Corpus-based Periodization of Japanese Kanbun Texts Based on Two-character Lexical Usage

Zifan Wu[1,2], Cheng Jin[3]
[1]The Graduate University for Advanced Studies, SOKENDAI; [2]National Institute for Japanese Language and Linguistics;
[3]Shanghai Jiaotong University
Tokyo 190-8561, Japan; Shanghai 200030, China
gs2024r007@ninjal.ac.jp; jincheng25@sjtu.edu.cn

Abstract

This study proposes a corpus-based periodization of Japanese Kanbun texts based on changes in two-character lexical usage. Traditional periodization, including the Ōchō Bungaku, Gozan Bungaku, and Edo Bungaku periods, has mainly been established from a literary-historical perspective. By contrast, this study examines a broader range of Japanese Kanbun texts and explores a quantitative, language-internal basis for periodization.

The analysis uses 691 volumes from the Earth Corpus, covering the eighth to nineteenth centuries and comprising approximately 5.17 million characters. Because reliable word segmentation remains difficult for Kanbun, candidate two-character strings were extracted as 2-grams. Significant lexical units were then selected using frequency, T-score, Z-score, mutual information, and logDice, yielding 1,558 items. Their century-based frequency distributions were analyzed through multiple comparison tests, K-means clustering, correlation analysis, principal component analysis, and hierarchical clustering.

The results indicate three major lexical stages: the 8th–10th, 11th–13th, and 14th–19th centuries. The 10th century represents a particularly important lexical watershed. These findings demonstrate that quantitative lexical analysis can complement traditional literary history and provide a new perspective for periodizing Japanese Kanbun.

Keywords: Japanese Kanbun, periodization, 2-grams, quantitative analysis

1. Introduction

Japanese Kanbun literature is an important cultural exchange between China and Japan, and an important academic resource for the study of East Asian culture. Historical segmentation is a major topic of historical research, and traditional studies divide the history of Japanese Kanbun before the Meiji period into three periods: the Ōchō Bungaku period, the Gozan Bungaku period and the Edo Bungaku period, from a literary-historical perspective (Yoshida, 1932; Inoguchi, 1980; Chen, 2011; Ma, 2003).

This study attempts to establish a comprehensive historical periodization of Japanese Kanbun on the basis of broad patterns of language use. Using the frequencies of two-character strings extracted from a large-scale corpus of East Asian Kanbun texts, we adopt a quantitative, statistically grounded approach and employ computer programming to conduct an automated cluster analysis of historical Japanese Kanbun texts.

The findings of this study may provide new methods and quantitative evidence for research on the vocabulary and historical development of Japanese Kanbun. They may also offer a new perspective on the spread of Literary Chinese and Chinese-character-based written culture across East Asia.

2. Related Research

In the first section, we have briefly described the traditional study of the staging of the history of Japanese Kanbun literature. Here, we will sort out studies that deal with the staging of the history of Japanese Kanbun literature in short. Early studies divided the history of Japanese Kanbun literature into four periods: the ancient period, the mid-ancient period, the late ancient period, and the modern period (Haga, 1928). This classification has clear hints of political history and is relatively simple. The practice of dividing the history of Japanese Kanbun literature before the Meiji period into four periods emerged in the mid-twentieth century, which can be summarized as the Age of Court Literature, the Age of Noble Literature, the Age of Monastic Literature, and the Age of Confucian Literature (Toda, 1957; Ogata, 1961). This classification differs in the specific periods into which they are divided, but the overall division is still in accordance with the staging of political history. In the second half of the 20th century, authoritative studies divided the history of Japanese Kanbun literature into three periods: the Ōchō Bungaku period, the Gozan Bungaku period, and the Edo Bungaku period (Inoguchi, 1980). This classification has also been affirmed and followed by Chinese researchers (Chen, 2011; Ma, 2003). The beginning of the Gozan Bungaku period almost coincided with the establishment of the Kamakura shogunate, as did the Edo Bungaku period and the Tokugawa shogunate. We can see that traditional research on the staging of literary history lacks the support of quantitative data, which becomes a starting point for this article.

Unlike the English text, the processing of Chinese text faces the problem of word segmentation. The commonly used Chinese and Japanese word segmentation tools do not meet the requirements of Kanbun. Therefore, this paper refers to the methods used in some corpus studies to obtain classical Chinese (or Kanbun) two-character words. Luo and Sun (2003) tested the performance of nine widely adopted statistical measures of such kind in Chinese word extraction on the individual basis and suggested that mutual information will be an effective way in Chinese word extraction. In this paper, we mainly refer to N-gram theory, and statistical methods based on frequency, mutual information, and hypothesis testing to obtain classic Chinese two-character words (Duan et al.,2012). We also tried to apply statistical measures used to determine keyword collocations and the typicality of word

combinations, such as Z-scores, to the extraction of two-character words (Wei, 2002). In addition, T-scores, Dice coefficients, and logDice coefficients, which are lexical association scores, were also applied to this attempt at word segmentation (Ishikawa, 2008; Rychly, 2008). The results of the word segmentation obtained by statistical methods will be the subject of this paper.

The use of large-scale corpora and statistical methods has become increasingly common in empirical linguistic research. Among these methods, cluster analysis has been applied to the comparison, classification, and diachronic analysis of linguistic data. Previous studies have used various clustering techniques to identify developmental stages in individual linguistic phenomena, to periodize modern Chinese on the basis of newspaper corpora, and to compare textual styles through word-frequency patterns (Gries & Hilpert, 2012; Gries & Stoll, 2009; Rao & Li, 2017; Fang & Lin, 2022). These studies demonstrate the potential of clustering methods for diachronic linguistic research. However, such statistical clustering methods have not been widely applied to the classification and periodization of Japanese Kanbun texts. The present study therefore applies cluster analysis to two-character frequency data in order to examine the historical periodization of Japanese Kanbun from a quantitative perspective.

3. Materials and Methods

Next, we present the corpus used in this study and how we extracted the target two character words from it (Section 3.1). Then we describe the statistical analysis used in this paper (Section 3.2). In this paper, the letters 'A-L' in the charts and figures represent the 8th century to the 19th century, respectively.

3.1 Corpus and Word Extraction

This paper uses the 691 volumes of Japanese Kanbun literature from the 8th to the 19th centuries collected in the Earth Corpus β version as data set. The corpus is the first diachronic corpus for the study of Chinese words (Wang, 2024). Its contents cover most of the important Japanese Kanbun works of different eras. The total number of characters in the data set is 5,174,103. The numbers of characters and documents in each century are shown in Figures 1 and 2. The corpus has been balanced in terms of content, and we can see that the total number of characters is at a relatively similar level across the centuries. However, in terms of the number of documents, the number of documents in the 15th and 16th century is lower than in the other centuries due to the inherent differences in the number of Kanbun works in different periods. Overall, however, the amount of data used in this study is very large and suitable for analysis using statistical methods.

A 2-gram is a string of characters containing two adjacent characters. In this paper, all 2-grams occurring in Japanese Kanbun works are used as candidate sets for two-character words, and statistical methods are used for automatic extraction of two-character words[1]. By 2-gram processing of all texts in the corpus, a total of 1048576 different two-character strings were obtained. Among them, 4386 were used with a frequency of 100 or more. By calculating the T-score, Z-score, mutual information (MI), and LogDice coefficient (LD) of these 2-grams, we obtained the internal associative strength of character strings. The significant collocations were filtered according to the statistical values with the following conditions: T-score 1.645, Z-score 2, MI 3, LD 5. Finally, 1558 significant collocations of two-character words were obtained (Table 1).

two-character words	Frequency (≥100)	Z-score	T-score	MI	LD
朝臣	5166	713.73	71.52	7.65	12.84
天皇	2678	353.65	51.22	6.60	11.62
天下	2267	165.38	45.91	4.80	10.82
大臣	2065	159.87	43.85	4.84	10.62
今日	1580	122.16	37.99	4.50	10.29
和尚	1545	444.75	39.16	8.02	12.00
菩薩	1464	1116.07	38.24	10.74	13.46
上人	1422	57.35	33.05	3.02	9.54
先生	1219	189.53	34.36	5.97	10.85
日本	1091	112.03	31.80	4.75	9.95
omitted below					

Table 1: An illustration of 2-character words extraction.

In statistical studies, stable and reliable data usually has great value of the reliability coefficient. Cronbach coefficient of 0.7-0.8 indicates high reliability of the scale, and 0.8-0.9 indicates very good reliability of the scale. According to the reliability test conducted by RStudio, the Cronbach's Alpha coefficient of the total volume table was 0.7679975, which exceeded the minimum acceptable level of 0.7, indicating good internal consistency and reliability of the data, which could be used for statistical inference research.

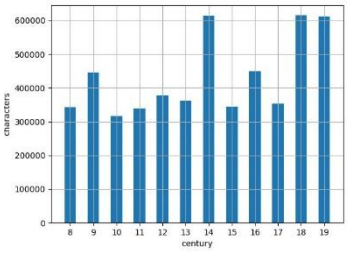

Figure 1: Numbers of characters each century.

[1] In this study, the term "two-character word" is used operationally to refer to a statistically selected two-character lexical unit, rather than a word identified through linguistic word segmentation.

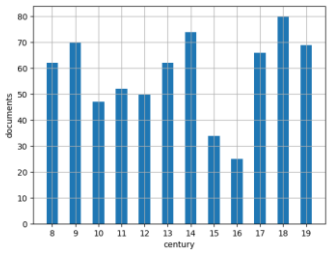

Figure 2: Numbers of documents each century.

3.2 Cluster Analysis Algorithm and Preprocessing

In this paper, we use K-means algorithm and hierarchical method to classify the texts in the corpus, and RStudio is used to complete the clustering. Steinhaus in 1956 (multidimensional data), Lloyds in 1957 (computer and pattern recognition field 'Lloyd algorithm'), Ball and Hall in 1965 and McQueen in 1967 (first named 'k-means') independently proposed K-means clustering algorithm in their respective fields of scientific research. Then K-means clustering algorithm has been widely studied and applied in different areas, and a large number of different improved algorithms have been developed. Although K-means clustering algorithm has been proposed for more than 50 years, it is still one of the most widely used partition clustering algorithms (Lloyd, 1957; Ball and Hall, 1965).

K-means algorithm is a typical distance-based clustering algorithm. It adopts distance as the evaluation index of similarity, that is, it believes that two closer objects have more similarities. The algorithm considers that a class is composed of objects close to each other, so the final goal is to get compact and independent classes. The algorithm essentially consists of two phases. First, the algorithm assigns cases to initial k classes. Second, the assignment is updated by adjusting the boundaries of the classes. The boundaries are based on the cases that fall into the current class. Repeat the update and allocation process several times until the changes are no longer improving the class's goodness. Then the process stops and the final clusters are obtained.

Assume that the characteristic space is a n-dimensional real vector space R^n. For vectors $x_i, x_j \in X, x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^T, x_j = \left(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)}\right)^T$, the Minkowski distance of x_i and x_j is defined as follows:

$$L_p(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}}$$

It is given that $p \geqslant 1$.

It is called Euclidean distance for $p = 2$:

$$L_2(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^2\right)^{\frac{1}{2}}$$

It is called Manhattan distance for $p = 1$:

$$L_1(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|\right)$$

It is called Chebyshev distance for $p = \infty$:

$$L_\infty(x_i, x_j) = \max_l \left|x_i^{(l)} - x_j^{(l)}\right|$$

In this paper, various distances are tested to classify the 2-character words in the corpus. Finally we adopt Euclidean distance which best matches the texts.

As for the hierarchical method, it performs a hierarchical decomposition of a given dataset until some condition is met. Specifically, it can be divided into two schemes: 'bottom-up' and 'top-down'. In a bottom-up scenario, each data point starts with a separate group, and in subsequent iterations, the combination of proximity is combined into a group by a certain distance measure until all records form a group or meet a certain condition. Representative algorithms are: BRCH algorithm, CURE algorithm, CHAMELEON algorithm, etc. This method is often applied to the merging stage after taking K-means algorithm with a relatively large K.

Figure 3 shows the Rank-frequency distribution of the two-character words in Japanese Kanbun literature from the 8th to the 19th centuries, and it can be seen that the word frequency of the top 50 and the back show a cliff-like drop. It stabilized after 300. From the boxplot of total frequency (Figure 4), it can also be seen that there are more data singularities, and the words that appear less frequently are densely distributed. Figure 5 shows the Rank-frequency distribution of the two-character words in Japanese Kanbun literature by century（The letters 'A-L' represent the 8th century to the 19th century, respectively), and the word ordering method is the same as the total word frequency (Figure 3), which shows that the high-frequency words in each century are similar, but the relatively low-frequency vocabulary has a distinct era style. From the boxplot of frequency by century (Figure 6), we can more intuitively discover the differences between eras. Both the word frequency interval and the degree of dispersion show different levels.

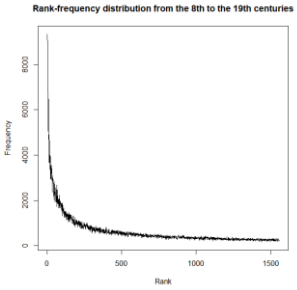

Figure 3: Rank-frequency distribution of the two-character words from the 8th to the 19th centuries.

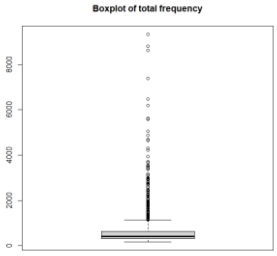

Figure 4: Boxplot of total frequency of the two-character words from the 8th to the 19th centuries.

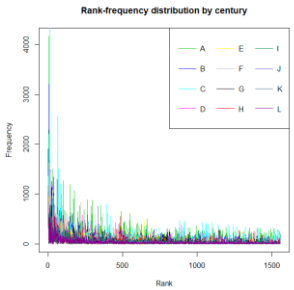

Figure 5: Rank-frequency distribution of the two-character words by century.

Through Bartlett's test for testing homogeneity of variance of the word frequency, we found significant differences in word frequency between eras. Analysis of variance and multiple comparisons are performed below. The results of multiple comparisons are shown in Figure 7 (red shows no significant difference, blue shows significant difference). By clustering eras with no significant differences through multiple comparisons, the following groupings can be obtained: (8, 9, 10, 13), (11,12), (14, 15, 16, 17, 18, 19). Excluding the revival of classical literature, the preliminary chronology should be divided into the 8th-10th centuries, the 11th-12th centuries, the 13th century, and the 14th-19th centuries. We find significant differences between the pre-middle and late Ōchō Bungaku period. The early appearance of the Gozan Bungaku period is also related to this, and the early style of the Gozan Bungaku period in the thirteenth century may have been a certain degree of literary revival at the beginning of the Ōchō Bungaku period, and then began its own stylistic evolution. The style of the Edo Bungaku period is not significantly different from that of the late Gozan Bungaku period.

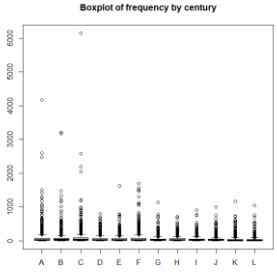

Figure 6: Boxplot of frequency of the two-character words by century.

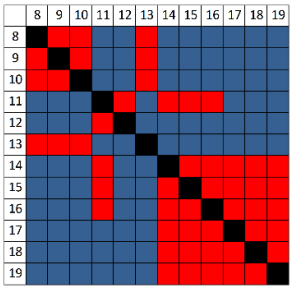

Figure 7: Multiple comparisons of significant differences in word frequency across eras.

4. Clustering

In order to further explore the differences and commonalities in the usage of words in different eras, we first chose the K-means algorithm to cluster the texts in the corpus in this paper. We started clustering after setting the number of clusters to three and obtained the following results: from the 8th to the 10th century, from the 11th to the 13th century, and from the 14th to the 19th century. We then re-determined the number of clusters using the scree-plot (Figure 8).

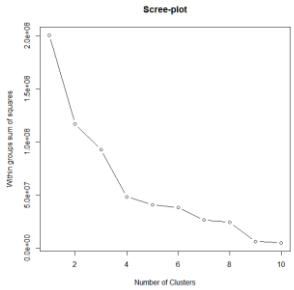

Figure 8: Scree-plot.

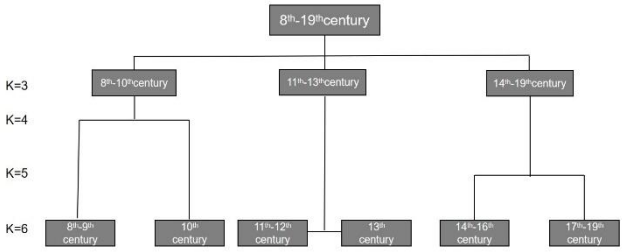

Figure 9: An illustration of the epoch division.

Increasing the number of cluster centres, we find that when the number of cluster centers reaches 4, the 10th century is cut off from the Ōchō Bungaku period. When the number of cluster centers reaches 5, the division of the Edo Bungaku period at the 16th century will be represented. And when the number of cluster centers reaches 6, the 13th century of the pre-Gozan Bungaku period is cut off separately (Figure 9). Furthermore, the data suggests a clear difference in literary style around the 10th century, saying that the split between the early-middle and late Ōchō Bungaku period is even more significant than the separation between the three major periods.

We find that the highest correlation between the 10th century and any other century is with the 9th century (0.47), followed by the 8th century (0.35), so we believe that the 10th century should belong to an upper cluster branch together with the 8th-9th centuries. However, the highest value of only 0.47 is still a relatively low level, suggesting that Japanese Kanbun literature of the 10th century was somewhat specific in its lexical style and that the century was a crucial transitional period in the history of Japanese Kanbun literature.The highest correlation between the 13th century and any other century is with the 12th century (0.66), followed by the 11th century (0.56), so we believe

that the 13th century should belong to the same upper cluster as the 11th- 12th centuries.

In addition, as a validation, the word frequencies of each century were clustered using hierarchical clustering to verify the above staging. Figure 10 shows the results of this hierarchical clustering, which are broadly in line with the above conclusions, with the 10th century still showing a strong specificity.

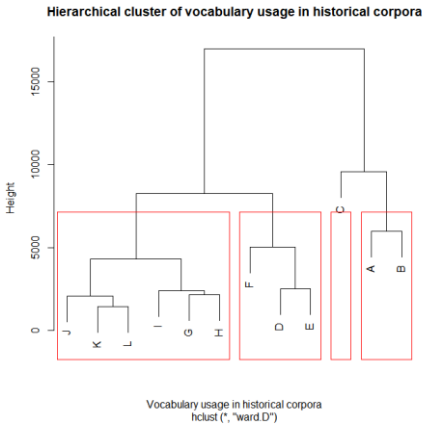

Figure 10: Hierarchical cluster of vocabulary usage in historical corpora.

	A	B	C	D	E	F	G	H	I	J	K	L
A	1.00	0.66	0.35	0.21	0.24	0.08	0.00	0.03	0.02	0.20	0.03	0.08
B	0.66	1.00	0.47	0.40	0.42	0.20	0.08	0.10	0.09	0.26	0.08	0.14
C	0.35	0.47	1.00	0.27	0.29	0.21	0.03	0.06	0.05	0.15	0.03	0.06
D	0.21	0.40	0.27	1.00	0.73	0.56	0.30	0.33	0.31	0.43	0.32	0.38
E	0.24	0.42	0.29	0.73	1.00	0.66	0.20	0.28	0.23	0.31	0.17	0.23
F	0.08	0.20	0.21	0.56	0.66	1.00	0.28	0.30	0.29	0.26	0.14	0.19
G	0.00	0.08	0.03	0.30	0.20	0.28	1.00	0.69	0.58	0.52	0.58	0.55
H	0.03	0.10	0.06	0.33	0.28	0.30	0.69	1.00	0.69	0.47	0.48	0.48
I	0.02	0.09	0.05	0.31	0.23	0.29	0.58	0.69	1.00	0.42	0.39	0.42
J	0.20	0.26	0.15	0.43	0.31	0.26	0.52	0.47	0.42	1.00	0.73	0.78
K	0.03	0.08	0.03	0.32	0.17	0.14	0.58	0.48	0.39	0.73	1.00	0.86
L	0.08	0.14	0.06	0.38	0.23	0.19	0.55	0.48	0.42	0.78	0.86	1.00

Table 2: Relevance matrix of the frequency of the two-character words in Japanese Kanbun literature from the 8th to the 19th centuries

5. Results

Based on the clustering results, we suggest to divide the history of Japanese Kanbun literature from the 8th to the 19th centuries into three periods from the point of view of lexical usage. The first period is from the 8th century to the 10th century, which can be further divided into two smaller periods, 1-1 for the 8th century to the 9th century and 1-2 for the 10th century. The second period, from the 11th century to the 13th century, can be subdivided into a 2-1 period from the 11th century to the 12th century and a 2-2 period consisting of the 13th century alone. The third period is from the 14th century to the 19th century, which can also be divided into two segments, 3-1 from the 14th century to the 16th century and 3-2 from the 17th century to the 19th century. Figure 11 briefly illustrates the staging. Compared to the traditional staging of the history of Japanese Kanbun literature, the staging of the three major periods in this paper has three main points of difference. One is the split in the Ōchō Bungaku period at around the 10th century. In terms of data, the differences in vocabulary usage around the 10th century are actually much greater than the differences between any two of the three major periods of traditional staging. And the 10th century itself has certain peculiarities. We therefore argue that the staging of the first and second periods reflects the fact that the 10th century was an important turning point and

transitional period in the history of Japanese Kanbun literature.

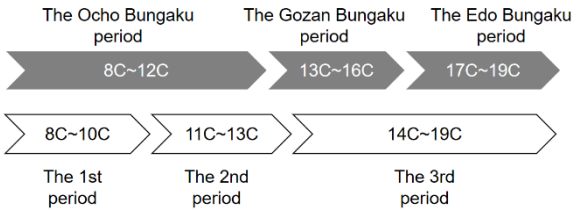

Figure 11: An illustration of the result of the epoch division.

Another notable difference concerns the conventional boundary between the Gozan Bungaku and Edo Bungaku periods. In the present analysis, the 14th–19th centuries show relatively similar patterns of lexical usage and are grouped into the same major cluster. By contrast, the 13th century, conventionally associated with the early Gozan Bungaku period, shows greater similarity to the 11th–12th centuries and is therefore classified within the second period. This suggests that diachronic changes in lexical usage were gradual and did not necessarily coincide with conventional literary-historical or political-historical boundaries.

6. Conclusion

In this paper, we have used corpus and statistical methods to conduct a period staging of the history of Japanese Kanbun literature. First we used statistical methods to extract 2 character words from the Japanese Kanbun corpus and obtained 1,558 statistically selected 2-character words as the survey data set. We conducted multiple comparisons of the data from the 8th to the 19th centuries in terms of lexical usage, and the results showed significant differences from traditional staging. In response, we further performed K-means clustering. We found that when the number of cluster centers was 3, the staging results obtained were closer to the results of multiple comparisons and differed more from traditional staging. The boundaries of the three major periods in traditional staging are only fully presented when the number of cluster centers reaches 6. At the same time, however, the 10th century and the 13th century are cut off separately. For the specificity of these two centuries, a correlation matrix was calculated and produced for the usage of vocabulary in each century. According to the correlations, the 10th century is closer to the 8th-9th centuries, but still maintains a strong specificity. The 13th century, on the other hand, is closer to the 11th-12th cen turies. By means of hierarchical clustering we also obtained similar conclusions. Finally, based on patterns of two-character lexical usage, we divide the history of Japanese Kanbun into three periods: the first period (the 8th–10th centuries), the second period (the 11th–13th centuries), and the third period (the 14th–19th centuries). The first and second periods are separated by a marked change in lexical usage around the 10th century, while the 13th century shows transitional characteristics between the second and third periods. The third period is characterized by relatively high lexical similarity across centuries. These results suggest that the 10th century represents an important watershed in the lexical history of Japanese Kanbun..

This study focuses on the period staging based on quantitative analysis. The three stages identified here should be understood primarily as stages of lexical similarity in the present corpus, rather than as a replacement for established literary periodization. Further quantitative and qualitative analyses can give us more insight into the question of what char acterized the usage of vocabulary in these periods, and this could be a path for future work.

7. Acknowledgements

This work was supported by JST SPRING, Japan Grant Number JPMJSP2104. We thank Ding Wang for providing us with the corpus.

8. Bibliographical References

A, Inoguchi. (1980). 日本漢詩鑑賞辞典. Tokyo: Kadokawa.

A, Inoguchi. (1984). 日本漢文学史. Tokyo: Kadokawa.

D, Wang. (2024). 日本汉字词语料库‘大地语料库’的建设与公开—. *Journal of Japanese Language Study and Research*, 2024(2):125–127.

F, Chen. (2011). 日本汉文学史（上）. Shanghai: Shanghai Foreign Language Education Press.

G, Ball and D. J, Hall. (1965). *Isodata, A Novel Method of Data Analysis and Pattern Classification*. Stanford Research Institute, Menlo Park.

G, Ma. (2003). Chinese Han Poems among Japanese Buddhist Monks in Wushan Period. *Tangdu Journal*. 19(1), 5-10.

G, Rao and Y, Li. (2017). Lexicon Clustering Based Modern Chinese Staging. *Journal of Chinese Information Processing*, 31(6):18–24.

H, Toda. (1957). 日本漢文学通史. Tokyo: Musashinoshoin.

K, Ogata. (1961). 日本漢文学史講義. Tokyo:Hyoronsha.

L, Duan, F, Han, and J, Song. (2012). A Comparative Study on the Automatic Extraction of Two-character Word from Ancient Chinese. *Journal of Chinese Information Processing*. 26(4), 34-42.

L, Lin and X, Wang. (2016). Variability-based Neighbor Clustering: An Integration of Diachronic Linguistics and Corpus Linguistics. *Foreign Language Education and Re search*. 4(1), 6-11.

M, Yoshida. (1932). 岩波講座日本文学：平安朝時代の詩. Tokyo: Iwanami Shoten.

N, Wei. (2002). Corpus-based and Corpus-driven Approaches to the Study of Collocation. *Contemporary Linguistics*. 4(2), 101-114.

P, Rychly. (2008). A Lexicographer-Friendly Association Score. In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, 6-9.

S, Ishikawa. (2008). コロケーションの強度をどう測るか-ダイス係数, tスコア, 相互 情報量を中心として-. 言語処理学会第14回大会チュートリアル資料,40-50.

S, Lloyd. (1957). Least Squares Quantization in PCM. Bell Telephone Labs Memoran dum, Murray Hill, NJ. Reprinted in: *IEEE Transactions on Information Theory*. 28(2), 129-137.

S, Luo and M, Sun. (2003). Chinese Word Extraction Based on the Internal Associative Strength of Character Strings. *Journal of Chinese Information Processing*. 17(3), 9-14.

Stefan Th. Gries and Martin Hilpert. (2012). Variability-based Neighbor Clustering: A Bottom-up Approach to Periodization in Historical Linguistics. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.). *The Oxford Handbook of the History of English*. New York: Oxford University Press. 134-144.

Stefan Th. Gries and Sabine Stoll. (2009). Finding Developmental Groups in Acquisi tion Data: Variability-based Neighbor Clustering. *Journal of Quantitative Linguistics*. 16(3), 217-242.

Y, Fang and Y, Lin. (2022). 传统小说与网络小说的计量风格分析. In W. Huang(eds.). *Quantitative Studies on Vocabulary and Syntax*. Hangzhou: Zhejiang University Press. 94-105.

Y, Haga. (1928). 国語と国民性　日本漢文学史. Tokyo: Fuzambo

9. Language Resource References

D, Wang. (2025). 大地语料库・大地コーパス.中国 国家社科基金项目“日本汉字词语料库建设与研究”（19AYY020）, http://www.xn--cesp9b.net/.