# KOTONOHA : A Corpus Concordance System for Skewer-Searching NINJAL Corpora

*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 7077–7083
Marseille, 11–16 May 2020


# KOTONOHA: A Corpus Concordance System for Skewer-Searching NINJAL Corpora

**Teruaki Oka[†], Yuichi Ishimoto[†], Yutaka Yagi[‡], Takenori Nakamura[†], Masayuki Asahara[†], Kikuo Maekawa[†], Toshinobu Ogiso[†], Hanae Koiso[†], Kumiko Sakoda[†], Nobuko Kibe[†]**
[†]National Institute for Japanese Language and Linguistics, Japan
[‡]Picolab Inc., Japan
kotonoha@ninjal.ac.jp

**Abstract**

The National Institute for Japanese Language and Linguistics, Japan (NINJAL, Japan), has developed several types of corpora. For each corpus NINJAL provided an online search environment, 'Chunagon', which is a morphological-information-annotation-based concordance system made publicly available in 2011. NINJAL has now provided a skewer-search system 'Kotonoha' based on the 'Chunagon' systems. This system enables querying of multiple corpora by certain categories, such as register type and period.

**Keywords:** corpus concordance system, skewer search, speech corpus, written corpus, diachronic corpus

## 1. Introduction

The National Institute for Japanese Language and Linguistics (NINJAL) has developed several corpora in a variety of registers for linguistic research (hereinafter called **NINJAL corpora** or **NINJAL corpus**):

(1) **BCCWJ**: Balanced Corpus of Contemporary Written Japanese
(2) **NWJC**: NINJAL Web Japanese Corpus
(3) **CSJ**: Corpus of Spontaneous Japanese
(4) **CEJC**: Corpus of Everyday Japanese Conversation
(5) **NUCC**: Nagoya University Conversation Corpus
(6) **CWPC**: Gen-Nichi-Ken Corpus of Workplace Conversation
(7) **COJADS**: Corpus of Japanese Dialects
(8) **CHJ**: Corpus of Historical Japanese
(9) **I-JAS**: International Corpus of Japanese as a Second Language

In the corpus design of NINJAL, original data such as speeches, videos, or documents are first textised. Speech or video data are manually transcribed into plain text data. Documents such as newspapers, books, and magazines are also digitised using OCR, and manual modifications are made to correct OCR errors. Second, morphological information is annotated to the plain texts.
The morphological information is given for each lexical unit prescribed by NINJAL, called the **Short Unit Word** (**SUW**, see Figure 1), which is designed for word frequency counts (Den et al., 2008). Since Japanese is not a word-segmented language (does not use white spaces between words), one and the same phrase often may be segmented in several ways. For example, 国立国語研究所 'The National Institute for Japanese Language and Linguistics' has the following segmentation possibilities:

＼国立国語研究所＼

＼国立＼国語研究所＼

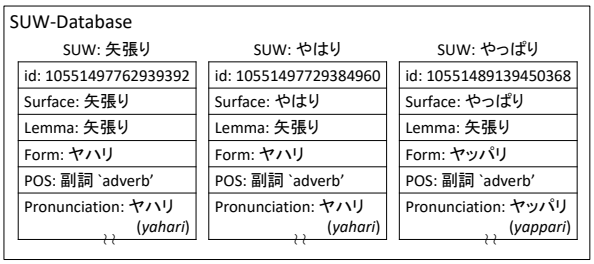


Figure 1: Examples of SUW. This shows the SUW: 矢張り (*yahari*) 'as was expected', its orthographic variant やはり (*yahari*), and their form variant やっぱり (*yappari*). SUWs are managed by a single database, and the actual entries have columns with more meta-information, such as accent information, but the presentation is simplified here.

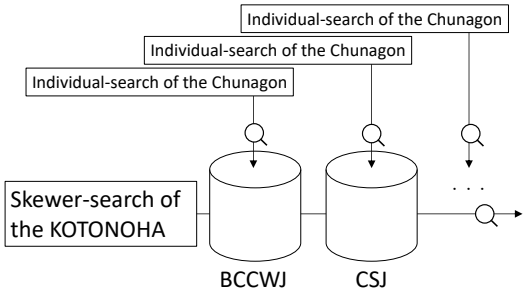


Figure 2: The two ways of searching the NINJAL corpora. Chunagon individually searches each NINJAL corpus, while KOTONOHA searches multiple NINJAL corpora simultaneously.

＼国立＼国語＼研究所＼

＼国立＼国語＼研究＼所＼ (← SUW-segmentation)

Thus, though the word-segmentation policy and meta information on words such as part-of-speech (POS) tags usually differ between corpora, we have regulated these differences

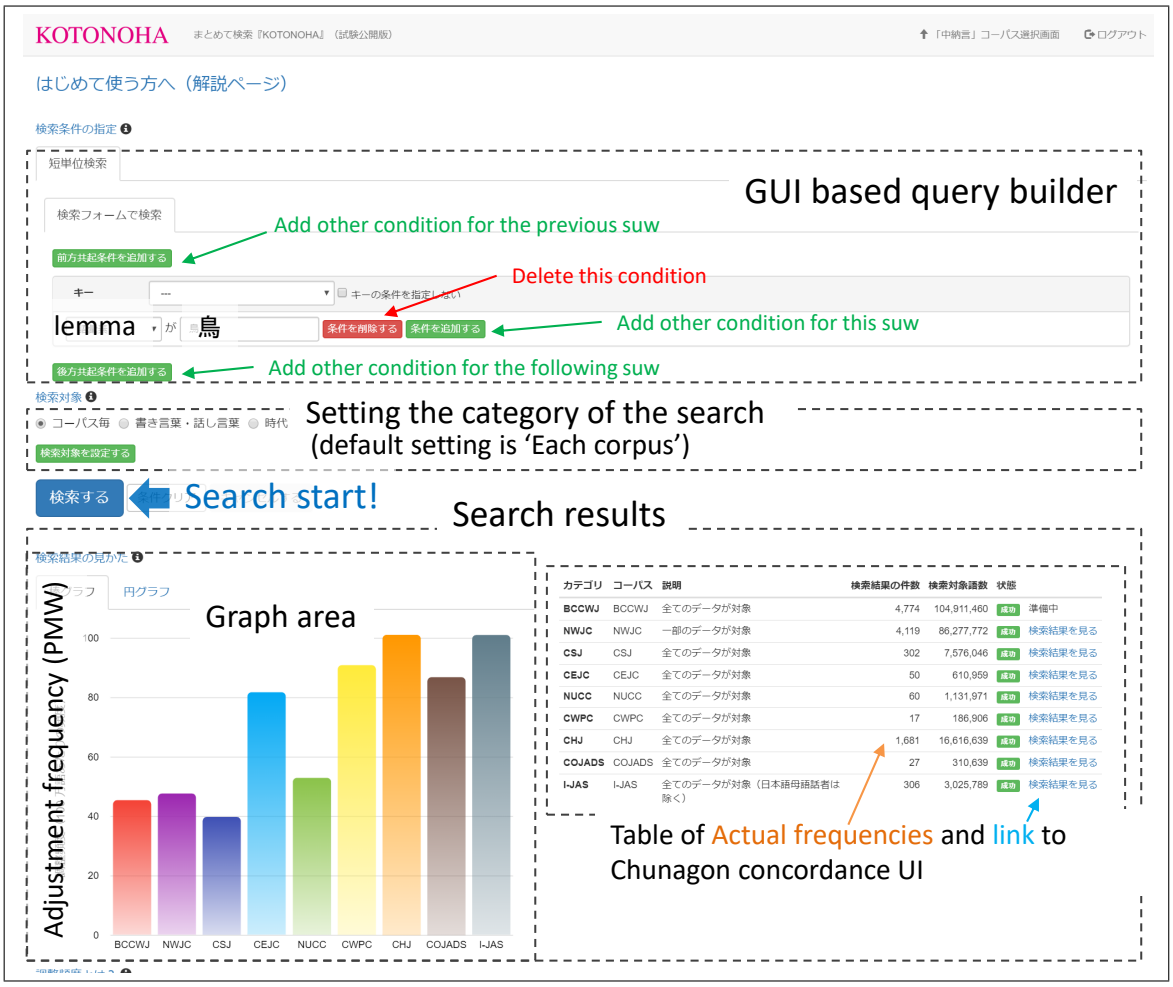


Figure 3: Online skewer-search system 'KOTONOHA' for the NINJAL corpora. Its query builder, similar to the SUW-search of Chunagon, can assign not only a surface string but also POS, pronunciation, and so on. Although in the initial state KOTONOHA has three search categories (grouping corpora), each corpus, register, and period, users are eble to add their own categories freely. The search results are displayed as a graph and a table. This graph is for the skewer-search results view as a bar graph for the query {lemma:鳥 } and 'each corpus' as the category. Although PMWs are used to draw graphs, actual frequencies are also shown in the table.

Table 1: The number of KOTONOHA-searchable SUWs for each corpus by category: Each corpus (as of November 2019).

| Group name in category: Each corpus | Corpus | Detail of target corpus | Number of SUWs |
|---|---|---|---|
| BCCWJ | BCCWJ | ALL | 104,911,460 |
| NWJC | NWJC | some data | 86,277,772 |
| CSJ | CSJ | ALL | 7,576,046 |
| CEJC | CEJC | ALL (Pilot Edition 2019) | 610,959 |
| NUCC | NUCC | ALL | 1,131,971 |
| CWPC | CWPC | ALL | 186,906 |
| I-JAS | I-JAS | ALL (Native speakers of Japanese excluded) | 3,025,789 |

using the SUW as a common word-annotation unit.
NINJAL also provides online corpus concordance systems: '**Chunagon**[1]' (Koiso et al., 2019) and '**KOTONOHA**[2]'. The corpus concordance systems enable query of NINJAL corpora.
The Chunagon has provided a suitable individual-search service for exploring each corpus, as described previously. As of November 2019, 19,000 users had been registered on the service, with an average of 215 active users per day. It has been used not only in Japan but also around the world, in countries such as China, Taiwan, South Korea, Hong Kong, the United States, France, Russia, and the United Kingdom, and has served a wide variety of research purposes, such as spoken language processing, natural language processing, phonetics, psychology, sociology,

[1] https://chunagon.ninjal.ac.jp/
[2] https://chunagon.ninjal.ac.jp/integrated/

Table 2: The number of KOTONOHA-searchable SUWs for each corpus by category: Register (written vs spoken) (as of November 2019).

| Group name in category: Register | Corpus | Detail of target corpus | Number of SUWs |
|---|---|---|---|
| Written | BCCWJ | ALL | 104,911,460 |
| | NWJC | some data | 86,277,772 |
| Spoken | CSJ | ALL | 7,576,046 |
| | CEJC | ALL (Pilot Edition 2019) | 610,959 |
| | NUCC | ALL | 1,131,971 |
| | CWPC | ALL | 186,906 |
| Written of L2 learners | I-JAS | Written language section: story-writing | 154,583 |
| Spoken of L2 learners | | Spoken language section: story-telling, role-play, interview, and picture-description | 2,605,123 |

Table 3: The number of KOTONOHA-searchable SUWs for each corpus by category: Period (diachronic search) (as of November 2019).

| Group name in category: Period | Corpus | Detail of target corpus | Number of SUWs |
|---|---|---|---|
| Nara | CHJ | sub-corpus: Nara | 98,499 |
| Heian | | sub-corpus: Heian | 856,827 |
| Kamakura | | sub-corpus: Kamakura | 822,905 |
| Muromachi | | sub-corpus: Muromachi | 358,419 |
| Edo | | sub-corpus: Edo | 204,519 |
| Meiji/Taishō | | sub-corpus: Meiji/Taishō | 13,259,330 |
| Present | BCCWJ | ALL | 104,911,460 |

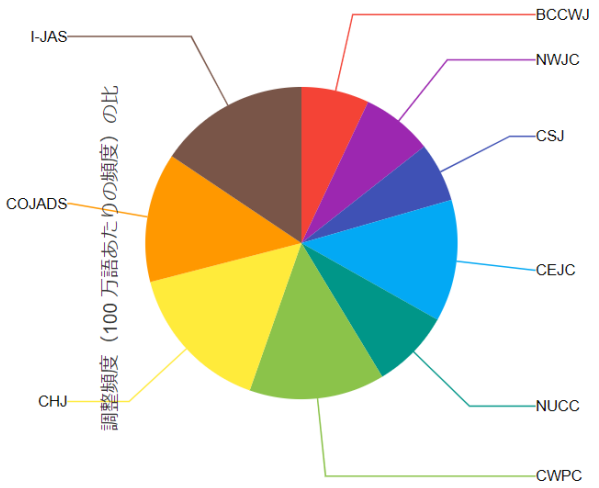


Figure 4: The skewer-search result viewed as a pie chart for the query {lemma:鳥 } and 'each corpus' as the category.

Japanese education, historical research, dialect research, and dictionary compilation.

On the other hand, the KOTONOHA serves as a platform providing a skewer-search (or cross-search) service to explore the NINJAL corpora as a whole (see Figure 2). KOTONOHA mainly shows the results using adjustment frequencies (**Per Million Words** (**PMW**)) and graphs. The skewer-search was implemented through the consistent annotation of SUWs in the NINJAL corpora. In this search, following the common annotation policy, the SUW is used as the common query, whereupon the NINJAL corpora, which include several genres of texts, are skewered in a single query search action, and the search results are aggregated and presented as one result. In the traditional Chunagon search, just one corpus is targeted, but the introduction of KOTONOHA enables new research methods like the following:

1) Using KOTONOHA to survey trends in linguistic phenomena in the NINJAL corpora (≒ Japanese language) from a broader perspective than only searching through the Chunagon.

2) Since the KOTONOHA is only the entrance from which the user may gain an overview, the details of what it returns can be viewed using the conventional individual Chunagon searches.

In this paper, we introduce the KOTONOHA skewer-search system; we then demonstrate its use through a simple case study.

## 2. Overview of the NINJAL Corpora

The Chunagon and KOTONOHA concordance systems enable queries of nine corpora (as of November 2019).

(1) **BCCWJ** (Maekawa et al., 2014) is the first 100-million-word balanced corpus in Japan. It consists of three subcorpora (a publication subcorpus, library subcorpus, and special-purpose subcorpus) and covers a wide range of text registers, including books in general, magazines, newspapers, governmental white papers, best-selling books, Internet bulletin-boards, blogs, school textbooks, minutes of


Figure 5: The CEJC concordance UI linked from the KOTONOHA result for the query {lemma:鳥 } and 'each corpus' as the category.

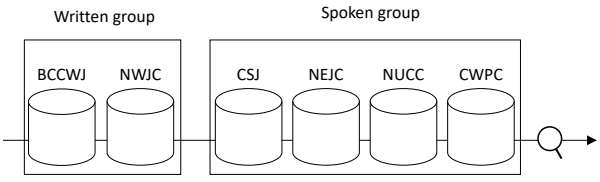


Figure 6: Skewer-search for register. In fact, I-JAS is also skewer-searched like Figure 7, but that is omitted here for simplicity.

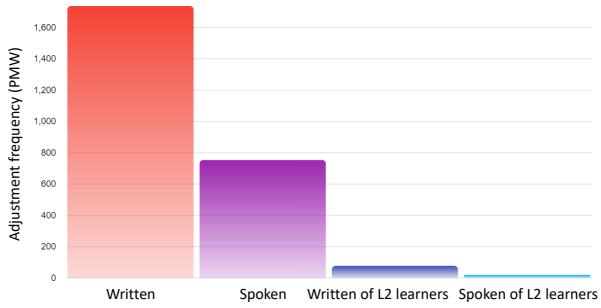


Figure 7: The skewer-search results viewed as a bar graph for である 'be' when we choose 'register' as the category.

the National Diet, publicity newsletters of local governments, laws, and verses.

(2) **NWJC** (Asahara et al., 2014) is a web corpus comprising twenty-five billion words. This corpus was developed without human intervention (normalisation, Japanese morphological analysis, and Japanese dependency analysis were performed automatically on collected web pages).

(3) **CSJ** (Maekawa, 2003) is a large-scale corpus of spontaneous Japanese. It contains speech signals and transcriptions of about 7 million words, along with various prosodic and syntactic annotations.

(4) **CEJC** (Koiso et al., 2018) is a large-scale corpus of everyday Japanese conversation. Prior to the publication of the entire corpus, scheduled for 2022, NINJAL published a part of the CEJC, about 50 hours total, on a trial basis in December 2018.

(5) **NUCC** (Fujimura et al., 2012) is a textised corpus of conversations of native Japanese speakers. It includes 129 conversations (total time about 100 hours).

(6) **CWPC** (Kashino et al., 2018) is a corpus based on the voice data of natural discourse at actual workplaces. The voice data were collected from 40 research cooperators (in their 20s to 50s) working in the Tokyo metropolitan area in Japan.

(7) **COJADS** (Kibe et al., 2018) is the first large-scale dialect corpus in Japan, drawn from the discourse voice data of dialects collected from all over Japan. The original voice data contained approximately 4,000 hours of recordings made at over 200 locations in all 47 prefectures nationwide.

(8) **CHJ** (Kondo, 2012) is a diachronic corpus whose goal in the future is to cover the whole historical range of the Japanese language, including texts in Old Japanese (the oldest attested form of the Japanese language). Some series – he Nara Period Series, Heian Period Series, Kamakura Period Series, Muromachi Period Series, Edo Period Series, and Meiji Era/Taishō Era Series – have been published as part of the projected whole.

(9) **I-JAS** (Sakoda et al., 2016) is a corpus containing cross-sectional research data of Japanese language learners with different mother tongues. It includes data from approximately 1,000 learners with 12 different native languages. It contains oral task data (story-telling, role-play, interview and picture-description) and written task data (story-writing, e-mail writings, and an essay).

Because some corpora include corpus-specific information (e.g. I-JAS includes error and correction information for L2 learners), the Chunagon provides a suitable individual-search system for each NINJAL corpus.

## 3. KOTONOHA: A Corpus Concordance System for Skewer-Searching NINJAL Corpora

As mentioned in Section 1., we developed the KOTONOHA online skewer-search system for the nine NINJAL corpora. In this section, we describe the functions of the KOTONOHA.

Figure 3 shows the UI of the KOTONOHA as displayed on a browser. The KOTONOHA accepts a query consisting of morphological information such as surface string, POS, and

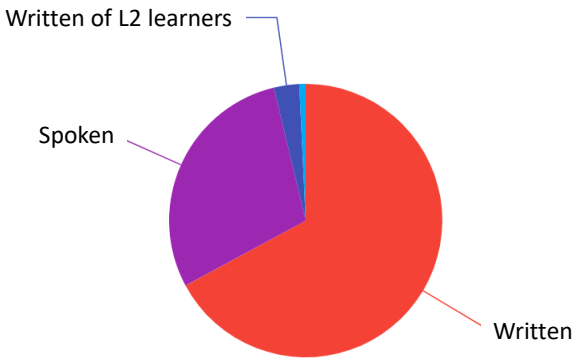


Figure 8: The skewer-search results viewed as a pie chart for である 'be' when we choose 'register' as the category.

Table 4: Actual frequencies of である 'be'.

| Corpus | Actual frequency |
|---|---|
| BCCWJ | 4,774 |
| NWJC | 4,119 |
| **CSJ** | **302** |
| CEJC | 50 |
| NUCC | 60 |
| CWPC | 17 |
| CHJ | 1,681 |
| COJADS | 27 |
| I-JAS | 306 |

conjugation information with the contextual SUWs. Even users without strong computer skills or extensive linguistic knowledge may easily operate the system because the query can be built using a GUI-based query builder similar to Chunagon. For example, Figure 3 searches 鳥 (*tori*) 'bird' by just using the query {lemma:鳥 }. This is a rather rough search query. The search results includes not only the noun 鳥 (*tori*) 'bird' but the suffix -鳥 (read *chou*), like 不死/鳥 (*fusi/chou*) 'phoenix'. Building a more detailed query like {lemma:鳥, POS:noun, pronunciation:トリ (*tori*)} enables the user to exclude the suffix forms above. Also the additional condition: {immediately preceding POS:adjective} narrows down the search results to those containing an adjective modifier, such as 青い/鳥 (*aoi/tori*) 'blue bird' or かわいい/鳥 (*kawaii/tori*) 'pretty bird'.

KOTONOHA can use the **category** function to group corpora (or their sub-corpora). The default setting is 'each corpus': each corpus becomes one group as such (Table 1). KOTONOHA has two other categories in its initial state: **register** (Written vs Spoken) and **periods** (diachronic search). Table 1, Table 2 and Table 3 show the total number of SUWs that the KOTONOHA can query for the following three categories. As can be seen from the tables, the sizes of these corpora are different. Thus, PMW is used to compare them in the KOTONOHA.

The lower part of Figure 3 shows the result view of KOTONOHA when searching on the query consisting of 鳥 'bird' as the lemma (previously described as 'rough query') and 'each corpus' as the category. The bar graphs and pie charts are drawn using the search results converted to the

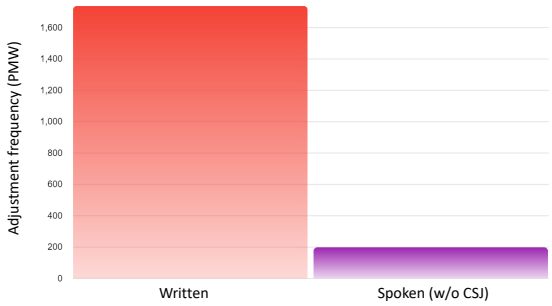


Figure 9: The skewer-search results viewed as a bar graph for である 'be' when we choose 'register without CSJ' as the category.

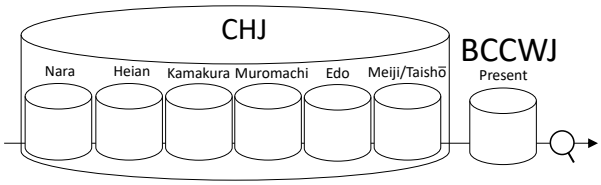


Figure 10: Skewer-search by period.

adjustment frequency (PMW) for each corpus. While the bar graph is shown in Figure 3, Figure 4 shows the pie chart for the same results by switching tabs on the UI.

The search-targeted corpus names, the actual SUW frequencies for each corpus, and the total number of SUWs in each corpus are displayed in table format on the right side of the graph. The table has links to the Chunagon concordance system for each corpus. By following the links, users can jump from the overall results to the Chunagon to see the actual contexts in which the search word appears (see an example in Figure 5). In addition, the users can listen to audio recordings of the contexts, although this function is limited to the following three speech corpora: CSJ, CEJC, and COJADS. We plan to apply the listening function to the search results for I-JAS soon. In this way, KOTONOHA has the functions to compare the NINJAL corpora and to lead the user to a suitable corpus concordance system.

## 4. Case Study: Use of KOTONOHA

In the above section, the search category is 'each corpus', but here we use 'register' as the category of the skewer-search. In this category, NINJAL corpora are categorised into written language, spoken language, written language of L2 learners, and spoken language of L2 learners (see Figure 6). For example, the sentence-final expression である *dearu* which means 'be' is often used in written rather than spoken Japanese.

- (ja) 操作は簡単 (である)
  (en) The operation (is) simple.
- (ja) KOTONOHA の利用は無料 (である)
  (en) The KOTONOHA service (is) free of charge.

We verified this fact with KOTONOHA. The results are shown in Figure 7 and Figure 8.

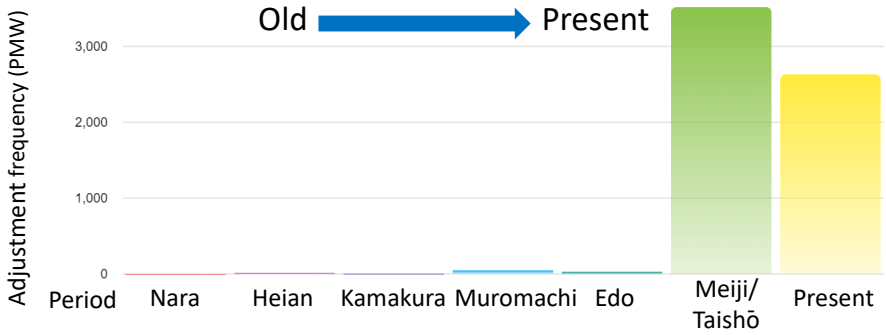


Figure 11: The skewer-search results viewed as a bar graph for the query である 'be' and 'period' as the category.

These graphs show at a glance that である appears more frequently in the written than the spoken language; on the other hand, there are many occurrences of である in the spoken language as well. Thus, when we check the detailed values in the table on the right side of the graph (see Table 4), we find many instances of である in CSJ in the spoken register. In fact, CSJ contains an extensive collection of utterance data from conference lectures, and it is in such places that である is often used. We then checked this using the Chunagon concordance view linked through the KOTONOHA, and as expected the occurrences largely appeared in conference lectures.

On the KOTONOHA system, the category of a search can be customised. We made a new category: register without CSJ and searched again using it. The results are shown in Figure 9. When CSJ was excluded from the spoken language register, it was confirmed that である overwhelmingly appeared in the written language register.

Historically, である has become widely used since the Meiji Era. Therefore, we next changed the category of the search to 'period' (see Figure 10).

The results are shown in Figure 11. As the graph shows, its use has grown rapidly since the Meiji Era.

## 5. Concluding Remarks

In this paper, we introduced our new online corpus concordance system 'KOTONOHA' that skewer-searches NINJAL corpora. KOTONOHA has 'each corpus', 'register' (spoken or written), and 'period' (diachronic search) as the default categories of the skewer-search. In addition to these, users can make a new category for a given purpose. KOTONOHA can already be used freely by accessing (`https://chunagon.ninjal.ac.jp/integrated/`) after registering with the online concordance system Chunagon (`https://chunagon.ninjal.ac.jp/`). We hope that KOTONOHA will generate new insights into the Japanese language that could not be found by studying just one corpus.

## 6. Acknowledgements

This work was supported by JSPS KAKENHI Grants Number 17H00917 and 18H05521, and is a project of the Center for Corpus Development, NINJAL. We would like to thank Editage (www.editage.com) for English language editing.

## 7. Bibliographical References

Asahara, M., Maekawa, K., Imada, M., Kato, S., and Konishi, H. (2014). Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan. *Alexandria*, 26(1–2):129–148.

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008)*, pages 1019–1024.

Fujimura, I., Chiba, S., and Ohso, M. (2012). Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference. Speech and Corpora,*, pages 393–398.

Kashino, W., Omura, M., Nishikawa, K., and Koiso, H. (2018). Supplemental arrangement for public data available in the chunagon versions of 'Gen-Nichi-Ken corpus of workplace conversation'. In *Proceedings of Language Resources Workshop 2018*, pages 494–509.

Kibe, N., Otsuki, T., and Sato, K. (2018). Intonational variations at the end of interrogative sentences in Japanese dialects: From the "Corpus of Japanese Dialects". In *Proceedfings of the LREC 2018 Special Speech Sessions*, pages 21–28.

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., and Usuda, Y. (2018). Construction of the corpus of everyday Japanese conversation: An interim report. In *Proceedings of the 11th edition of the Langauge Resources and Evaluation Conference*, pages 4259–4264.

Koiso, H., Asahara, M., Carlino, S., Nishikawa, K., Aoyama, K., Ishimoto, Y., Wakasa, A., Watanabe, M., Yoshikawa, Y., Kibe, N., and Maekawa, K. (2019). Speech corpora in NINJAL, Japan demonstration of corpus concordance systems: Chunagon and Kotonoha. In *Proceedings of the 3rd International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS 2019) Speech communication: Technology, learning, and pathology*.

Kondo, Y. (2012). the NINJAL diachronic corpus project - Oxford VSARPJ project joint symposium corpus based studies of Japanese language history. In *NINJAL Project Review*, volume 3, pages 84–92.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: its design and evaluation. In *Proceedings of The ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pages 7–12.

Sakoda, K., Konishi, M., Sasaki, A., Suga, W., and Hosoi, Y. (2016). International corpus of Japanese as a second language. In *NINJAL Project Review*, volume 6, pages 93–110.