

Frontiers of Descriptive Language Research Using Japanese Corpora:  
Introduction and Critical Review of the Special Issues  
on Japanese Corpus Studies in the Journal  
*Nihongogaku (Japanese Linguistics)*

Yasunori NISHINA

**Abstract**

In this review article, I critically review the two special issues of the research journal *Nihongogaku*: the December 2016 issue *Data Changing Japanese Language Research and Education* and the summer 2020 issue *Japanese Word History and Modern Lexicology through Corpora* with my original findings.

**1. Introduction**

*Nihongogaku (Japanese Linguistics)*, a journal published by Meiji Shoin, is a unique specialised journal that introduces a wide range of cutting-edge research on the Japanese language. Each issue has a special theme from a wide range of topics, such as contemporary linguistics, historical transition, or Japanese language education, and introduces the latest research results and findings. Published since 1982, the journal recently celebrated its 40th full year; it was a monthly journal until 2019, but became a quarterly from 2020. This review focuses on the December 2016 issue (special feature: *Data Changing Language Research and Education*) and the summer 2020 issue (special feature: *Corpus-based Word History and Modern Lexicology*), because the journal decided to include a special feature on corpora and corpus linguistics in such a long-established Japanese linguistics journal. I will critically review the book with my own research findings.<sup>1</sup>

**2. On data-driven language research and education**

First, the December 2016 issue of *Nihongogaku* (special issue: *Data Changes Japanese Language Research and Education*), which is one of the targets of the reviews in this paper, contains six stimulating studies, mainly by researchers affiliated with the National Institute for Japanese Language and Linguistics (hereafter, NINJAL). Introducing some of them, *The Virtual Lecture: Introduction to Language Resource Studies* (Maekawa, 2016), an introductory and enlightening discussion of what corpora bring to Japanese language studies, explains the importance of statistical language models, machine learning and natural language processing techniques in an easy-to-understand way for readers who are not experts in these areas.

Next, in the UK, the creation of language education syllabuses based on corpus data has attracted attention and been practised for more than 30 years. Finally, it has been discussed in the creation of its Japanese version in *Data-based Syllabus for Japanese Language Education: The Beginner's Grammar Syllabus That Came into View* (Yamauchi, 2016). In particular, in this paper, the importance of “reviewing and creating the syllabus based on objective data rather than teachers’ intuition and experience” (p.38) is successfully presented based on the data of the KY corpus, which is a transcript of the speech of the OPI (Oral Proficiency Interview), a maximum 30-minute conversation test for 90 students whose mother tongue is English, Korean or Chinese.

Furthermore, in *Language Problems Captured by SNS Big Data—Using the Medical Field as an Example* (Aramaki, 2016), based on the viewpoint that “the distance between the two disciplines of medical science and linguistics is getting shorter” (p.29), the author presents a study of the social network that collects human health information from texts on SNS and statistically analyses it. While SNS linguistic data is easy to study, the accuracy of information was also examined in this paper. For instance, there are geographical problems such as a bias towards urban dwellers, user bias regarding the age range of SNS users, and the fact that some senders may be distributing false information.

Among these six papers, I will discuss here the paper entitled *The Science of the Japanese Language as Changed by Corpora: How Japanese Language Research Will Change* (Yamazaki, 2016), which should be particularly interesting to (corpus) linguists, by introducing its concrete contents and incorporating my own original data.

## **2.1. Yamazaki (2016)**

With the advent of computer corpora, Japanese language research is undergoing dramatic changes. Yamazaki (2016) presented an easy-to-understand graph of the number of references with *coopasu* (コ ー パ ス, *corpus*) in the title from 1980 to 2014 in five-year increments, showing a gradual increase in the number of references. In particular, he points out that the total number of references for the five-year period from 2010 to 2014 amounted to nearly 250. Yamazaki (2016) conducted a survey using the Bibliographic Database of Research on Japanese Language and Japanese Language Education at the NINJAL, so the author also used the same database to investigate the situation year by year since 2015 (retrieved on 15th August 2022). The results are as follows:

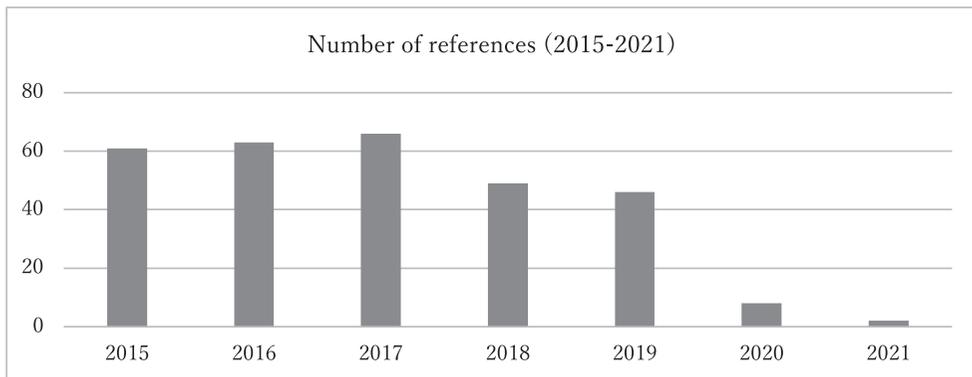


Figure 1. Trends in the number of references with *coopasu* (コーパス, *corpus*) in the title (in one-year increments since 2015)

If we observe the data in five-year increments as in Yamazaki (2016), the number of articles containing *coopasu* (コーパス, *corpus*) from 2015 to 2019 amounts to 285 in total (retrieved 8th September 2022), indicating that the number of articles has been steadily growing. Figure 1 shows that the number remained almost unchanged from 2015 to 2017, slightly decreased from 2018, and appears to be declining sharply from 2020 onwards. For data from 2020 onwards, the NINJAL database may have missed the inclusion of the actual data because it was searched soon after the literature was published. Therefore, the actual number is expected to be much higher. In any case, looking at the data for the 20 years from 2000 to 2019 (in five-year increments) shown in Figure 2, it is safe to conclude that the number of articles has been steadily increasing. This is due to the fact that, as of 2011, more than 80% of the articles published in the most prestigious international journal of linguistics *Language* were already empirically-based (Sampson, 2013), and the ‘quantitative turn’ in cognitive linguistics (Janda, 2013) is a recent trend in language research in general.

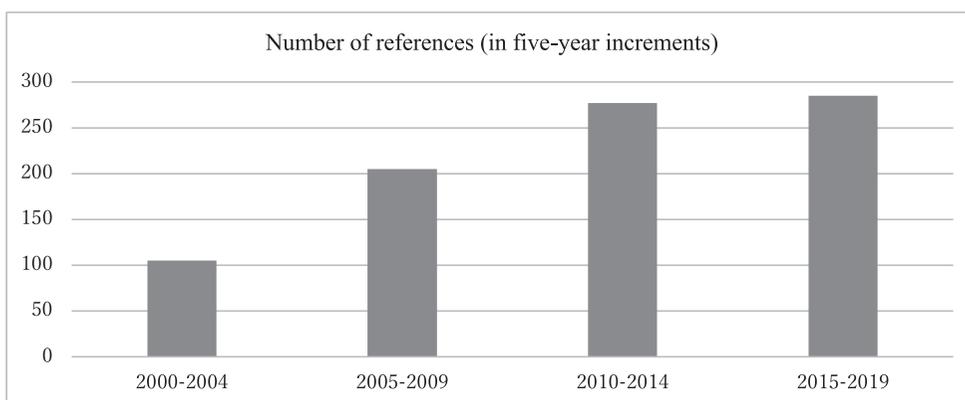


Figure 2. Number of references with *coopasu* (コーパス, *corpus*) in the title (in 5-year increments)

Yamazaki (2016) introduces four aspects of research that change with the use of corpora, one of which is the ease of trial research. For example, he points out that by using the analysis method of lexical profiling, which is also employed in the parallel corpus search tool developed by Nishina & Akasegawa (2021, 2022a, 2022b), linguistic facts about words that have not been noticed even by native speakers can be revealed. As an example, he presents a list of collocations of ‘*Noun* (名詞) + *no* (の, *of*) + *hassei* (発生, *occurrence*)’ extracted from the NINJAL-LWP for BCCWJ (hereafter, NLB, <https://nlb.ninjal.ac.jp/>)<sup>2</sup> in order of frequency as follows:

[Search results for ‘*Noun* (名詞) + *no* (の) + *hassei* (発生)’ in the NLB (partially modified from the table in Yamazaki, 2016, p.15) \*Numbers in brackets indicate frequency.]

*jiko* (事故, *accident*) (189), *saigai* (災害, *disaster*) (84), *jiken* (事件, *incident*) (71), *higai* (被害, *damage*) (59), *jishin* (地震, *earthquake*) (57), *kasai* (火災, *fire*) (54), *kougai* (公害, *pollution*) (53), *shougai* (障害, *disability*) (45), *hanzai* (犯罪, *crime*) (42), *gan* (癌, *cancer*) (33), *songai* (損害, *damage*) (27), *jitai* (事態, *occurrence*) (22), *zassou* (雑草, *weeds*) (21), *bouryoku* (暴力, *violence*) (20)

An overview of the specific nouns that occur shows that the majority of them are “nouns with a negative image” (p.15). However, Yamazaki also points out that few Japanese dictionaries reflect this linguistic fact in their descriptions. In Nishina (2023), it was also pointed out that many of the everyday meanings of *maneku* (招く, *invite*) are used in negative semantic prosody, such as the non-lexical meaning ‘bad things happen’. The linguistic facts about the behaviour of Japanese words obtained from computer corpora have not yet been reflected in dictionary descriptions in Japanese dictionaries and Japanese-English dictionaries. For example, when such Japanese data are

reflected in the descriptions of Japanese-English dictionaries, the descriptive information such as the English translations and example sentences will also be refined. Therefore, as long as Japanese is used in dictionaries and study reference books, it is extremely important to analyse and reflect objective Japanese language data in the descriptions.

The corpus contributes to the discovery of linguistic facts that could not be extracted by human intuition. Yamazaki points out that “by using (corpus) tools and programming yourself, you can easily collect data that used to take a long time by manual work, which makes it possible to do research on a trial basis and gain new findings from them” (p.15). In the field of corpus linguistics, English corpus studies are far ahead of Japanese corpus studies. Therefore, it is likely that the methodological, analytical and other findings gained from English corpus studies will be applied to Japanese lexicological studies in the future.

### 3. An example of corpus-based contemporary lexicological studies of the Japanese language

This section critically reviews the summer 2020 issue of *Nihongogaku* (special issue on *Corpus-Based Word History and Modern Lexicology*). The book contains 25 short papers on Japanese corpus studies in line with the theme above. These include 15 on word history research using historical corpora of Japanese, nine on lexicological research using modern corpora of Japanese, and one commentary paper on lexicological information portals.

In the word history studies, all the papers utilise the Corpus of Historical Japanese (commonly known as CHJ, <https://clrd.ninjal.ac.jp/chj/>) released by the NINJAL. The studies are also based on the frequency and/or the ratio of occurrence of the words/synonyms under analysis. In modern lexicological research, the previously mentioned BCCWJ, CSJ (Corpus of Spoken Japanese, <https://clrd.ninjal.ac.jp/csj/index.html>), CEJC (Corpus of Everyday Japanese Conversation, <https://www2.ninjal.ac.jp/conversation/cejc.html>) and the Tsukuba Web Corpus (<https://tsukubawebcorpus.jp/>), are used, and all these studies are also based on the frequency/ratio of occurrences of the words/synonyms under analysis. In general, the methodology is based on a qualitative investigation of each example after obtaining a rough overall picture of the target linguistic items of investigation based on the frequency/ratio information.

In the following subsections, from among these 25 studies, we critically review Inoue’s (2020) discussion of the occurrence verbs *okiru* (起きる), *okoru* (起こる), *shoujiru* (生じる) and *shouzuru* (生ずる), Maruyama’s (2020) investigation of the adjective *amai* (甘い) and Ogura’s (2020) study of the prefixes of negation *hi-* (非), *fu-* (不), *mi-* (未), and *mu-* (無), which are of particular interest with regard to modern lexicology. With regard to the study of word history, the study by Kim (2020) on the incorporation of the foreign word *kureemu* (クレーム) into basic words will be also reviewed.

### 3.1. Inoue (2020)

Inoue (2020) analyses the usage of the occurrence verbs *okiru* (起きる), *okoru* (起こる), *shoujiru* (生じる) and *shouzuru* (生ずる) from multiple perspectives using data from the BCCWJ, CSJ and Mainichi Shimbun. First, an analysis of the frequency ratio (%) of the use of the genitive complement of *okiru* (起きる) and *okoru* (起こる) in the BCCWJ showed that *koto* (こと) and *nani(ka)* (何(か)), which occur in the genitive complement position, co-occurred more frequently with *okoru* (起こる) (16.9%) than *okiru* (起きる) (13.5%). On the other hand, the result also showed that for *mondai* (問題), *jiken* (事件), *genshou* (現象) and *jishin* (地震), *okiru* (起きる) was predominant over *okoru* (起こる) (see Table 1 below). As a consequence, it is pointed out that in this day and age, “*mondai* (問題) and the like are becoming *okiru* (起きる) more than *okoru* (起こる)” (p.89).

Table 1. Frequency of use of genitive complements for *okiru* (起きる) and *okoru* (起こる)  
(adapted from Inoue, 2020, p.88)

|   | Genitive Complements (主格補語) | <i>okiru</i> (起きる) | <i>okoru</i> (起こる) |
|---|-----------------------------|--------------------|--------------------|
| 1 | <i>koto</i> (こと)            | 117 (13.5%)        | 324 (16.9%)        |
| 2 | <i>nani(ka)</i> (何(か))      | 48 (6.9%)          | 220 (11.5%)        |
| 3 | <i>mondai</i> (問題)          | 55 (6.4%)          | 51 (2.7%)          |
| 4 | <i>jiken</i> (事件)           | 46 (5.3%)          | 66 (3.4%)          |
| 5 | <i>genshou</i> (現象)         | 31 (3.6%)          | 44 (2.3%)          |
| 6 | <i>jishin</i> (地震)          | 27 (3.1%)          | 30 (1.6%)          |
|   | Total                       | 865                | 1919               |

Comparing these two occurrence verbs (発生動詞), it would seem that *okiru* (起きる) has relatively more negative semantic prosody than *okoru* (起こる). On the other hand, in order to examine the overall trend, it is necessary to scrutinise all the genitive complements that occur and to calculate the proportion of nouns with negative polarity among them. The top six in Table 1 would explain only 37.46% of the total for *okiru* (起きる) and only 38.30% for *okoru* (起こる).

Inoue (2020) also uses the BCCWJ to conduct a register analysis of *shoujiru* (生じる) and *shouzuru* (生ずる). By dividing the registers to be searched into white papers (白書) / legal documents (法律) and wisdom bags (知恵袋) / blogs (ブログ), a comparison of written vs. spoken language is made (N.B. Wisdom bags are online question and answer platforms that are popular in Japan). The results are shown in Table 2. An overview of the frequency ratios (%) of occurrences of *shoujiru* (生じる) and *shouzuru* (生ずる) shows that *shouzuru* (生ずる) is slightly more dominant in white papers (白書) and legal documents (法律) (namely, written language), while *shoujiru* (生じる) is overwhelmingly more frequent in wisdom bags (知恵袋) and blogs (ブログ)

(namely, spoken language). It should be noted that the sample size of white papers (白書) / legal documents (法律) only amounts to about 30 per cent of that of wisdom bags (知恵袋) / blogs (ブログ). In order to avoid data distortion as much as possible, one approach would be to cut 70% of the wisdom bags (知恵袋) and blogs (ブログ) data by random sampling, adjust the sample numbers so that they are directly comparable to each other, and recalculate the frequency proportions.

Table 2. Comparative study of the frequency of use of *shoujiru* (生じる) and *shouzuru* (生ずる) (adapted from Inoue, 2020, p.90)

|                       | white papers (白書),<br>legal documents (法律) | wisdom bags (知恵袋),<br>blogs (ブログ) | Word Differentiation |
|-----------------------|--|-----------------------------------|----------------------|
| <i>shoujiru</i> (生じる) | 191 (39.9%)                                | 137 (90.1%)                       | 328 (52.0%)          |
| <i>shouzuru</i> (生ずる) | 288 (60.1%)                                | 15 (9.9%)                         | 303 (48.0%)          |
| Total                 | 479 (75.9%)                                | 152 (24.1%)                       | 631 (100%)           |

Inoue (2020) further investigated the stylistic phase of the three occurrence verbs (*okiru* (起きる), *okoru* (起こる) and *shoujiru* (生じる)) by register (white papers (白書), newspapers (新聞), books (書籍), diet proceedings (国会議事録) and blogs (ブログ)) using the BCCWJ. The results showed that in white papers (白書), the use of *shoujiru* (生じる) was more than 80%, whereas in books (書籍), diet proceedings (国会議事録) and blogs (ブログ), the use of *okoru* (起こる) was in the 50% range, which was more predominant than others. Therefore, it is concluded that “*shoujiru* (生じる) is written, whereas *okoru* (起こる) is spoken” (p.91), pointing out that the degree of spoken language becomes stronger in the order of *shoujiru* (生じる) < *okiru* (起きる) < *okoru* (起こる).

In order to determine whether these verbs are spoken or written, the results of the search and analysis would be more robust if at least the legal documents (法律) and wisdom bags (知恵袋) registers used in Table 2 were also included in the analysis, data from spoken corpora such as CSJ were also presented, and, as already mentioned, the number of samples for each register were also adjusted.

### 3.2. Maruyama (2020)

Maruyama (2020) investigated the actual usage of the modern Japanese adjective *amai* (甘い, *sweet*) from the perspective of a corpus-based lexicography. The corpus used in this study was the BCCWJ. After an overview of the number of occurrences per million words in each of the seven registers, the subsequent nouns were examined in detail, as shown below (the numbers in brackets in the examples below are frequencies). Regarding the results of the register analysis in the

BCCWJ, the high number of occurrences in blogs (ブログ), publications (出版) and magazines (雑誌) indicated that this adjective “tends to appear in texts with a relatively soft style” (p.95). As for subsequent nouns, the high-ranking occurrences of *mono* (物, *object*), *kaori* (香り, *aroma*) and *koe* (声, *voice*) at the top of the list point to the spread of their use from the sense of taste to other senses, such as smell and hearing.

[Top 10 subsequent nouns for *amai* (甘い) in the BCCWJ (partly modified from Maruyama, 2020, p.95)]

*mono* (物, *object*) (441), *kaori* (香り, *aroma*) (135), *koe* (声, *voice*) (62), *nioi* (匂い, *smell*) (54), *shiru* (汁, *juice*) (41), *kotoba* (言葉, *word*) (39), *kangae* (考え, *thought*) (36), *koto* (事, *thing*) (23), *masuku* (マスク, *mask*) (21), *aji* (味, *taste*) (18)

Using the examples of nouns following the attributive form of *amai* (甘い) and classifying them according to the ten word senses listed in *the Meikyo Kokugo Dictionary* (『明鏡国語辞典』), the most common use was in the main sense of taste (65% in blogs and 40% in magazines), followed by pleasant sensations other than taste and smell. By register, *amai* (甘い) was followed by nouns such as *kotoba* (言葉, *word*), *koe* (声, *voice*) and *masuku* (マスク, *mask*) in blogs, and *masuku* (マスク), *pinku* (ピンク, *pink*) and *deiteiuru* (デイトール, *detail*) in magazines. In particular, examples of the use of *amai* (甘い) in fashion magazines were introduced as characteristic examples corresponding to pleasant sensations other than taste and smell (e.g. *amai kyami* (甘いキヤミ, *sweet camisole*), *amai toppusu* (甘いトップス, *sweet tops*)).<sup>3</sup>

As a discussion, this study was in line with the basics of descriptive linguistics. With regard to the extraction of the subsequent nouns of *amai* (甘い), a different aspect of the behaviour of this adjective might be revealed by analysing the nouns with the highest scores using statistical indices. Using the NLB, as in Yamazaki (2016), a list by LD score and MI score can be obtained. For reference, the list is given below (retrieved 29 September, 2022).

[Top 10 subsequent nouns for *amai* (甘い) based on MI and LD scores from the NLB]

**MI score** *yakudoku* (薬毒, *medicinal poison*) (14.79), *nioyaka* (匂やか, *smelling*) (14.20), *pomaado* (ポマード, *pomade*) (12.54), *lemonsui* (レモン水, *lemon water*) (12.20), *henkyuu* (返球, *returned ball*) (11.88), *sasayaki* (囁き, *whisper*) (11.06), *sushimeshi* (すし飯, *sushi rice*) (10.79), *sufule* (スフレ, *soufflé*) (10.70), *kaori* (香り, *aroma*) (10.65), *fukuiku* (馥郁, *fragrance*) (10.62)

**LD score** *kaori* (香り, *aroma*) (9.33), *shiru* (汁, *soup*) (8.44), *masuku* (マスク, *mask*) (7.90), *nioi* (匂い, *smell*) (7.83), *okashi* (お菓子, *sweets*) (7.68), *yuuwaku* (誘惑, *temptation*) (7.41), *sasayaki* (囁き, *whisper*) (7.09), *mitsu* (蜜, *honey*) (7.01), *toiki* (吐息, *sigh*) (6.85), *tamagoyaki* (卵焼き,

*fried eggs*) (6.68)

### 3.3. Ogura (2020)

Ogura (2020) investigated the actual use of the prefixes of negation *hi-* (非), *fu-* (不), *mi-* (未) and *mu-* (無) using the BCCWJ. In all cases, in terms of the type ratio of words to be combined, among Japanese, Chinese, foreign and mixed words, the proportion of Chinese words was the highest, with *fu-* (不), *mi-* (未) and *mu-* (無) accounting for around 95%, whereas only *hi-* (非) was in the 80% range. In particular, many of the mixed words combined with *hi-* (非) were found to contain foreign words as constituents, such as *alucoulu inryou* (アルコール飲料, *alcoholic beverage*) and *suteloidosei* (ステロイド性, *steroidal*), and 100 (78.1%) of the 128 mixed word types to be combined with *hi-* (非) were found to have foreign words as constituents. In addition, *hi-* (非) also has the second highest proportion with foreign words, after *mi-* (未), and has the highest number with foreign words (e.g. *hi-akutibu* (非アクティブ, *non-active*), *hi-howaitokaraa* (非ホワイトカラー, *non-white-collar*), etc.). Ogura pointed out that this is a characteristic of *hi-* (非).

The number of corpus studies investigating Japanese prefixes is rather limited, and Ogura's (2020) findings are a linguistic fact that could not have been discovered by our linguistic intuition alone as native speakers of Japanese. On the other hand, the distribution by register is also of interest. As in the studies by Inoue (2020) and Maruyama (2020), by conducting a survey in the order of registers in the BCCWJ with the highest degree of written language (e.g. white papers (白書) → newspapers (新聞) → books (書籍) → diet proceedings (国会議事録) → blogs (ブログ)), or by using the CSJ, a more detailed picture of language use from the perspective of written vs. spoken language would become more evident.

### 3.4. Kim (2020)

Kim (2020) investigated the foreign word *kureemu* (クレーム, *complaint*), which seems to have failed to become a basic word, using an expanded and revised version of a diachronic newspaper corpus of approximately 22 million characters (for more information on this corpus, see Kim (2011)). Figure 3 shows that *kureemu* (クレーム) gained momentum against the synonyms *kujyou* (苦情, *grievance*) and *monku* (文句, *complaint*) between 1970 and 1991, but that its momentum then waned between 2000 and 2010, and that in the newspaper genre, its incorporation into a basic word seems to have failed. Kim (2020) suggests that the reason for this failure is that “the form of the word *kureemu* (クレーム) has taken on the same negative emotional meaning as *monku* (文句), which is ultimately mediated by the form *~wotsukeru* (〜をつける, *to put ~ on*)” (p.75).

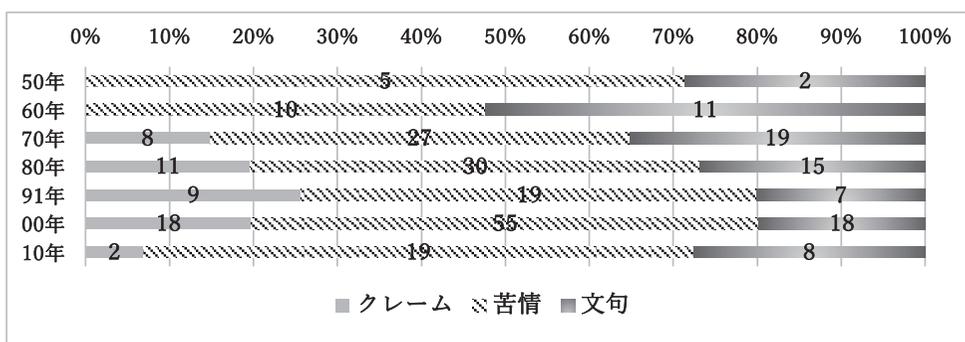


Figure 3. Percentages of use of *kureemu* (クレーム), *kujyou* (苦情) and *monku* (文句) (modified from the figure in Kim, 2020, p.73)

Kim (2020) analyses *kureemu* (クレーム) in terms of frequency comparison between independent usage (自立用法) and combined usage (結合用法), comparison of meaning (complaints about economic activity vs. complaints about matters not related to the economy), forms such as noun phrases, verb phrases and *tyuusihou* (中止法, suspended form), and type of sender (individual vs. non-individual such as organisation or group). The final result is that it has failed to become a basic word, but from the author's point of view, it is highly likely that foreign words such as *kureemu* (クレーム) are still commonly used in the colloquial register in the business world. If this is the case, then it is highly likely that it has not failed to become a basic word, but rather that it has inevitably become segregated from its synonyms and usage in the language economy of modern society. Further research on this topic is expected in the future.

#### 4. Conclusion

##### 4.1. Concluding remarks

The corpus-based studies of modern Japanese linguistics introduced in this review are on the same paths (namely, research findings and methodologies) that English corpus studies have taken in the past. In this sense, it can be said that the Japanese corpus is a treasure trove of linguistic facts that have yet to be discovered, and that at present only the tip of the iceberg is being uncovered. Further active interactions between English and Japanese corpus studies are expected in the future. For a quick overview of recent developments in Japanese corpus linguistics, these two special issues are highly recommended. I also recommend reading Lee, Ishikawa & Sunakawa (2018), Maekawa (2013) and Tanomura (2014).

##### 4.2. Future potential of Japanese corpus studies

Finally, I would like to explore a little about the future potential of the Japanese language

corpus. Table 4 introduces the 11 Japanese corpora that can be searched with the software (Chunagon and KOTONOHA) currently available from NINJAL. These corpora include a corpus of Japanese language learners. For example, C-JAS collected data from three native speakers of Korean and Chinese, respectively, over a three-year period, containing 47 60-minute dialogues (free conversations). Such a corpus could be used in future longitudinal surveys of Japanese language education. These corpora should be used to research Japanese language and develop teaching materials and syllabuses for learners of Japanese now and in the near future.

Table 3. List of Japanese Corpora (See details at <https://clrd.ninjal.ac.jp/corpus-list.html>)

| Corpus Name   | Acronym |
|---|---------|
| Balanced Corpus of Contemporary Written Japanese      | BCCWJ   |
| NINJAL Web Japanese Corpus                            | NWJC    |
| Corpus of Spontaneous Japanese                        | CSJ     |
| Corpus of Everyday Japanese Conversation              | CEJC    |
| Showa Speech Corpus                                   | SSC     |
| Nagoya University Conversation Corpus                 | NUCC    |
| Gen-Nichi-Ken Corpus of Workplace Conversation        | CWPC    |
| Corpus of Historical Japanese                         | CHJ     |
| Corpus of Japanese Dialects                           | COJADS  |
| Corpus of Japanese as a Second Language               | C-JAS   |
| International Corpus of Japanese as a Second Language | I-JAS   |

Also, the Sketch Engine (<https://www.sketchengine.eu/>) can search the Japanese Web 2011 (jaTenTen11), a corpus of 10 billion Japanese words collected from the internet. The Word Sketch search function in the Sketch Engine is particularly outstanding. This function employs the Lexical Profiling technology already explained, and instantly lists examples corresponding to specific patterns and collocations of a particular word extracted from the corpus. Figure 4 shows part of a screen shot of a Word Sketch search for the expression *liajyuu* (リアージュ), a recent favourite expression of Generation Z Japanese.

| particle       | noun/noun                   | suffix           | 1verb                     | 0pronom                   | prefix       |
|----------------|-----------------------------|------------------|---------------------------|---------------------------|--------------|
| じゃん<br>リア充 じゃん | 爆発<br>リア充 爆発 する             | 奴<br>リア充 奴       | 憧れる<br>リア充 に 憧れて          | 集窟<br>リア充 の 集窟            | 非<br>非 リア充   |
| ばかり<br>リア充 ばかり | リア充<br>リア充 爆発 リア充 爆発 リア充 爆発 | 振る<br>リア充 ぶって    | 離れる<br>見える<br>リア充 に 見える   | 溜まり場<br>振り<br>リア充 の ぶり    | 業<br>クソ リア充  |
| って<br>リア充 って   | いけメン<br>リア充 イケメン            | 臭<br>リア充 臭 が     | 囲む<br>リア充 に 囲まれた          | 定義<br>リア充 の 定義 が          | 超<br>超 リア充   |
| なんて<br>リア充 なんて | カップル<br>リア充 カップル            | 擬き<br>リア充 もどき    | 勝つ<br>リア充 に 勝てる           | 集まり<br>リア充 の 集まり          | 今<br>今 リア充 な |
| なんか<br>リア充 なんか | アピール<br>リア充 アピール            | ぼい<br>リア充 っぽい    | 負ける<br>リア充 に 負け           | いけメン<br>仲間<br>リア充 の 仲間 入り |              |
| しか<br>リア充 しかない | 満腹<br>リア充 満腹                | 寄り<br>リア充 寄りの    | 成る<br>リア充 に なっ            | 自慢<br>不幸                  |              |
| か<br>リア充 か     | オーラ<br>リア充 オーラ              | 共<br>リア充 ども      | 近づく<br>分かる<br>リア充 に は わから | 祭典<br>群れ<br>リア充 の 群れ に    |              |
|                | アビ<br>リア充 アビ する             | だらけ<br>リア充 だらけ の |                           | 乗り                        |              |
|                | 桃源<br>非 リア充 桃源 郷            | 振り<br>の リア充 っぷり  |                           |                           |              |
|                | 先<br>館                      | 向け<br>リア充 向け     |                           |                           |              |
|                | 自慢<br>リア充 自慢                | 向き<br>リア充 向きの    |                           |                           |              |
|                |                             | 臭い               |                           |                           |              |

Figure 4. Word sketch of *liajyuu* (リア充) from jaTenTen11

As can be seen from some of the examples on Word Sketch, expressions that include the word *liajyuu* (リア充) basically express envy, longing or jealousy towards *liajyuu* (リア充) people from the perspective of non-*liajyuu* (非リア充) people, and it is rare for *liajyuu* (リア充) people themselves to use this expression. Therefore, the expression *liajyuu* (リア充) can be said to be imbued with negative semantic prosody (Louw, 2000; Hunston, 2002)<sup>4</sup>. Exploring and teaching the discourse features of such vocabulary and phrases is necessary for the production of natural Japanese for non-native speakers of Japanese, and such information should be reflected in teaching materials and dictionaries for future Japanese language education.

#### Acknowledgements

This research was funded by JSPS KAKENHI 20K00692 and 23K00599, and benefited from a long-term overseas research fellowship from Kobe Gakuin University. The author would like to thank Prof. Bjarke Frellesvig and Ms. Kaori Nishizawa of the University of Oxford, UK, and all those involved. This paper is also part of an invited lecture given at the seminar ‘Frontiers of Japanese Corpus Linguistics’ organised by The British Association for Teaching Japanese as a Foreign Language (BATJ) on 10th June 2023.

#### Notes

- 1 Please note that the quotations given in this paper are the author’s English translations of the original texts.
- 2 LWP stands for Lago Word Profiler (Nishina, 2023) and BCCWJ for Balanced Corpus of Contemporary Written Japanese (<https://clrd.ninjal.ac.jp/bccwj/>).
- 3 Maruyama (2020, p.97) points out that *the Digital Daijisen* (『デジタル大辞泉』) describes the adjective *amai* (甘い, *sweet*) as used in fashion as “fashion that emphasises girlish cuteness with ribbons, frills,

pastel colours and the like”.

- 4 Louw (2000) pointed out that semantic prosody is “a form of meaning which is established through the proximity of a consistent series of collocates” (p.57). Hunston (2002) also notes that semantic prosody belongs to the whole phrase, not to individual words, cannot be ascertained without a large number of examples from a corpus, and is usually negative, rarely positive in its evaluation.

## References

- Aramaki, E. (2016). “Language issues viewed through SNS big data: The medical field as an example”. *Nihongogaku*, 35(13), 28-37. [Translation by the author] (荒牧英治 (2016) 「SNS ビッグデータでとらえる言語問題－医療分野を例に－」『日本語学』第35巻第13号：28-37.)
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Inoue, T. (2019). “Analysis of Linguistic Registers of Occurrence Verbs: *Okiru, Okoru, Shojiru, Shozuru*” *The Bulletin of University of Kochi Faculty of Cultural Studies*, 68, 37-50.
- Inoue, T. (2020). “Special issue: Corpus-based word history and modern lexicology *Okiru, Okoru (Shojiru, Shozuru)*”. *Nihongogaku*, 39(2), 88-92. [Translation by the author] (井上次夫 (2020) 「特集 コーパスによる語史と現代語誌 起きる・起こる (生じる・生ずる)」『日本語学』第39巻第2号：88-92.)
- Janda, L. A. (2013). “Quantitative Methods in Cognitive Linguistics: An Introduction.” In Janda, L. A. (ed.), *Cognitive Linguistics—The Quantitative Turn: Essential Reader*. Berlin: Mouton de Gruyter, pp.1-32.
- Kim, E. (2011). “Shift of the loanwords to basic words in the Japanese newspaper vocabulary in the second half of the 20th century”. *Handai Nihongo Kenkyu*, 3, 1-175.
- Kim, E. (2020). “Special issue: Corpus-based word history and modern lexicology *Kujyou and Monku*”. *Nihongogaku*, 39(2), 72-76. [Translation by the author] (金愛蘭 (2020) 「特集 コーパスによる語史と現代語誌 クレーム (苦情・文句)」『日本語学』第39巻第2号：72-76.)
- Lee, J., Ishikawa, S., & Sunakawa, Y. (2018). *New Introduction to Corpus Research for Japanese Language Teaching*. [Translation by the author] (『新・日本語教育のためのコーパス調査入門』). Tokyo: Kuroasio Publishers.
- Louw, B. (2000). “Contextual prosodic theory: Bringing semantic prosodies to life.” In Heffer, C., Sauntson, H., & Fox, G. (Eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham, England: ELR, pp.48-94.
- Maekawa, K. (ed.) (2013). *Corpus and Japanese Language Education*. [Translation by the author] (『コーパスと日本語教育』) Tokyo: Asakura Publishing Company.
- Maekawa, K. (2016). “Virtual lecture: Introduction to language resources”. *Nihongogaku*, 35(13), 2-11. [Translation by the author] (前川喜久雄 (2016) 「仮想講義「言語資源学入門」」『日本語学』第35巻第13号：2-11.)
- Maruyama, T. (2020). “Special issue: Corpus-based word history and modern lexicology *Amai*”. *Nihongogaku*, 39(2), 94-97. [Translation by the author] (丸山岳彦 (2020) 「特集 コーパスによる語史と現代語誌 甘い」『日本語学』第39巻第2号：94-97.)
- Nishina, Y., & Akasegawa, S. (2021). “Toward the development of Parallel Link (tentative name) (Ver. 1.0), an online search tool for Japanese-English and English-Japanese parallel corpora (interim report)”. *Proceedings of the JAECS Conference 2021*, 25-30.
- Nishina, Y., & Akasegawa, S. (2022a). “Toward the development of Parallel Link (ver.1.0), an online analysis tool for Japanese-English and English-Japanese parallel corpora: Review of parallel corpora, analysis tools

- and researches followed by re-compilation of existing parallel corpora”. *English Corpus Studies*, 29, 63-78.
- Nishina, Y., & Akasegawa, S. (2022b). “Parallel Link (Ver. 1.20), a tool for searching Japanese-English and English-Japanese parallel corpora: About the interface, search functions, and research applications”. *Proceedings of the JAECs Conference 2022*, 7-12.
- Nishina, Y. (2023). *Aspects of Parallel Corpus Linguistics: From Monolingual Corpus Studies to Bilingual Corpus Studies (Society of Global Communication Studies of Kobe Gakuin University Research Series Vol.2)*. Tokyo: Kaitakusha.
- Ogura, H. (2020). “Special issue: Corpus-based word history and modern lexicology *Hi- (Fu-, Mi-, Mu-)*”. *Nihongogaku*, 39(2), 98-101. [Translation by the author] (小椋秀樹 (2020) 「特集 コーパスによる語史と現代語誌 非-(不-・未-・無-)」『日本語学』第39巻第2号: 98-101.)
- Sampson, G. (2013) “The empirical trend: Ten years on.” *International Journal of Corpus Linguistics* 18(2), 281-289.
- Tanomura, T. (ed.) (2014). *Corpus and Japanese Linguistics*. [Translation by the author] (『コーパスと日本語学』) Tokyo: Asakura Publishing Company.
- Yamauchi, H. (2016). “Data-based Japanese language teaching syllabus: The emerging picture of the elementary grammar syllabus”. *Nihongogaku*, 35(13), 38-47. [Translation by the author] (山内博之 (2016) 「データに基づく日本語教育シラバスー見えてきた初級文法シラバス像ー」『日本語学』第35巻第13号: 38-47.)
- Yamazaki, M. (2016). “The science of the Japanese language as changed by corpora: How Japanese language research is changing” *Nihongogaku*, 35(13), 12-17. [Translation by the author] (山崎誠 (2016) 「コーパスが変える日本語の科学ー日本語研究はどのように変わるかー」『日本語学』第35巻第13号: 12-17.)