WILEY | Hindawi

Retraction

Retracted: Development and Teaching Application of Japanese Corpus Based on Big Data

Wireless Communications and Mobile Computing

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Liu and H. Zhang, "Development and Teaching Application of Japanese Corpus Based on Big Data," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 5965123, 9 pages, 2022.

WILEY | Hindawi

Research Article

Development and Teaching Application of Japanese Corpus Based on Big Data

Xuelin Liu and **Hua Zhang**

Foreign Language School, Weifang University, Weifang, Shandong 261061, China

Correspondence should be addressed to Hua Zhang; zhanghua198316@126.com

The establishment of English and French corpora has provided great convenience for language application and research. However, in the field of Japanese language research, the construction of Japanese corpus has been slow due to the limitation of Japanese markup format. This paper discusses the scientificity, rationality, feasibility, and construction of "microcorpus" and "specialized corpus" for the application of Japanese corpus in Japanese language teaching. Textual corpus is a corpus that contains all text but no media such as audio and pictures, and it is mostly used in teaching. The Japanese language corpus is a corpus created by analyzing the existing Japanese classroom teaching situation in response to the current reform of Japanese language teaching mode. It is important for teaching Japanese in colleges and universities because it can help students to learn independently and improve their learning initiative and motivation. The teaching-based text corpus is a small corpus for Japanese language teaching and learning, which is derived from the teaching process. The corpus can be created independently on an ordinary PC machine through software, and the collected corpus can be imported into the corpus and used, which provides another way to improve students' interest and learning ability on their own. Based on the search software, this paper selects representative novels from the Aozora Library and creates a simple corpus to serve the teaching of Japanese interpretation, which improves the teaching effect. At the same time, this paper discusses the boundaries and difficulties of individual self-built corpora with a view to making breakthroughs and progress in future research and development.

1. Introduction

In the last decade, with the rapid development of hardware storage technology and Internet technology, the construction of corpora at home and abroad has also shown a good momentum of development [1]. A search on China Knowledge Network using the keywords "corpus" shows that the number of research papers in this field has increased from 481 in 2008 to more than a thousand at present [2]. However, by entering the keywords "Japanese corpus", only 17 references were retrieved, and the earliest article on Japanese corpus was published in 2009, while scholars had already started to build the JDEST computer corpus of technical English as early as 1982, which indicates that the construction and development of Japanese corpus lags far behind that of English corpus [3, 4]. In the past decade, the National Institute of Linguistics (NIL) has made rapid progress in the construction of corpora and has built a dozen large corpora, such as the Modern Japanese Written Balance Corpus (『現代日本語書き言葉均衡コーパス』), the Modern Japanese Spoken Language Corpus (『日本語話し言葉コーパス』), and the Kokugakuin Japanese Web Corpus (『国語研日本語ウ⊠ブコーパス』) [5]. The corpus is a great convenience for researchers and learners from abroad. The corpus provides rich linguistic materials for language research and has been an important area of linguistic research since the 1960s, especially in the linguistic communities of Western countries [6].

The establishment of English and French corpora has provided great convenience for language application and research. However, in the field of Japanese language research, the construction of Japanese corpus is slightly delayed due to the limitation of the form of Japanese markers. In terms of corpus characteristics, there are

different types of corpus according to different classification methods [7]. The corpus can be classified according to its timeliness and can be divided into a common time corpus and an ephemeral corpus. In terms of selection, the corpus can be divided into a sample corpus and a full-text corpus [8]. In terms of applied genres, the corpus can be divided into single-genre corpus and multigenre corpus. From the point of view of the way of application, it can be divided into written corpus and spoken corpus [9]. The same corpus can be categorized into different categories, such as Brown corpus is a coextensive corpus, a sampling corpus, a multiliterate corpus, a written language corpus, and also a monolingual corpus. Compared to the corpora of European and American languages, the construction of Japanese corpus is relatively late. Although computers were used in data collection such as newspaper terminology surveys in Japan in 1952, such surveys failed to develop smoothly to the stage of building a corpus [10]. In terms of corpus construction, the main force of corpus construction in Europe and America is universities. After the 1994 release of the British National Corpus (BNC), the world's most representative corpus of contemporary English, the Japanese corpus has lagged behind in its development [11]. Although the Japanese corpus has begun to take shape, it is not used very frequently in China at present [12]. At this stage, except for some universities that have started to use the computer-assisted language teaching function and testing function of the corpus, most universities in China lack a comprehensive understanding of the Japanese corpus, which affects its usefulness [13]. Therefore, it needs to be paid attention to and studied.

The corpus can be applied to the teaching of several Japanese language courses, and it is especially effective to enhance the use of the corpus in Japanese audio-visual classes. By showing teaching videos selected from the video corpus that fit with the textbook and explaining how the same word is used in different contexts, students can create contexts in which they can quickly understand and master language knowledge, develop their intercultural communication skills, and achieve the effect of learning by example. As the integration of multimedia technology and Japanese language teaching courses continues to develop, the original audio-visual teaching resources and teaching mode can no longer meet the existing teaching needs. The application of corpus is the main way to solve this problem. By introducing the corpus resources such as CSJ and JV-Finder into the teaching, and by combining "contextual teaching" and "inquiry learning," the teaching effectiveness of Japanese audio-visual classes can be effectively improved. In this paper, we analyze the necessity of using the corpus and propose countermeasures for its application in the context of teaching practice.

2. Related Work

In December 1948, the Ministry of Education, Culture, Sports, Science and Technology established the National Research Institute of Japanese Language. This research institute conducted a lot of research and studies on the Japanese language as early as the 1950s, but due to objective constraints, the construction of the Japanese corpus has been slow for nearly half a century [14]. In the late 1990s, with the rapid development of computer technology, the Japanese corpus entered a period of rapid development. In 2009, I started to build a small-scale, simple teaching corpus, which is now about 20 million words in size. The construction of this corpus is still in progress. The expected goal is to build a small-scale, easy-to-use, and fast-retrieving corpus [15].

The source of the corpus is the articles of some famous authors in the Aozora Library. The software carrier for corpus retrieval is Ant Conc, which has three major functions: word search, word list generation, and subject word [16]. The corpus can be presented in Japanese in this software after encoding and compression, and there is no garbled code phenomenon. The reasons for choosing Aozora Corpus are as follows: first, there is no copyright problem. In Japan, the copyright of any literary work is strictly protected, and no institution or individual can reproduce or use the work without the author's permission. This is the reason why older corpora are commonly used in corpus construction [17]. As mentioned above, the copyright of the corpus collected by the Aozora Corpus has disappeared, and according to the Japanese copyright law, there is no longer any problem of copyright for these works, and they can be used freely. The produced corpus can be made available to teachers and students for free, which greatly facilitates Japanese language learning and research. Second, the corpus is rich.

It has been 13 years since the construction of the Aozora Library began in 1997, and it has reached a considerable size, containing 10,752 literary works. Most of the works of some of these writers have been entered into the database [18]. The wide range of subjects in the Aozora Library, including novels, essays, travelogues, book reviews, memoirs, and various other subjects, ensures the diversity of the corpus and allows for the study of specific linguistic phenomena from a variety of different linguistic materials. Again, the famous authors ensure the scientific nature of the corpus. In the process of corpus extraction, famous literary works by famous authors such as Natsume Soseki, Mori Ouwai, Arishima Takuro, and Tayama Hanabukuro are mainly extracted [19]. In grammar teaching and vocabulary teaching, the corpus can be used to directly retrieve the corpus of famous authors, avoiding the errors caused by teachers' own sentence construction. Moreover, the corpus composed by famous authors has high literary and artistic quality, and in specific teaching, the extraction and study of the corpus can greatly improve the students' Japanese language level and literary literacy.

Firstly, the scale problem: THE construction of a corpus is an extremely large-scale project, and there are limitations in the scale and performance of a corpus built with one's own financial and material resources due to computer memory and storage performance [20]. The Ant Conc search software can only search small-scale databases, and it takes more time to search large-scale databases.

The second issue is the selection of the corpus. The choice of the corpus, whether it is the spoken language used by Japanese people in their daily lives, the dialogues of

characters in Japanese dramas, or the business Japanese used in Japanese companies, plays a crucial role in the nature of the corpus [21]. Due to financial support and manpower problems, this corpus is only a raw corpus and does not involve the issue of corpus balance. Therefore, in terms of material selection, mainly representative novels were chosen. In terms of time span, works from four different periods, Meiji, Taisho, Showa, and Heisei, were selected to cover literature from each historical period, making the corpus more scientific and representative.

Finally, the issue of timeliness of the corpus: the corpus is not the current Japanese language, but rather a literary work that has been fixed in the form of a literary composition and is well known to the Japanese people. A point to be made about this issue is that the timeliness of language materials is not like the shelf life of food, which changes qualitatively in a short time. Language is a relatively fixed thing, as can be well illustrated by the fact that we can read Dream of the Red Chamber, written more than 250 years ago, with relative fluency.

3. Application of Big Data-Based Corpus in Japanese Language Teaching

3.1. Applicable Principles. Teachers must be clear about the learning stage of their students. For example, scientific experiments and judgments must be made about what kind of corpus is appropriate for students at the beginning level and what kind of corpus is appropriate for students at the intermediate and advanced levels. Teachers must be clear about the suitability of the corpus for teaching, whether the selected corpus is suitable for teaching and for what level (or students). After the corpus is compiled, all Japanese teachers are required to discuss and put forward optimization suggestions so that it can meet the needs of Japanese teaching for students of different grades. If the corpus is used randomly for teaching without screening, it may lead to unexpected problems. For example, showing a more complex or deeper corpus to a beginner level student is bound to put more pressure on the student. Students may feel that there are too many vocabulary words, and sentences are too long to understand, which may affect their motivation and have a negative effect. If the selected corpus is consistent with the students' learning stage (level), it will be good to stimulate students' interest and motivate them to actively participate in classroom interaction, which will improve the effectiveness of classroom teaching. Although it is not possible to quantify the amount of the Japanese corpus used in Japanese language teaching, it is important to make it "appropriate." For example, for vocabulary that has multiple meanings and is used in various ways, the corpus can be used to help students experience and understand the meaning and usage of the word in a specific context. For example, some sentence patterns need to be understood in a "context," so we can use the corpus to retrieve the corresponding sentence patterns and help students enter a certain "context." Another example is the teaching of nouns such as "Japan" and "China," where the use of a corpus is not very
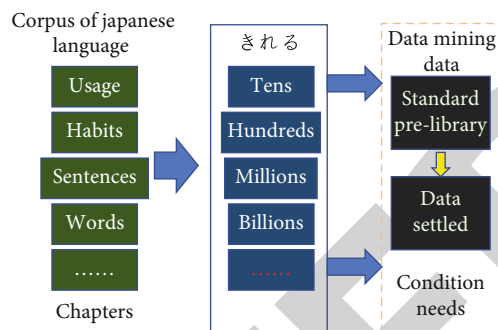

Figure 1: Data mining in corpus.

meaningful. Japanese teachers must have some control over this.

3.2. Combined Application of Big Data and Corpus. The corpus is powerful. By setting search conditions, the search object can be retrieved in the specified corpus package. Therefore, corpus producers tend to pursue the volume of the corpus, thinking that the larger the better. A corpus of "millions," "tens of millions," and "billions" of words is indeed very convincing for studying language phenomena. For example, it proposes the idea of compiling a dictionary of Japanese corpus examples. Taking the Japanese verb "きれる" as an example, Dai retrieved various meanings and usages of "きれる" from a corpus of more than 70 million words and analyzed them. The analysis and study will enable us to supplement the current Japanese dictionary. The larger the corpus, the better it is in terms of its use in Japanese language teaching, if it is intended for general Japanese language learners rather than Japanese language researchers. If a corpus of "millions," "tens of millions," and "billions" of words is called a "pan-corpus" here, then a smaller corpus may not be a good idea. If the corpus of "millions," "tens," and "hundreds" of millions of words is called "pan-corpus," the smaller corpus of "tens" and "hundreds" of thousands of words may be called "microcorpus." From the actual needs of teaching, we can build some "microcorpus" packages, such as "Japanese sentence patterns," "Level 2 vocabulary," and "Kawabata Yasunari's novels." These "microcorpus" packages can be built. These "microcorpus" packages can be used in Japanese language teaching, as shown in Figure 1, because they can be directly linked to the learning points. On this basis, we can also use the "pan-corpus" package to further enrich the teaching content if necessary.

Build a corpus of Japanese language teaching according to the direction of the major. For example, when building a corpus for teaching business Japanese majors, the curriculum and teaching contents of the majors' teaching should be fully considered. In addition to the contents of textbooks, the corpus can also include novels, scripts, and conversation books depicting workplace and business activities without infringing on intellectual property rights. At the same time, in order to make the corpus practical, school-enterprise cooperation can be strengthened, and documents, technical data, and e-mails related to business activities of enterprises in Japan can be included in the corpus without infringing on
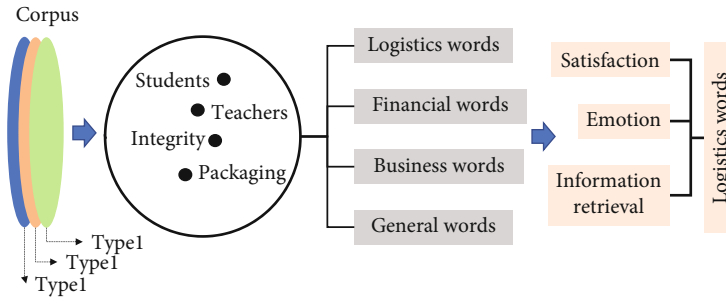
FIGURE 2: Conceptual diagram of the corpus.

commercial secrets and obtaining permission from partner enterprises. Such a corpus is more professional in teaching and is more conducive to the development and improvement of students' language ability and professionalism, so that students can engage in the "real world" of the workplace in the classroom, getting rid of the tedium of the textbook and effectively increasing students' interest in learning. At present, it is difficult to reach a cooperative relationship with Japanese enterprises, and not all Japanese teaching can achieve cooperation with Japanese enterprises.

The database is the basis and data source of the whole corpus, which stores all the corpus, so the creator needs to design a well-structured database model. In order to facilitate the use of the corpus, it is important to store the corpus in a way that is consistent with the user's usage and retrieval habits and not dependent on the specific machine, which is the key to designing a reasonable database. In Japanese, the basic building block of language is words, which form various kinds of utterances and thus chapters. Therefore, when designing a database, the composition of metadata needs to take into account the forms of words, sentences, and chapters. Based on the characteristics of Japanese language composition and the characteristics of the Japanese corpus in high school, the E-R conceptual model shown in Figure 2 is designed as follows.

3.3. Advantages of Japanese Corpus Application in Japanese Language Teaching. In the teaching of basic Japanese, business Japanese, and Japanese general reading courses in the second and third years of JSBC in higher education institutions, I tried to apply the Japanese corpus and found that the application of the Japanese corpus for object retrieval in the teaching of Japanese vocabulary and Japanese sentence patterns was significantly effective. As shown in Figure 3, as a comparison, the Japanese learning performance of students who did not use the corpus was significantly worse than that of students who did use it. It is typical that students learn to use and apply words and sentence patterns rather than memorizing them by heart, for example, "laughed happily." After searching the corpus and observing the words "うれし," and "嬉しく," and "うれしく," we found that they are mostly used in the third person. It was found that when used in the third person, they are mostly "嬉しそうに," "嬉しに," and "うれしそうに," "うれしげに," instead of "嬉しく" and "うれしく." When used in the first person, the opposite is true. Students often

use the conjunctive form of "しく" and "うれしく" when expressing the third person "I laughed happily". In the study of sentence patterns, the corpus often presents contextualized fragments instead of ordinary sentences, which helps students master the language context of the sentence pattern and improve their own language expression. After nearly a year of practice, I found that the corpus has a very positive effect on students' vocabulary memorization and use and on their understanding and application of sentence patterns.

Compared to textbooks, the corpus is richer in language materials, covering all aspects of literature, politics, economics, society, and history. Students not only master language points such as words and sentence patterns in the learning process, but also grasp the inherent patterns through these language points presented in different linguistic contexts. Students master the diversity of Japanese expressions and understand the essence of Japanese language and culture in specific Japanese contexts, so that they can combine various expressions of Japanese according to the context and make their Japanese expressions "authentic." The use of the corpus in Japanese language teaching will effectively promote cultural transfer and provide students with more exposure to Japanese literature, politics, economics, society, history, etc., increasing their understanding and awareness of these areas and enabling them to understand the differences and similarities between Chinese and Japanese cultures. On the basis of students' ability to express themselves correctly and their knowledge and understanding of Japanese culture, they will be able to further improve their intercultural communication skills.

4. Results and Analysis

When the Japanese corpus is applied to teaching, since the presented corpus is often a fragment, students must pay attention to the content (sentences) before and after the target in order to understand and master the target content more deeply during the learning process. In this case, the linguistic logic of the preceding and following texts is perceived and comprehended overtime. After such a learning process, it is more conducive for students to develop chapter awareness. At the same time, the corpus often presents a variety of topics, so that students can "see more and more" of each type of text. When reading different texts, students are less likely to resist reading and remain more interested
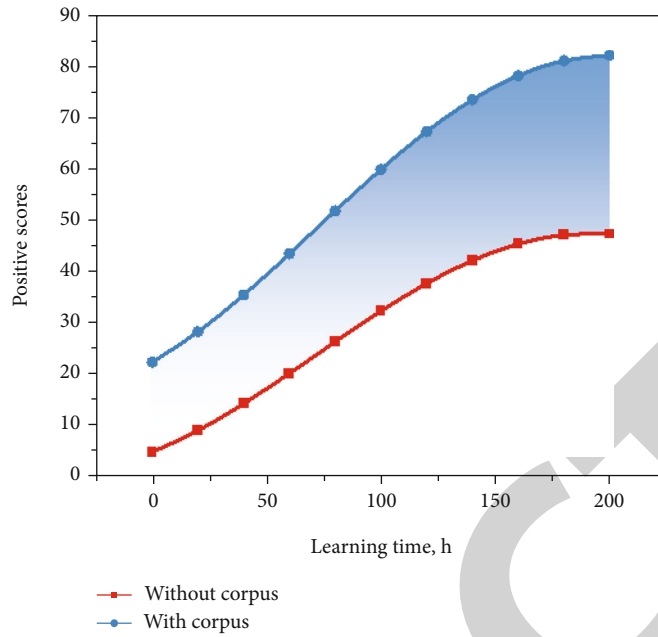
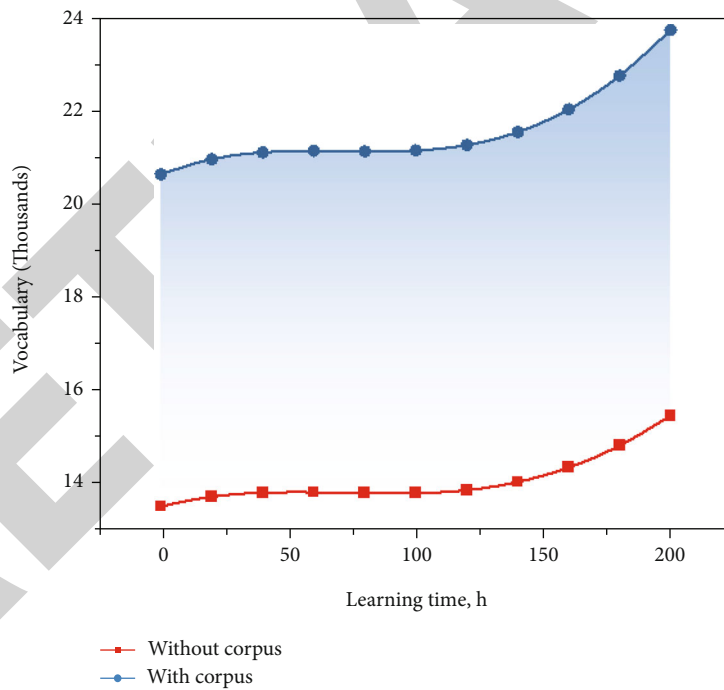FIGURE 3: Comparison of Japanese language learning performance.

FIGURE 4: Changes in students' vocabulary.

in reading. In addition, reading through the corpus will also improve students' vocabulary, as shown in Figure 4. Taking all factors together, the Japanese corpus applied to teaching helps students improve their reading skills. In the process of creating and using the Japanese corpus, teachers will make many new discoveries, encounter many new problems, and bring about many new thoughts. Teachers continue to gain new knowledge and enrich themselves, explore ways to solve problems, improve their teaching and research abilities, and

engage in deep thinking to find a new starting point for further work in the future.

The application of the Japanese corpus in Japanese language teaching has undoubtedly contributed to the improvement of Japanese learners' Japanese language proficiency, intercultural communication skills, and professionalism, as well as to the improvement of teachers' teaching and research abilities. However, due to the differences in individual teachers' knowledge, acceptance, production, and use of
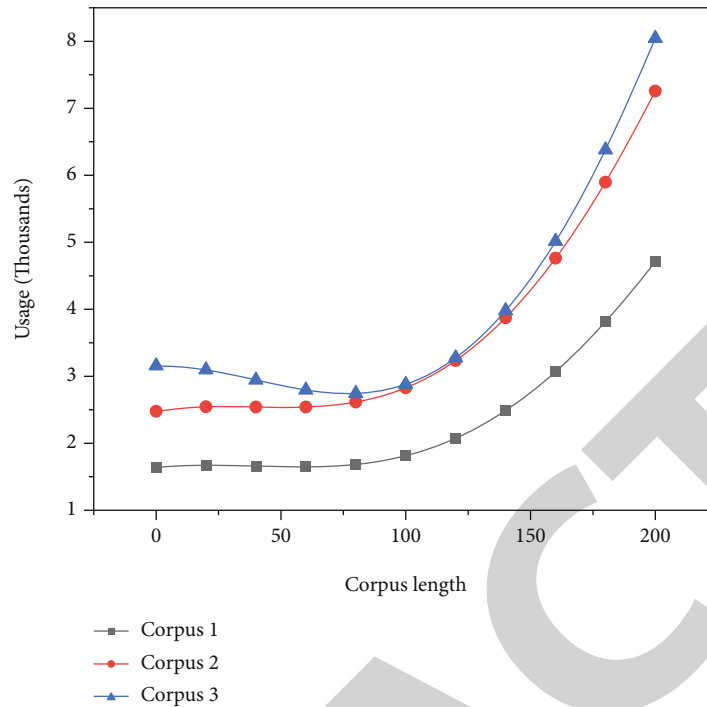
Figure 5: Corpus usage.

the Japanese corpus, the popularization of the use of the Japanese corpus in Japanese language teaching has been difficult to achieve for a while, and it is basically used by individual teachers at present. At present, there is an urgent need for colleges and universities to organize their own Japanese teachers to edit and use the corpus. It is an extremely long process to first promote the use of the corpus in the University and then carry out the application and communication of the Japanese Corpus in different schools and gradually realize the promotion of the Japanese Corpus in Japanese teaching. As shown in Figure 5, there is a large difference between the usage of the corpus by teachers and students. Meanwhile, the construction of "microcorpus" and "professional corpus" still needs the participation and support of more Japanese language educators and school-enterprise cooperation enterprises. For example, the corpus of ERD is mainly from news and magazines, the corpus of spoken language is mainly from audio and speech, and the corpus of Chinese and Japanese translation is mainly for linguistic research. Among the many corpora that have been established, there is no specific corpus for Japanese language teaching, so it is essential to establish a Japanese corpus where the corpus material comes from the teaching process. The introduction of a corpus in the Japanese language teaching process can make up for many of the deficiencies in the language teaching classroom, and the use of corpus retrieval methods for language teaching has advantages over traditional classroom teaching, mainly in the following ways.

(1) The corpus enables students to integrate context into the learning process, so that they no longer rely solely on teachers' subjective descriptions and perceptions and therefore better reflect the realities of language use and make the language knowledge they learn more relevant

(2) The corpus can foster students' self-exploration in the learning process, stimulate their learning initiative, and turn learning into research

(3) The corpus contains many corpus used in teaching, which can visually show the use of language units such as vocabulary, corpus sentences and chapters, and the frequency of examinations, so that students can learn to explore language patterns on their own

In the process of building a corpus, the collection and processing of the corpus is one of the fundamental and most important tasks. Since the corpus in the corpus should be continuously updated, the creator needs to continuously collect new corpus and add it to the corpus throughout the corpus survival cycle, as shown in Figure 6. At the same time, the higher-level Japanese corpus is a specialized corpus, and the difficulty level and specialized characteristics of the selected corpus should be considered in the selection of the corpus. The corpus in the teaching-based Japanese text corpus mainly comes from the exercises and test papers used by teachers and students in some of the senior high school institutions in their daily study and also covers some of the test questions of the Japanese Language Application Proficiency Test Level 2. At present, the collected corpus is mainly textual, and the spoken corpus such as audio will be collected gradually in the process of using the corpus. In order to make the corpus better serve the Japanese language
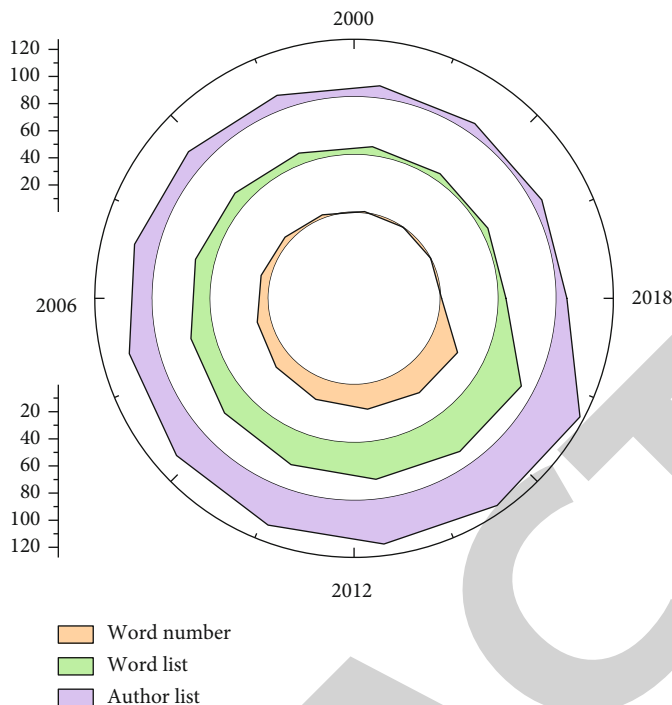
Figure 6: Lexical updates in the corpus.

teaching, the following matters should be noted when collecting the corpus.

(1) Locating a reasonable language domain. When collecting the corpus, it is necessary to select the corpus of the appropriate language domain for different types of language courses, such as choosing written terms or literary works as language materials for advanced Japanese reading courses, but not everyday conversation materials

(2) Selecting a typical corpus. Since the purpose of Japanese language teaching is to develop students' English language skills, the selection of the corpus must pay attention to its practicality and typicality

(3) Grasp the difficulty of the corpus. The selection of the teaching corpus must grasp the difficulty level of the corpus. Choosing a corpus that is too difficult will lead to a decrease in students' interest in learning, while a corpus that is too simple will not be able to serve the corresponding purpose. To ensure the typicality of the corpus, the creator should collect the corpus from a variety of text sources with different styles and specialized topics; to ensure the practicality of the corpus, the corpus can be selected from various kinds of test and examination questions used by students in their daily learning process

In this corpus, Microsoft SQL Server 2008 is used as the underlying database, considering the scale and generality of the corpus. When defining the two-dimensional table structure, the vocabulary, corpus sentences, chapters, authors, and other entities are defined as a two-dimensional table,

for example, the vocabulary table should contain "name," "form," and "similar words." The specific two-dimensional table design is shown in Figure 7.

It should be noted that since several of the software programs used to create the Japanese corpus are in Japanese, machines with Chinese operating systems need to change the language environment of the machine first. In addition, the higher the hardware and software configuration of the user's machine, the faster the corpus retrieval will be, as shown in Figure 8. A reasonable database design is the basis for building a good corpus, and subsequent operations such as corpus retrieval and update will involve the operation of metadata in the database. With the continuous development of computer storage technology and big data retrieval technology, a large amount of Japanese learning corpus is flooded on the Internet, which greatly opens up the horizons of Japanese learners and enriches their learning tools. However, there is an urgent need to solve the problem of how to sift through such a large amount of corpus information and effectively improve our learning efficiency. Based on Ant Conc search software, this paper selects representative novels from the Aozora Library to create a simple corpus to serve the teaching of Japanese interpretation and improve the teaching effectiveness. At the same time, this paper discusses the boundaries and difficulties of individual self-built corpora, with a view to making breakthroughs and progress in future research and development.

Corpora provide rich linguistic material for language research and have been an important area of linguistic research worldwide, especially in developed Western countries, since the 1960s. The establishment of English and French corpora has provided great convenience for language application and research. However, in the field of Japanese
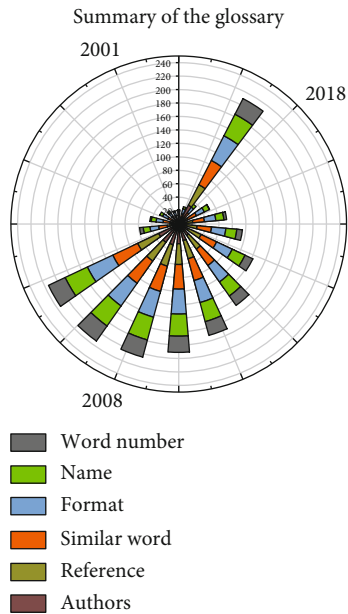
Summary of the glossary
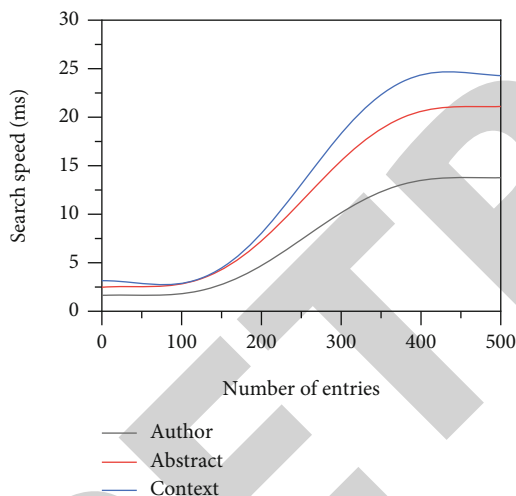

FIGURE 7: Summary of the glossary.

FIGURE 8: Variation of corpus search speed.

language research, the construction of Japanese corpus is slightly delayed due to the limitation of Japanese markup format. In terms of corpus characteristics, there are different types of corpus according to different classification methods. The corpus can be classified according to its timeliness and can be divided into a common time corpus and an ephemeral corpus. In terms of selection, the corpus can be divided into a sample corpus and a full-text corpus. In terms of applied genres, the corpus can be divided into single-genre corpus and multigenre corpus. From the point of view of the way of application, it can be divided into written corpus and spoken corpus. The same corpus can be classified into different categories, such as Brown corpus is a coextensive corpus, a sampling corpus, a multitext corpus, a written language corpus, and also a monolingual corpus.

5. Conclusions

(1) This paper presents a small text-based corpus for Japanese language teaching, which is derived from the teaching process. The corpus can be created on the machine by software, and the collected corpus can be imported into the corpus for use, which provides another way to improve students' interest and learning ability

(2) With the continuous development of the integration of multimedia technology and Japanese teaching courses, the original audio-visual teaching resources and teaching models can no longer meet the existing teaching needs. The application of corpus is the main way to solve this problem. Introducing CSJ, JV finder, and other corpus resources into teaching and combining "situational teaching" with "Inquiry Learning" can effectively improve the teaching effect of Japanese audio-visual class. The corpus application results in this paper show that the use of corpus in teaching can improve students' Japanese learning achievements, greatly improve students' Japanese vocabulary, and improve students' efficiency in learning Japanese

(3) In this paper, we analyze the necessity of using corpus and propose countermeasures for its application in the context of teaching practice. There are still some problems in the editing, content selection, and regular updating of the corpus. Solving these problems will help promote the use of the Japanese Corpus and realize the optimization of Japanese teaching

Data Availability

The figures and tables used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] S.-j. Lee and S.-J. Kim, "The characteristics of organic brands using big data corpus analysis-based on sanitary pads in Korea," *ACM Journal of Computer Documentation*, vol. 70, pp. 203–217, 2020.

[2] X. Xu, "Retracted article: machine learning-based prediction of urban soil environment and corpus translation teaching," *Arabian Journal of Geosciences*, vol. 14, no. 11, pp. 1–15, 2021.

[3] S.-J. Kim and S. Lee, "A study of the brands characteristics using big data corpus analysis-focused on South Korean

sanitary pad brands," *Archives of Design Research*, vol. 32, no. 4, pp. 113–127, 2019.

[4] K. M. Lee, C.-S. Han, K.-I. Kim, and S. H. Lee, "Word recommendation for English composition using big corpus data processing," *Cluster Computing*, vol. 22, Supplement 1, pp. 1911–1924, 2019.

[5] C. Holcomb and D. A. Buell, "First-year composition as "big data": towards examining student revisions at scale," *Computers and Composition*, vol. 48, pp. 49–66, 2018.

[6] B. Li, A. Drozd, Y. Guo, T. Liu, S. Matsuoka, and X. Du, "Scaling Word2vec on big corpus," *Data science and engineering*, vol. 4, no. 2, pp. 157–175, 2019.

[7] J. Wan, P. Zheng, H. Si, N. N. Xiong, W. Zhang, and A. V. Vasilakos, "An artificial intelligence driven multi-feature extraction scheme for big data detection," *ieee access*, vol. 7, pp. 80122–80132, 2019.

[8] W. Gong, "An innovative English teaching system based on computer aided technology and corpus management," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 14, no. 14, pp. 69–80, 2019.

[9] N. Li, W. Luo, K. Yang, F. Zhuang, Q. He, and Z. Shi, "Self-organizing weighted incremental probabilistic latent semantic analysis," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 12, pp. 1987–1998, 2018.

[10] L. Changbong, "Multicultural aspects analysis using news articles corpus," *Journal of Humanities and Social Sciences*, vol. 12, no. 1, pp. 1405–1418, 2021.

[11] S. B. R. Kumar, K. Wang, and S.-M. Shen, "Corpus-based topic derivation and timestamp-based popular hashtag prediction in Twitter," *Journal of Information Science and Engineering*, vol. 35, no. 3, pp. 675–696, 2019.

[12] J. Yao, H. Chen, and Y. Liu, "Research on constructing "parallel contrast corpus of grammatical errors"," *Journal of Language Teaching and Research*, vol. 11, no. 5, pp. 756–763, 2020.

[13] K. S. Sarfo-Kantankah, "Corruption is a big issue: a corpus-assisted study of the discursive construction of corruption in Ghanaian parliamentary discourse," *Legon Journal of the Humanities*, vol. 29, no. 1, pp. 226–258, 2018.

[14] M. Davies and J.-B. Kim, "The advantages and challenges of "big data": insights from the 14 billion word Iweb corpus," *Linguistic Research*, vol. 36, no. 1, pp. 1–34, 2019.

[15] W. Zhang, Y.-C. Zhu, and J.-P. Wang, "An intelligent textual corpus big data computing approach for lexicons construction and sentiment classification of public emergency events," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30159–30174, 2019.

[16] V. Tournay, M. Jacomy, A. Necula, A. Leibing, and A. Blasimme, "A new web-based big data analytics for dynamic public opinion mapping in digital networks on contested biotechnology fields," *Omics A Journal of Integrative Biology*, vol. 24, no. 1, pp. 29–42, 2020.

[17] S. Wu, "Subtitling swear words from English into Chinese: a corpus-based study of *big little lies*," *Open Journal of Modern Linguistics*, vol. 11, no. 2, pp. 277–290, 2021.

[18] J. Lou and Y. Zhang, "Semantic change analysis of Korean verbs based on massive culture corpus data," *Personal and Ubiquitous Computing*, vol. 24, no. 1, pp. 115–125, 2020.

[19] E. Arık and B. T. Arık, "The concept of love from the perspectives of big data and corpus semantics: meaning, reference, and metaphor," *Scientific Programming*, vol. 39, no. 1, 178 pages, 2019.

[20] Y. Chen and F. Yan, "International visibility as determinants of foreign direct investment: an empirical study of Chinese provinces," *Social Science Research*, vol. 76, pp. 23–39, 2018.

[21] J. R. Gallagher, Y. Chen, K. Wagner, X. Wang, J. Zeng, and A. L. Kong, "Peering into the Internet Abyss: using big data audience analysis to understand online comments," *Technical Communication Quarterly*, vol. 29, no. 2, pp. 155–173, 2020.