elan is voice annotation[3,4], but annotation with it encounters two very difficult problems, one is that the speech synchronization function does not substantially achieve accurate synchronization; the other is that grammatical annotation is very troublesome. The core function of Toolbox is grammatical annotation[5], and again, Toolbox has some problems, such as word separation techniques and annotation methods that need to be improved. And for voiced languages such as Chinese dialects, the method provided by Toolbox is not suitable, and the solution is given on the platform, but the processing is complicated and very difficult to operate for many researchers, in addition, there is also a lot of repetitive work, which makes the annotation work less efficient.

In the face of these problems, an annotation method that is suitable for the in-depth study of Chinese minority languages and aims to improve the efficiency of text-based corpus processing is extremely urgent and important for linguists.

II. DATA RESOURCES FOR TEXT ANNOTATION

Faced with the complex and rich minority languages, the corpus collection has integrity and authenticity, which involves pragmatics and context, which is of great significance for grammar research. The annotation method given in this paper will build a large number of linguistic resources, such as phonetic, grammatical, textual, dictionary and word list data resources. And how to ensure the quality and efficiency of the data resources is the key and purpose of this paper.

2.1. Voice Resources.

Language survey collects a large number of speech and video files to preserve the most realistic corpus speech resources. In voice annotation, taking into account various factors such as linguist research use and speech text synchronization, voice annotation will have the following three characteristics:

(1) Each language unit will have different phonetic annotations for multi-layer display, such as

international phonetic annotations, transcriptions, ethnic characters, and so on. And you can also add or subtract some specific phonetic annotations by yourself.

(2) Records of different languages (dialects) of the same text are arranged in a hierarchical manner, thus enabling a quick comparison of the similarities and differences of several dialect points.

(3) Each language unit of text has phonetic material corresponding to it and calling each other.

In the pre-annotation work of voice resources, in order to reduce the manual workload, the audio-assisted segmentation processing given in this paper, as well as the fast voice annotation method, all lay the foundation for linguists who want to implement a voice dictionary repository later, such a dictionary is very helpful for language research, multilingual teaching and so on.

2.2. Textual Resources.

That is, the corpus, a more realistic record of the full picture of the language. In this paper, textual resources are divided into two main parts, raw corpus and processed corpus.

(1) Raw corpus, i.e., text resources that have not been processed by grammatical annotation. This paper provides multiple ways to obtain the corpus, firstly, the fast voice annotation method given in this paper, which is 3 to 5 times more efficient than other annotation methods in the past, such as traditional manual entry and Elan transcription. Secondly, the data representation in this paper provides default data format and customization methods, which provide researchers of different languages with multiple ways to perform voice annotation. And it is fully compatible with the data format of text resources already processed by annotation tools such as Elan and Toolbox, providing researchers with an interface to import them. Finally, there are linguists' original saved corpus resources, such as file formats in Word, Excel, PDF, TXT, which can be easily imported.

(2) Processed corpus, i.e., interlinear cross-referenced annotated text, is also the language research results after grammatical annotation. This kind of annotated corpus is of great value for minority language research, and some results have been achieved in China, while foreign countries have studied this kind of interlinear cross-referenced language grammar annotation much earlier, and they attach great importance to in-depth research on minority languages. The completed processed corpus of the annotation is more meaningful in the grammatical description of the grammatical features, with several more annotated lines than the usual example descriptions, and these also provide help for non-native researchers, which can be shared across languages and resources. For example, the examples used by Jiang Di in the article "Identification of Irregular Verbs in Tibetan" use interlinear cross-referenced examples, so that even if one does not know Tibetan, one can clearly understand and appreciate the grammatical meaning of the statements and each word in the Tibetan corpus, as well as the phenomenon of morphological changes of Tibetan monophthong verbs as explained in the author's article.

2.3. Dictionary.

Dictionary is an important tool for language research, and it has a bridging role in this paper. The researcher is able to automate the work of text grammar annotation relatively quickly, where the grammatical information of the dictionary is derived from the dictionary. Each language entry contains Latin transcriptions, international phonetic symbols, grammatical attributes, etc., represented by interlinear cross-referencing. The researcher can also add translations for each dictionary entry in multiple languages, containing pragmatic features, etc., depending on the language study.

The dictionary construction method takes the interaction of words and texts, the text annotation work is automated, and the linguistic grammar research work is left to linguists. New dictionary entries are automatically recognized and identified in the text annotation, and the grammar researcher stores the new entries in the dictionary based on the judgment of the back-and-forth semantic analysis, so that new words can be obtained in a large number of texts. If an entry attribute is wrong or a new dictionary attribute or grammatical feature is found, error correction can also be done in the dictionary to achieve the purpose of automatic error correction in the whole text.

In addition to new entries added by researchers, dictionary resources can also be imported into the dictionary by simply marking the relationship between the dictionary attributes of the entries and their corresponding dictionary attributes, as well as dictionary data obtained through the "crowdsourcing technology" provided by the open port in this paper, and entry information completed by the TOOLBOX compatible tool and survey software such as "Feifeng".

2.4. Word List.

The word list belongs to the result resource, and the amount of information in the word list increases as the corpus increases. The word list records the words and morphemes appearing in the corpus, as well as their grammatical properties, and makes the word frequency statistics, the provenance of the words in the text. The word is is counted based on the grammatical properties of each linguistic unit in the text, in accordance with the dictionary's collected linguistic units.

III. METHODS OF PHONETIC ANNOTATION AND GRAMMATICAL ANALYSIS

3.1. Realization Idea of Annotation Method

This paper adopts a new text annotation method for corpus processing that respects the objectivity of real texts and aims to facilitate the use of researchers, providing a solution with flexibility and practicality. The general framework of the new annotation method for phonetic-grammatical study of Chinese minority language text resources is shown in Figure 1 below.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
May-June 2022 Page No. 1097-1104
Article History: Received: 24 February 2022, Revised: 05 April 2022, Accepted: 08 May 2022, Publication: 30 June 2022

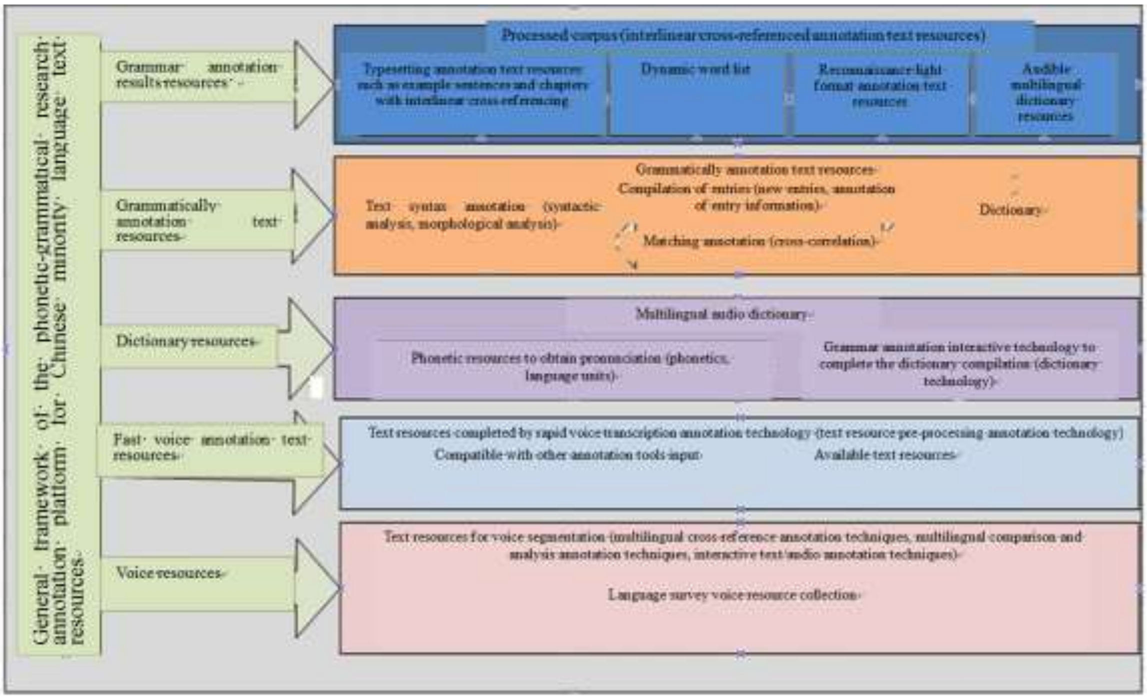
General framework of the phonetic-grammatical research annotation platform for Chinese minority language text resources
Grammar annotation results resources
Grammatically annotation text resources
Dictionary resources
Fast voice annotation text resources
Voice resources
Processed corpus (interlinear cross-referenced annotation text resources)
Typesetting annotation text resources such as example sentences and chapters with interlinear cross-referencing
Dynamic word list
Reconnaissance-light format annotation text resources
Audible multilingual dictionary resources
Grammatically annotation text resources
Text syntax annotation (syntactic analysis, morphological analysis)
Compilation of entries (new entries, annotation of entry information)
Matching annotation (cross-correlation)
Dictionary
Multilingual audio dictionary
Phonetic resources to obtain pronunciation (phonetics, language units)
Grammar annotation interactive technology to complete the dictionary compilation (dictionary technology)
Text resources completed by rapid voice transcription annotation technology (text resource pre-processing annotation technology)
Compatible with other annotation tools input
Available text resources
Text resources for voice segmentation (multilingual cross-reference annotation techniques, multilingual comparison and analysis annotation techniques, interactive text/audio annotation techniques)
Language survey voice resource collection

Figure 1 General framework of the new annotation method for phonetic-grammatical study of text resources in Chinese minority languages

3.2 Phonetic Transcription and Grammatical Analysis Annotation Methods for Chinese Minority Language Text Resources

3.2.1. Voice annotation.

In this paper, the voice-to-text annotation is based on a dictionary, and the data sources are obtained through "crowdsourcing" with the help of the researchers' accumulated data word lists, dynamic output word lists, and open ports for users. Then, using the adjacency matrix method, we load the data into memory and construct a graph to build a chain of data sources with increasing information content to assist the annotator in transcribing the voice, thus reducing the workload of voice-to-text annotation. The voice annotation designed in this paper cannot achieve the same effect as the ASR voice transcription, because for language research, the annotation is mainly done hierarchically around native language, international phonetic symbols, Chinese, English and several others, so that the completed language resource database is more meaningful and can be studied by non-native speakers.

We have done the voice annotation work for Tibetan, Yi, Zhuang and other minority languages with the text resources completed by rapid voice annotation, which accelerates the recording efficiency and saves the recording time than Elan voice annotation tool, and the experimental results respond well, but further improvement is needed to improve the annotation efficiency. The long-form Yi language corpus, *Elder Brother and Younger Brother*, was transcribed by native researchers through this platform to quickly transcribe the completed annotated text, and the efficiency was improved by 3 times.

3.2.2. Grammatical annotation.

Text processing is at the heart of grammar annotation, and this section provides grammar researchers with annotation methods that dynamically assemble the grammatical rules and dictionary semantics of a language from a real corpus using an interlinear cross-referenced format. In the grammatical annotation, the construction of a dictionary, text division, syntactic analysis and morphological analysis are mainly completed. These grammatical annotation methods reduce a lot of repetitive and tedious workload for researchers and improve the annotation processing of the whole corpus for grammatical research. In the syntactic analysis, we have improved the word division method, and in the morphological analysis, we have given a new text annotation method mainly for the characteristics of Chinese language.

The text and the dictionary are interactive, and in the text annotation processing, the unrecognizable words are automatically marked out for the researcher to quickly view and add new words to the dictionary, as follows: the text and the entries are interactively marked to identify new entries, and the interlinear cross-referenced format, where the place marked with * is the new entries, "Those who laugh at others will eventually be laughed at by others" (in Uyghur).

The key technology of text processing is text segmentation, and the segmentation method and the segmentation process of word recognition are very important [4]. The dictionary data in this paper is an accumulation process, and if the segmentation method depends entirely on the dictionary is impossible, especially in the face of minority languages, the annotation efficiency of natural language processing will be very low if it is based on dictionary matching only. In this paper, an improved segmentation method is given to provide different annotation methods for syntactic analysis and morphological analysis.

Text segmentation based on the requirements of linguists, who believe that the larger the granularity of the segmentation result, the better. In this paper, a dictionary strategy is used, so the size of the segmentation unit is determined by the granularity of the dictionary. In the syntactic analysis of the word separation algorithm combined with the improved KMP pattern matching algorithm and dynamic matching rule algorithm, etc. to complete the text segmentation.

Morphological analysis, this paper uses some special ways to develop rules and add special symbols to mark certain specific, common and repetitive grammatical phenomena, mainly to facilitate the labeling of certain specific, common and repetitive grammatical phenomena, and these rules are the ideas provided by linguists who have been engaged in grammar research for many years.

For the characteristics of Chinese minority languages, linguists suggest that grammatical annotation can be done through morphological structures. This paper mainly focuses on flexion, overlap and tonal change grammatical phenomena to solve part of the subordination and annotation problems, to assist the researcher to complete the annotation automatically and thus improve the annotation efficiency, and further improvement and enhancement of the annotation methods are needed later.

The flexion phenomenon, in this paper, in the dictionary construction, the dictionary entries map the surface and deep correspondence, which can be represented separately in separate lines or spaced apart by ";". In the following example, the Tibetan corpus, "Stingy Master", where the original words gyis, smras, and khyer are marked with a change in the participle line, and the dictionary attribute line is marked with the grammatical meaning. This is the surface and deep correspondence processing method to solve the annotation problem of the flexural phenomenon. If new deep semantics are found, then added to the dictionary entries, the grammatical information of the dictionary entries of the investigated language becomes more and more perfect. This training mode, automatic annotation will be more and more efficient.

The phenomenon of overlap is a grammatical phenomenon with strong regularity in languages, and minority languages are more complex and somewhat different, with a great variety of overlapping forms, and this paper only addresses simple overlapping forms, such as AA, ABB, AABB. The following example sentence from the text of the corpus shows the Yi language "Have you eaten yet?” dzɯ^{33} dzɯ^{33} and dzɯ^{33} da^{21}, two representations of predicate verb roots, are automatically recognized when annotated by interlinear cross-referencing.

The phenomenon of tonal change, first ignore the tone, direct match dictionary entries, and then this mandatory automatic labeling with a special symbol to identify, this paper uses font color as a distinction, assuming that the original tone of existing dictionary entries is ŋo^{21} ɡɯ^{55}, then ŋo^{22} ɡɯ^{55}, ŋo^{21} ɡɯ^{33} all directly match the annotated proto-tone entries. In the experiments, it is found that this forced method is very efficient than the normal separate annotation, and also more convenient and efficient than the processing method provided in Toolbox for adding dictionary entries separately.

In this paper, grammatical annotation is carried out on more than one hundred corpuses, which is mainly tried by teachers of Chinese Academy of Social Sciences and students of Guangxi University and Guangxi University for Nationalities, and the results respond to be more practical, simple to operate, and solve some problems in segmentation processing.

3.2.3. Results resources.

The system provides a variety of selectable and layout able outputs, which are accomplished through search filtering techniques and line-break processing, mainly to facilitate the publication of researchers' results and resource sharing, and is capable of exporting whole or selected passages of interlinear cross-referenced processed corpus, as well as grammatical screening examples, dictionaries and word lists.

IV. CONCLUSION

The acquisition of high-quality authentic corpus resources is the basis for conducting language research. This paper provides linguists with annotation methods for voice transcription and grammatical analysis

Article History: Received: 24 February 2022, Revised: 05 April 2022, Accepted: 08 May 2022, Publication: 30 June 2022

through a series of information processing techniques, which can automatically complete the annotation work in batches, improving the efficiency of annotation and alleviating the repetition and large amount of annotation work. Meanwhile, the Chinese Language Resources Conservation Project being implemented by the Ministry of Education of the People's Republic of China will provide a large amount of voice and real text corpus for the linguistic community, and this paper also provides an effective annotation method for this project. All these works will be beneficial to promote the in-depth development of Chinese linguistics and the in-depth study of textual resources.

REFERENCES

[1] Sun Hongkai. Chief editor. Language of China. Commercial Press 2007
[2] http://www-01.sil.org/computing/toolbox
[3] Elan manual. http://www.mpi.nl/corpus/manu-als/manual-elan.pdf. 2012
[4] Jiang Di. Construction Methods of Tibetan Grammar Dictionary: Interaction between Entry and Text. Advance in Chinese Information Processing. 2006
[5] Li Bin. Chinese Dialect Multimedia Corpus Built by ELAN and Its Application. Hunan Normal University. 2013