A Tentative Idea of Multi-modal Corpus Applied to National Minority Chinese Proficiency Test in China

Zhou Xuan, Chang Yinghao, Cao Deqing
Beijing Language and Culture University
No. 15, Xueyuan Road, Haidian District, Beijing
luoxiting2012@163.com

Abstract

To the masses of Minorities in the frontier areas of China, proficiency in the Chinese language tool is not only conducive to the common development and prosperity of all nationalities in our country, but also an important cornerstone for ensuring the smooth running of the "Belt and Road" strategy. However, for a long time, the primary means of monitoring Chinese learning in frontier minority areas is the Chinese Minority Chinese Proficiency Test (MHK). The disadvantage of this method lies in the fact that the test scores of written and oral exams are mostly rough and monotonous. They fail to fully assess the mastery of Chinese in minority areas and accurately reflect the integration of minority languages and Chinese. Therefore, this paper argues that there is an urgent need to establish a dynamic text corpus of Chinese Minorities in frontier areas, which can not only be used to monitor the use of Chinese and bilingual integration in local areas, but also be used in test selection, difficulty control, vocabulary development, and provide reference for the Minority Chinese test, provide more targeted suggestions for Chinese teaching in the "Belt and Road" region, and provide data support for the security of national language and character at the same time.

Keywords: Multi-modal Corpus , Education Measurement, Language Testing

China is a multi-ethnic country which has many ethnic minorities. The Chinese learning situation of ethnic minorities in the border region matters whether the policy of national integration is successful. In recent years, with the promotion of China's international status, there are more and more researches on the popularization of Chinese. But most studies have focused on the spread of Chinese in foreign countries, rather than systematic monitoring of the use of Chinese in border areas. The promotion of Chinese is an important indicator of China's improvement in soft power and the mastery of Chinese of the border ethnic minority is a necessary condition for the internal political stability of the country. With the implementation of One Belt And One Road strategy, the languages of ethnic minorities near the border are more important since Chinese is the cornerstone of One Belt And One Road's economic and trade communication. In addition, once the One Belt And One Road strategy was carried out, the language security needs to be real-time monitored since opportunities of external exchanges for ethnic minorities will increase.

At present, the main means of measuring the mastering situation of Chinese for minority is a series of tests represented by National Minority Chinese Proficiency Test(MHK). MHK is a national standardized test built under the guidance of second language teaching theory and combined with the characteristics of ethnic minorities of Chinese learners in China. It mainly examines a test taker's ability of using Chinese in communication, especially the ability of study, work and social communications. The test items of MHK include listening comprehension, reading comprehension, written expression and oral expression. By testing different language skills, MHK can comprehensively test the ability of candidates to communicate in Chinese. Language testing is an important means to evaluate language ability, but a single report score does not specify the problem. The shortcoming of this approach is that the test scores can neither comprehensive evaluation of ethnic minority areas to master Chinese, also can't accurately reflect the fusion of the ethnic languages and Chinese. Therefore, this paper argues that minority nationalities Chinese corpus needs to be established, which can be used to monitor the local Chinese and bilingual fusion status and can also be used to help test material selection, difficulty control, glossary revision and oral evaluation standard development for the National Minority Chinese Proficiency Test.

1. Corpus Linguistics Bibliometrics Analysis in Recent Ten Years

Bibliometrics is a quantitative method which studied the external characteristics of the literature and the output must be quantized information content. As a branch of library information science, mathematical and statistical methods are used by bibliometrics to describe, evaluate and predict the current situation and development trend of a subject. (Qiu Junping, Wang Yue Fen, 2008: 1). In the past, bibliometric methods were mostly used in the field of natural science, and then gradually radiate to the humanities and social sciences. From the perspective of literature metrology, we take a quantitative research approach to the research of the subject area. From the metrology point of view, we can calculate the theoretical indicators of each subject in the field of literature and describe the development on a certain dimension of one subject.

1.1 The Object of Bibliometric Analysis and Highly Cited Papers

We can get hundreds of search results by selecting the "advanced search method" on online literature retrieval platform, taking "Philosophy and Humanities", "Social Science", "Information Technology" as the subject area, selecting CSSCI as a journal source, choosing the time span from 1998 to 2017 and taking "corpus" as the key words to carry on the accurate retrieval. We finally retrieve 270 articles through the artificial screening, removing the non-disciplinary research articles such as "conference essay, meeting review, book review, notice and notice" in the result. With the help of bibliometrics, this paper analyzes the annual changes of the published papers and the data of the research topics, and calculates the frequency and proportion of different dimensions. In addition, according to the quoted frequency sort, we

select the most cited citations from 1998 to 2017, and make a review of representative papers related on language test, thus outline the hot spots and trends of the research in the field of Chinese corpus in the past 20 years.

1.2 Changes of Yearly Quantity Published Articles

Quantity of published articles is the number of essays published under the specific conditions. Taking a year as a node of time and counting the number of published papers in each year in 10 years, we can see the intensity of the discipline research in the field of domestic Chinese corpora in recent years. Changes of yearly quantity published articles between 1999 and 2018 are as follows.

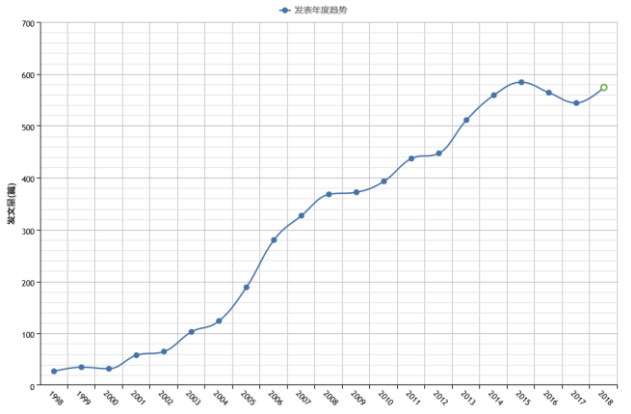

Figure 1: Changes of yearly quantity published articles between 1999 and 2018

From Figure 1, we can see that, the volume of articles involved in corpus research is on the rise during the 20 years from 1998 to 2017 on the whole. Among them, there was a decrease from 2015 to 2017, an upward trend from 1998 to 2015 and a downward trend from 2015 onwards, indicating that 2009 is a watershed in the field of corpus linguistics. From the correlation between the number of published papers and the development of disciplines, we can roughly infer that the corpus linguistics volume increased steeply from 2015 to 2015, indicating that the discipline flourished during this period and was constantly making new breakthroughs period. After 2015, the number of published papers has picked up. It doesn't mean that corpus linguistics is no longer important. Actually, it is due to corpus linguistics has made breakthroughs in recent years and the fundamental difficulties have been overcome. As a result, the number of published papers has slowed down.

Among them, it is noteworthy that quantity published journal articles in 1998 was 4 and the number in 2017 and 2014 was about 540. The number of papers published during the six years from 2004 to 2009 increased by leaps and bounds every year, with an average increase of nearly 100 per year. Finally, the first small peak was ushered in in 2009, reflecting that these years are the spring of corpus research. However, it started to grow again for the second time in 2013. In 2015, the number of published papers on linguistic test reached its peak, which shows that the study of corpus linguistics still shows great vitality after 20 years development.

1.3 The Change of Hot Spots and Trends in Corpus and Language Testing Field

The distributive situation of the research topics can reflect the concentrated hot spots and development trends of a subject. By focusing on the concentrated distribution of the research topics over the past 20 years, we can find the research hot spots in the field of the corpus in recent years. From the vertical perspective, we can outline the development trend of corpus linguistics.

We used literature visual analysis tools to analyze 5950 retrieved essays which were used as samples. K-means clustering analysis was used to analyze and the threshold value was set as 50, and "key words" were analyzed for papers from 1999 to 2018. Finally, we get clustering results of class topics. By selecting the high volume of text from the amount of distribution of keywords and time extension analysis, we hope to outline the research hot topics and vein in the field of corpus focus.

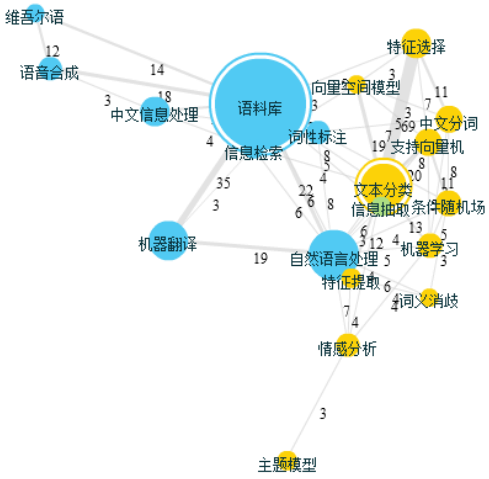

Figure 2: Clustering results of class topics of corpus

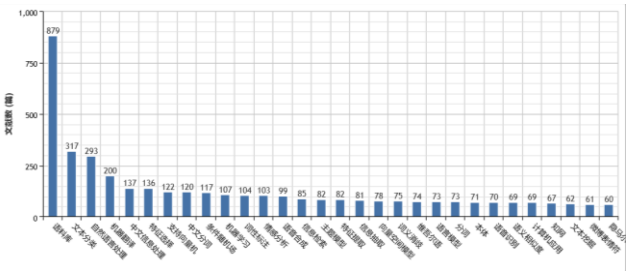

Figure 3: Research hot spots on corpus research

Since corpus linguistics belongs to a branch of natural language processing and computational linguistics, topics of natural language processing and Chinese information processing are removed from retrieved essays. From figure 2, it can be found that the research focus of the corpus research mainly involves text classification, machine translation (more than 200 papers) in the past 20 years, followed by machine translation, machine learning, emotional analysis, speech recognition and other fields. This phenomenon shows that corpus linguistics not only focus on the text corpus, but also focus on voice and other non-text corpus. Technically speaking, with the development of machine learning, more and more machine learning is used to deal with the urgent problems in corpus. Among them, there are 74 articles in Uyghur language, which indicate that the corpus has been explored in the minority languages, and Uyghur language

is a very important part.

From the corpus bibliometrics analysis, we can see that the corpus research has made great progress in recent years, but the main concern is language ontology and corpus construction technology application. Scholars pay more attention to how language resources are collected, the way to collect, how to set up the corpus, the development of computer application technology. The research of application of corpus on teaching and language testing is relatively rare. Due to the particularity of the subjective test of language testing, the application of corpus technology is needed.

2. Application of Corpus in Language Testing

2.1 Corpus Auxiliary Language Test Proposition

As the first process of test formation, proposition often requires a lot of time and labor, is a very important but difficult part in language testing. Language test includes many sections, reading section is the longest and long-lasting part. In the proposition process of MHK, the collection of reading texts and the proposition of the item has become an indispensable part. However, in the long term item-constructed process we find that most of the corpus material used by the proponents is collected from the Internet. Although the network is more convenient and faster than the paper media collection, the quality of the network material is uneven. Finding the high quality corpus that meets the requirements in a short period of time is not an easy task. The propositional personnel often need to spend a lot of time to filter and modify the corpus. Compared with the proposition, the collection of corpus will take more time for the proposition, which makes the proposition inefficient and low efficiency.

In addition, compared to the collection of reading text in the process of traditional language test proposition, the collection of the listening text is much more difficult, which makes the following questions are very common. Firstly, how to produce a text that meets the needs of everyday situations and the output of natural language flow, this would make the proposition staff often feel a headache. Secondly, if the text of the listening all rely on proposition personnel's original creation, it would lead to dialogue's mismatch with the logic of daily conversation and the output of natural language flow due to the lack of mental capacity. Because of the shortage of high-quality audiometry texts and titles, the high repetition rate of conversation and dialogue content is a common problem in listening test texts, which may increase the exposure of topics in an untruthful manner, which is not conducive to the long-term and effective implementation of large-scale examinations. Third, it increases the workload of initial examination and the difficulty of work, resulting in the low utilization rate of language test talents.

In view of this, if multi-modal corpus for daily conversation of multiple scenes in accordance with linguistic norms can be established, annotated and analyzed in many aspects of sound and body posture, stored in different categories for retrieval and screening, it will greatly reduce the burden on the proposition stuff. At the same time, effectively improve the quality and speed of the proposition, reduce the basic trial links, and improve the utilization rate of language test talents.

2.2 Language Test Validity Argument and the Difficulty of Expansion

Validity is the most important indicator to ensure the quality of the test questions and test the validity of the test. For years, the validity of the language test is validated using the theoretical framework of validity test. For example, the validity of the occupational proficiency test demonstrates the validity of the test question by using the correlation between the test scores and academic equivalence criteria. Regardless of the validity model, the validity of the test questions is validated from the "post-test correlation" between the post-test score and a conventionally established standard. If the validity is not high, there is no way to retrieve. If we can control the validity of a test in a certain way before the question test, this is not only the expansion of the validity argument research in language testing, but also is a great help to grasp the validity of the test questions in advance, especially beneficial for the ethnic Chinese proficiency test.

Test difficulty is to ensure that the parallel test fairness of an important regulatory indicators. In the current MHK test, essay, oral and other subjective questions is to take a parallel test paper to examine the candidates. Candidates who take the test may do essay or oral exams on different topics, but it is difficult to assess the consistency of test questions. At present, the control of the difficulty of exam questions can only be controlled by experts and experienced proposition stuff based on experience which cannot be quantified. If we build a corpus based on the MHK language test, we can make statistics on the ability distribution of participants who participated in the MHK language test, observe the candidates' knowledge of knowledge points in different languages, and have an intuitive data support for the difficulty of the difficulty of the test questions Subjective questions difficult to predict and test the equivalent of the problem.

2.3 Promote the Basic Unity of Language Proficiency Evaluation Criteria

The study of language ability in the field of language testing has undergone three different stages: the stage of skill/component speaking, the stage of speaking of overall ability and the stage of establishing communicative competence model. However, the academic community has not formed a unified understanding and evaluation criteria so far. As an unavoidable problem in the field of language testing, we all see each other's own ways to make language testing has long been unable to obtain a more authoritative test system of global recognition, but also to test the candidates caused great distress.

It seems that our discussion does not seem to make much sense if we just stay on verbal extermination. However, different scholars' ideas, disagreements and debates may be able to obtain fair judgment of third-party fair referees in this era. That is, using the platform of multi-modal corpus and utilizing the application of high penetration electronic terminals to collect and analyze high and low level language proficiency people's written information and video information, through the actual observation of the differences in performance, to speculate on the essence of language ability and the appropriate evaluation criteria, the resulting language proficiency evaluation criteria will be more scientific. Moreover, the research

results supported by the actual data will be approved by more experts and scholars and the general public, and promote the basic unification of the language proficiency evaluation standards.

The basic unification of language proficiency evaluation criteria and the requirement of quantitative assessment meeting the level of adaptive test capability will also provide theoretical basis and data support for the adaptability of language test.

2.4 Integration of Measurement, Study and Research

For many years, there has been a phenomenon of the test disconnect between language testing and language teaching. After a long period of learning and preparation, the improvement of language proficiency of many candidates cannot be tested by the test. The test results may reflect the candidate's language ability in some aspects, may also ignore some aspects of the language skills of candidates, the test is not targeted.

This long-term state separation of study and measurement greatly reduces the effectiveness of teaching and testing, and the establishment of a contiguous and consistent multi-modal corpus will help to promote the consistency of learning tests, improve the effectiveness of language tests and enhance learning Testability of the study can be landed. We can establish a multi-modal corpus including image, sound and other forms of multi-modal corpus through teaching to increase students 'interest in learning, discover candidates' problems through testing, and promote teaching pertinence through questions. We can also collect information from multi-modal corpus about the correct and wrong data of students. The consistency of test and teaching can help the test effectively distinguish between candidates with different levels of competence, reflect the improvement and decline of students. In this way, we can form a highly integrated and efficient test mode.

3. Suggestions for establishing corpus based on MHK

3.1 Collection of Language Resources Should Be True and Wide

The collection of the real corpus resources can not only come from the real corpus of the students in the test, but also from the first-line students' homework during the teaching. The resources should be extensive in order to ensure the effective application of the later corpus. In addition, it should be based on the different levels of difficulty and the order of collection should be successively.

3.2 Corpus Construction Must Be Targeted

Today, the formed corpus has been built, but the corpus that can be applied directly to the Chinese language test, especially the minority Chinese language test corpus, does not exist. The corpus construction must be well-targeted and able to solve the practical problems faced by ethnic minority Chinese proficiency testing. Otherwise, there is no need for construction.

3.3 Constructed Corpus Should Be Open

The use of many corpora is not open, which is not conducive to the development and updating of the corpus. Therefore, we suggest that the nature of the corpus is open and shared, and all the corpora need to be used by corpora in all fields. At the same time, the corpus should be dynamic to facilitate the timely updating of the corpus and the timely updating of the corpus can be realized automatically by the corpus individual users, so you can save the cost of lasting maintenance. Professional corpus maintenance staffs only need to be responsible for compliance with norms of audit and technical maintenance.

4. Conclusion

The application of corpus in the field of language testing is a multidisciplinary research subject. It requires experts in many fields such as linguistics, computer science and language testing to discuss together in order to clarify the needs and construction direction of the corpus. Researchers in the field of language testing provide concepts and requirements for corpus construction, computational linguistics researchers provide technical and ontological research support in order to ensure that the corpus is scientific and advanced.

At the same time, the corpus construction is not an overnight thing and requires scientific planning and effective implementation. Many problems such as how to obtain the operating cost of open corpus and how to coordinate the corpus development patent are all discussed. This article hopes to play a valuable role for later researches.

References

Lijun Chen, Fanzhu Hu (2010). Language Resource: A Tourism Resource urging Development. 24(6), pp. 22-27.

Daming Xu (2008), Language Materials Management And Planning For Language Material Discussion. Journal of Zhengzhou University, pp. 12-15.

Dong Yongyi (2016). Study On The Relationship With Corpus And Language Testing. Language Construction, pp 87-88.

YunDuan Nong (2008). How Corpus Be Applicated in Modern Language Testing. Examination Weekly, pp 218.

Liang Bo (2013). The Application of Mini-text in Compiling Of English Language Testing Materials. Journal of Wuhan Institute of Shipbuilding Technology, pp 86-89.

Lin Lin (2016). Study on the Application Of Corpus Linguistic and Corpus—Comment on the series of teaching practice series and corpus of foreign language teachers in national colleges and universities. News and writing, pp. 117.

Qin Peng (2007). Development and Application of Software Tools for Monitoring Language Resource of Print Media, Beijing Language and Culture University.

Wang Hui, Wang Yalan. 2016. Language Situation of "the Belt and Road" Countries. Language Strategy Research(2), pp. 13-19.