Universe
Scientific Publishing

10.18686/ahe.v7i4.7269

The Methods of Data Collection and Data Analysis in Japanese Linguistics

Yuanyuan Bai

School of Japanese Studies, Dalian University of Foreign Languages, Dalian 116044

Abstract: In general, data collection, analysis, and processing are indispensable parts of research in Japanese linguistics. The researcher has to adopt different research methods depending on the object and the content of the study. There are studies that use only one method of collection and analysis, and studies that combine different methods depending on the situation. And when using various collection methods and analysis methods, there are different requirements to be met to ensure the rationality and accuracy of the study. Therefore, this paper introduces the data collection methods and analysis methods for spoken and written Japanese, and discusses the characteristics of different methods with specific research examples.

Keywords: Data collection; Data analysis; Spoken Japanese; Written Japanese

1. Methods of data collection

Linguistic data can be obtained in a variety of ways, such as corpus, surveys (interviews, questionnaires, etc.). Therefore, it is necessary to choose the most appropriate method from them according to the object and the purpose of the study.

1.1 Corpus

A corpus is an important way to collect linguistic data. The advantage of a corpus is that it contains a large amount of rich examples, from which it is possible to observe and analyze these data to get a clear picture of the features and tendencies of language use. It is necessary to collect as much data as possible through the corpus, since it is incomplete to understand the language through the introspection of native speakers and analysis of a language cannot be based on intuition alone.

The BCCWJ is a corpus of written Japanese that contains a total of 134. 3 million words and includes data from books, magazines, newspapers, textbooks, blogs, and other sources. In the past, most of the traditional studies of Japanese linguistics were based on linguists' introspection, but nowadays many studies use corpus as one of the ways of data collection. A search of Japanese linguistics studies conducted with corpus on CNKI shows that there are more investigate on word collocation and word meaning, in addition to studies on grammar. For example, Jiang Hong (2012) uses a corpus to analyze the semantics of Japanese adjectives by examining their collocations, using not only the BCCWJ corpus but also the Asahi Shimbun Cross-Search to compile a rich set of examples, and on this basis, Jiang Hong explores the expansion process of word meaning from a cognitive linguistic perspective.

There are also corpora for spoken Japanese, such as the CSJ, the CEJC, the NUCC, and the BTSJ. The BTSJ, developed by Mayumi Usami in 2007, is currently the largest natural conversation corpus in Japan, and the latest version in 2022 contains 474 conversations with a total length of 118. 5 hours. The corpus includes the speaker's social attributes such as gender, age, subordination and affinity, as well as the content of the conversations in different scenarios. It is also refined to mark phonological sounds, such as intonation, laughter, silence and so on. The construction of the BTSJ provides a great value of innovation and convenience for Japanese linguistics related pragmatics research through the corpus.

As mentioned above, the corpus can be used to study not only the native Japanese language, but also Japanese learners and a contrastive study between Chinsese and Japaneses, and there are various corpora to choose from in these areas. In terms of study on Japanese language learners, the CLJC, the I-JAS, the C-JAS, etc. are frequently used. According to the search results on the CNKI, the CLJC is more widely used. The corpus is mainly derived from the compositions of the CJT-4 and CJT-8 and the Chinese to Japanese test papers, which are representative and scientific. Mao Wenwei (2011) points out that in the study of second language acquisition, in order to accurately and comprehensively grasp the development of learners' language ability, it is necessary to analyze not only through the learner corpus, but also to compare it with the native language corpus. Because the learner corpus only reflects the acquisition of the learners, only by using the native language corpus to understand the specific use patterns of the target language and then comparing the results with the learner corpus can we grasp the characteristics of the learners' acquisition and the problems that exist, and then propose targeted solutions.

1.2 Questionnaire Survey

Questionnaires are mainly used in studies related to Japanese education. Questionnaires are administered not only by having the subjects fill out paper questionnaires, but also by interview and, with the development of the Internet, by distributing questionnaires directly online. The questionnaires can be divided into self-administered and interviewed according to the way of answering. Self-administered questionnaires are distributed to the subjects and then read and answered by the researcher before being returned. Interview Questionnaire is that the interviewer interviewed the subjects about the questions on the questionnaire and recorded their responses. In addition, questionnaires can be divided into structured and unstructured questionnaires according to the level of control of the researcher. Structured questionnaires have a rigorous set-up, with uniform rules for question wording and question order, and the answers to the questions are also restricted in advance, making the results easy to count and quantify, and the researcher has some control over the reliability of the results. Unstructured questionnaires, on the other hand, are more open-ended and only need to be broadly defined according to the research topic, with no strict restrictions on question wording and question order, and the response form is freely stated by the subjects.

Compared to a corpus, a questionnaire has the advantage of providing a more detailed picture of Japanese learners' performance and effectiveness in language learning. In order to ensure that the results of the survey are relatively objective, there are some points to note when creating the questionnaire. First of all, the researcher should try to avoid the use of eliciting expressions when describing the questions, and the standard of words such as "many" and "often" will be different for different people, so the researcher can set the number to determine the specific standard. In addition, the design of the questionnaire also needs to follow the principle of necessity, and must be closely focused on the research topic to set the exact questions, if the questions are too complex, it will reduce the recovery rate of the questionnaire, and vice versa, if too simple, the results collected lack of depth, the research problem can not be fully expressed from it.

Studies in Japan mainly focus on the motivation, learning awareness, or vocabulary and grammar mastery of foreign students, and also comparative analysis of their learning situation with native Japanese speakers. In addition to Japanese language education, questionnaires can also be used in studies related to Japanese pragmatics.

1.3 ATR and VTR

ATR (audiotape recorder) and VTR (videotape recorder) are methods of collecting spoken discourse using a tape recorder or video recorder.

ATR was mainly used to record internal conversations among Japanese people, and later it was also widely used in research in the field of Japanese education. It can be used in the research of Japanese education to analyze the communication strategies of target language speakers and to study the speech interaction behaviors of foreigners and Japanese learners. However, there are certain problems if only ATR is used. For example, one of the problems is the processing of written language and audio information beyond display. Therefore, in addition to using ATR, it is necessary to supplement and improve the recording of comprehensive data by combining participant observation and other methods.

In contrast to ATR, which is used for data recording methods in verbal conversation studies, VTR has a very important role for research in sociolinguistic fields such as nonverbal behavior. For example, Koichi Shimahara (2014) analyzed the content of miscellaneous conversations in contact scenarios between Japanese learners and native Japanese speakers at their first meeting in order to learn the acquisition of topic switching by advanced learners of Japanese. The process of collecting the corpus is based on audio and video recording of conversations between native Japanese speakers and Japanese learners during their first meeting. and to form textual data based on the sound data. In addition, the learners were interviewed after the conversation and asked whether they had received education on topic change methods.

2. Methods of data analysis

Data analysis is the process of organizing data after collection and analyzing what phenomena or problems are reflected in the data. It is a process of grasping the essence of the results presented by the data. Methods of data analysis are generally divided into qualitative analysis and quantitative analysis.

In terms of the development of research methods, there are first qualitative and then quantitative methods of analysis. In general, the method of qualitative analysis, which interprets data in an abstract and intuitive way, is often used in traditional linguistics. Unlike quantitative analysis, qualitative analysis can begin even before the end of data collection. For example, in unstructured interviews, because one has to talk to the respondents while considering how the conversation will proceed, the content of the conversation needs to be analyzed in advance. Therefore, studies like social surveys in pragmatics, discourse analysis, and second language acquisition cannot be conducted without the qualitative analysis method.

In contrast to qualitative analysis, quantitative analysis is a quantitative, concrete and objective way of data processing. For example, after the corpus is collected, quantitative analysis can be conducted to grasp the whole picture of a language phenomenon. At the same time, even for native speakers, the judgment of a certain language form and syntactic structure may differ from one user to another, and the results obtained by qualitative analysis alone may lack accuracy or persuasiveness at this time, therefore, quantitative analysis is also an indispensable part in linguistic research nowadays.

From the current linguistic research, these two approaches are often combined to play a complementary role to each other. It can be said that numbers determine meaning, while meaning also determines numbers, and qualitative and quantitative are in a dynamic balance.

3. Conclusion

This paper has focused on methods of collecting and analyzing data about spoken and written language in Japanese linguistics research. The corpus can be used to collect both written and spoken language, and has the advantage of being rich in types and materials. Questionnaires are often used in studies of second language acquisition and pragmatics. In addition, audio or video recording can be used in the collection of spoken discourse. The data analysis methods are mainly divided into qualitative analysis, which is relatively abstract and intuitive, and quantitative analysis, which is more concrete and objective. In addition to the methods introduced in this paper, there are other research methods, and it is important to choose the appropriate method according to the different research. These methods are not isolated from each other, and the combination of multiple methods is often needed in specific research.

References

[1] Gotoo Hitosi. Corpus linguistics and Japanese language studies[J]. Japanese Linguistics, 2007, (22): 47-58.

[2] Itsuko Fujimura, Naohiro Takizawa. Techniques of linguistic research: Data collection and analysis[M]. Hituzi Syobo, 2011.