Putting *Ranuki* into Perspective: A Multi-Corpus Analysis of Linguistic Variation and Change

Shin-ichiro Sano
Keio University

1 Introduction

The recent development of corpora as a resource for linguistic research has been producing a growing body of corpus-based studies. Cross-linguistically, this trend sheds light on unnoticed aspects of language, and accordingly led to new findings. Moreover, it is further along in development: beyond the written vs. spoken dichotomy, a variety of corpora that differ, for example, in time periods, register or style are currently available. This enables cross-corpus comparisons of linguistic features, such as comparing the distribution of a variable phenomenon in different corpora with different style and different recording period (Sano 2019a, b, to appear).

This study focuses on *ranuki* that is one of the well-known and widely-studied variable phenomena in modern Japanese, and analyzes its properties by employing multiple corpora with different features. In particular, among major sociolinguistic categories/labels that may affect the distribution of *ranuki*, this study examines the effects of time period, register, and style. Furthermore, based on the chronological trend and a gender difference in the distribution of *ranuki* observed in one of the spoken corpora, this study predicts the properties of *ranuki* as a linguistic change.

The generalizations drawn from the patterns observed in the analysis are summarized as follows: a) time period: *ranuki* is more likely to be observed in recent speech than in old speech, reflecting the progress of linguistic change; b) type of speech: *ranuki* is more compatible with conversations than with monologs; c) formality: *ranuki* is more likely to be observed in less formal speech than in formal speech; d) spoken/written distinction: *ranuki* is more compatible with spoken language than with written language. The results also suggest e) the resistance of formal style to linguistic change, f) the status of *ranuki* as a linguistic change (change from below), g) the property of written language in social media that is different from written language in a traditional sense. Moreover, the results offer some implications about h) *ranuki*'s status as a linguistic change at both the community and the individual level (age-grading), i) the possible interplay between language acquisition (natural acquisition vs. explicit learning) and the distribution/diffusion of *ranuki*. Additionally, the gender-biased pattern in the use of *ranuki* (more preferred by female speakers than by male speakers) confirms the general role of women in the development of linguistic change.

The remainder of this article is organized as follows: Section 2 sketches background information about *ranuki* and research on *ranuki*, and clarifies the research agenda. Section 3 introduces the characteristics of corpora employed in this study, and the method with which the data was collected and analyzed. Section 4 presents the results of the analysis and the discussion. Section 5 concludes the discussion with some future perspectives.

2 Background

2.1 *The variable ranuki* *Ranuki* (also known as *ra*-Deletion) is a morphophonological variation in potential forms. It is an ongoing change observed since the 1920s (Kindaichi et al. 1995) as part of the global change in potential forms in Japanese over the past 400 years (e.g., Sano 2012, 2015).

Potential forms in Japanese comprise a verb stem and a potential suffix. The verb stems are classified into two types according to their ending, being either vowel-final (e.g., *mi-* 'see', *tabe-* 'eat') or consonant-final (e.g., *ik-* 'go', *nom-* 'drink'). Potential forms show morphophonological alternations according to the stem type, as vowel-final stem verbs take the potential suffix *-rare*, while consonant-final stem verbs take *-e*. *Ranuki* affects

[*] I would like to thank Satoshi Nambu, Hanae Koiso, and Takehiko Maruyama, for their helpful comments and suggestions on this corpus-based research project. This project is partly supported by KAKENHI grant #19K00558, #16H03426, and NINJAL collaborative research project "A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation." The author is solely responsible for any faults and errors that may arise in this article.

only verbs with vowel-final stems, where it variably deletes the syllable *ra* from the potential suffix *-rare*, resulting in the reduced form *-re*. This produces a morphophonological variation in potential forms with vowel-final stems comprising *-rare* (full form) and *-re* (reduced form) (Ito and Mester 2004). As exemplified in Table 1, the potential forms of vowel-final stems such as *mi-* and *tabe-* are traditionally full forms, such as *mi-rare* and *tabe-rare*; if these forms undergo *ranuki*, novel reduced forms such as *mi-re* and *tabe-re* are produced.

Table 1. Examples of full and reduced potential forms of vowel-final stem verbs

stem (vowel-final)	full form	*ranuki* (reduced) form
mi- 'see'	mi-rare 'can see'	mi-re 'can see'
tabe- 'eat'	tabe-rare'can eat'	tabe-re 'can eat'

Although *ranuki* is still diffusing, there is an established attitude towards it, as (i) it has been more than 100 years since its appearance, and *ranuki* is now widely used, and (ii) *ranuki* has been covered in the media and in education by highlighting the non-standard status of the reduced form, so the distinction between the full form (standard) and the reduced form (non-standard) is recognized even by non-experts.

2.2 *Research on ranuki* *Ranuki* has long been subject to research from a variety of perspectives, including traditional grammar, sociolinguistics, and phonology, as its properties provide significant clues to the solution of linguistic issues, to a better characterization of Japanese grammar, and to the demonstration of linguistic variation and change in progress. Thus, the research focusing on *ranuki* has made a significant contribution to the development of both empirical and theoretical linguistics (Inoue 1998; Inoue and Yarimizu 2002; Ito and Mester 2004; Kanda 1964; Kinsui 2003; Matsuda 1993, 2008; Nakamura 1953; Okazaki 1980; Sano 2011 et seq.; Shibuya 1993,).

In the variationist sociolinguistic paradigm, for example, *ranuki* has been studied with a focus placed on the link between the dichotomy of full (standard) vs. reduced (vernacular) forms and pre-determined socio-demographic categories. Additionally, the characteristics of *ranuki* that the prior work revealed mainly concerned linguistic or internal factors, such as length/frequency of the preceding verb stem, phonotactic restrictions (identity avoidance), sentence type, and embeddedness (Matsuda 1993; Sano 2011, 2013, 2019b). Recent usage-based studies have shed light on previously unnoticed extra-linguistic aspects of *ranuki*, such as the fact that the reduced form is more preferred by younger and female speakers, speakers with a lower level of education, and in informal style (Sano 2011). Furthermore, beyond this category-based approach, the recent multi-dimensional approach has demonstrated that the use of *ranuki* patterned by formality reflects gender and workplace stereotypes in Japan (Sherwood 2014, 2016). Additionally, it has shown that in performing social actions based on the community's ideology, speakers derive *ranuki*'s new indexicalized functions (non-referential meanings) for their interactional purposes (e.g., interpersonal relationships and interactional atmosphere) associated with indirect indexicality, in addition to the status imposed on the form by linguistic norms/prescriptive grammar (Sano 2018).

Building upon the findings of prior work, this paper aims i) to provide an additional case study of *ranuki* via multiple-corpus comparison with a main focus on time period, register, and style; and ii) to predict the nature of *ranuki* as a linguistic change by comparing the distribution observed in multiple corpora with different characteristics.

3 Method

3.1 *Corpora* This study employed five corpora that differ in their characteristics. This section presents a brief sketch of the characteristics of each corpus.

1. *Balanced Corpus of Comtemporary Written Japanese (BCCWJ)*

The written data were retrieved from the Balanced Corpus of Contemporary Written Japanese (data version 1.1, hereafter BCCWJ, Maekawa 2008), which is a collection of published Japanese articles (e.g., newspapers, books, magazines, SNS/blogs). The BCCWJ was released in 2011. The range of the target data covers written texts published from 1976 to 2005 (books, that constitute the major part of this corpus, are from 1986 to 2005). The size of the corpus amounts to 100 million words, in which the core component with detailed annotations amounts to 1 million words. This study focused on the core component.

2. *Corpus of Showa Spoken Japanese (CSSJ)*

The *Corpus of Showa Spoken Japanese* (Maruyama 2016. henceforth CSSJ) is a database of speech samples amounting to about 50 hours that were recorded from the 1950s to the 1970s (mainly in the 1950s) during the Showa era (1926-1989). In addition to conversations, the speech samples in the CSSJ also include monologs, such as public speaking. The "monitor-released" version comprising about 17 hours of monologs was published in 2019. This database of audio recordings from about 60 years ago and earlier provides important information about the chronological transition of spoken language. This study employed the "pilot" data that includes all speech samples (about 50 hours) consisting of conversations and monologs.

3. *Corpus of Spontaneous Japanese (CSJ)*

The *Corpus of Spontaneous Japanese* (Maekawa et al. 2000, henceforth CSJ) released in 2004 is a large-scale database of spontaneous speech. The CSJ consists of about 7.52 million words amounting to 651 hours of speech, in which the subset called *Core* (about 45 hours, 50,000 words) comes with more detailed annotation. The majority of the speech samples (about 90 %) are monologs, although other types of speech such as dialog are also included. Monologs consist of two types of speech samples: Academic Presentation Speech is a collection of academic talks like conference presentations. The other type is Simulated Public Speaking that comprises presentations on pre-selected topics given by non-professional speakers. Speech samples in the CSJ were recorded from 1999 to 2003.

4. *Meidai Conversation Corpus (MCC)*

The *Meidai Conversation Corpus* (data version 2016.12, hereafter MCC, Fujimura et al. 2012) is a collection of 120 conversation samples between native speakers of Japanese that amounts to about 100 hours of speech. The MCC was compiled under a project of the Grant-in-Aid for Scientific Research from 2001 to 2003, and was released in 2003. It is now available on *Chuunagon* (Ogiso and Nakamura 2013) administered by the National Institute for Japanese Language and Linguistics (audio-data is not provided). Meta-linguistic information annotated to the MCC includes speaker attributes (e.g., age, gender, birthplace), relationship between speakers, and setting, which are useful for sociolinguistic research and discourse analysis.

5. *Corpus of Everyday Japanese Conversation (CEJC)*

The Corpus of Everyday Japanese Conversation (Koiso 2017 et al., CEJC) is a database of multi-party conversations in daily-life situations that are free from artificial and task-oriented purposes. In addition to transcription files and audio files, the CEJC also comes with video data. The final version of the CEJC will amount to about 200 hours of conversations; however, this study targets the monitor version (only available for now) that comprises 50 hours of speech, and was released in 2018.

Table 2 summarizes the characteristics of each corpus that are relevant to the present analysis.

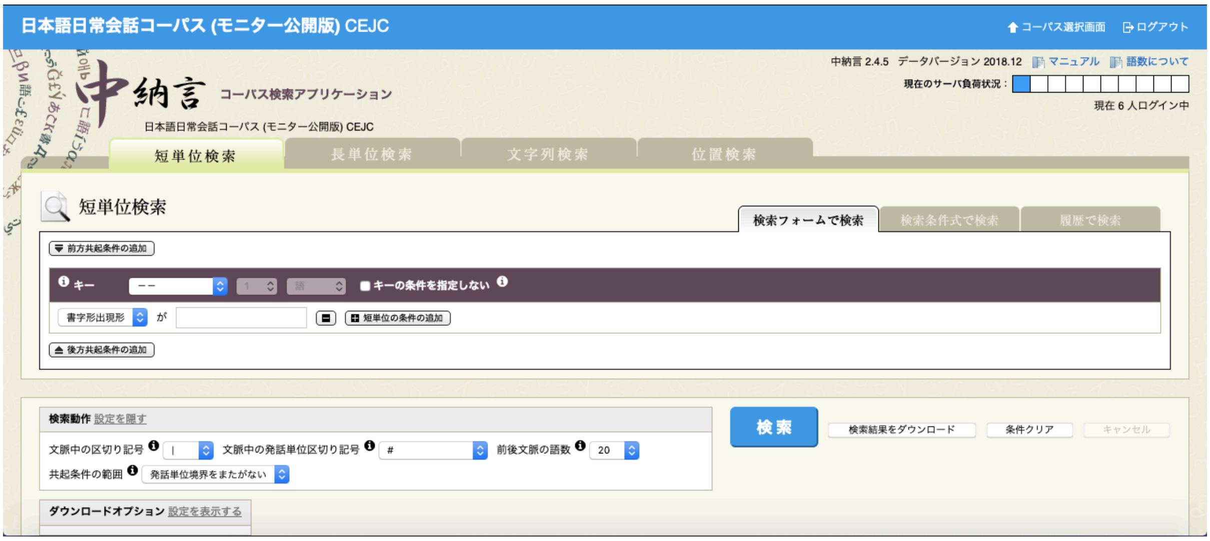
Table 2. Major characteristics of the five corpora

	spoken/written	time period	type of speech	formality
BCCWJ	written	intermediate (1976-2005)	NA	NA
CSSJ	spoken	old (1950s-1970s)	conversation +monolog	intermediate
CSJ	spoken	recent (1999-2002)	monolog (+dialog and reading)	formal
MCC	spoken	recent (2001-2003)	conversation	informal
CEJC	spoken	very recent (2016-)	conversation	informal

As Table 2 shows, the first item that classifies the five corpora into two groups is the spoken/written distinction. Only the BCCWJ is a corpus of written language, while the other four corpora record spoken language. Because written language is not produced as a "speech," and the formality of written language cannot be straightforwardly compared with the one of spoken language, the BCCWJ was left out from the analysis regarding type of speech and formality. The five corpora are ordered according to the time period during which speech samples were recorded, where the BCCWJ and the CSSJ cover a wide range (about 30 years): the CSSJ is old, the CSJ and the MCC are recent, and the BCCWJ is in between these two groups. The CEJC is the most recent among these corpora. The four corpora of spoken language are different in terms of the type of speech they record: The CSJ consists mostly of monologs, and is stylistically formal; on the other hand, the MCC and the CEJC consist of conversations, and are stylistically informal. The CSSJ consists of both conversation and monologs, and thus it is intermediate between formal and informal styles.

In this study, I assume that the BCCWJ represents the written language of modern Japanese in general. I also regard the CSSJ as representing old spoken Japanese, the CSJ as recent formal spoken Japanese, the MCC as recent informal spoken Japanese, and the CEJC very recent informal spoken Japanese. Note that the CEJC is the most recent corpus that records utterances or interactions produced under relaxed and natural style/settings. This is one of the most important features that the database has for the observation of vernacular/non-standard forms like *ranuki*.

3.2 *Data collection procedure* There are several different interfaces to access the information in each corpus: The data in the BCCWJ, the CSJ, the MCC, and the CEJC are accessible via the online search engine *Chuunagon* (Figure 1). Besides that, the data in the BCCWJ, the CSJ, and the CEJC are also distributed via storage media (DVD, USB memory stick, or portable HDD depending on the corpus and its versions). This study collected the data in the BCCWJ, the MCC, and the CEJC using *Chuunagon* (version 2.2.2.2 for the BCCWJ and the MCC, and version 2.4.5 for the CEJC). Instead of searching in *Chuunagon* anew, the data in the CSJ is based on the report in Sano (2011). The details of the data collection procedure are presented therein. The corpora implemented into *Chuunagon* provide the information in the following format. Transcription follows the Japanese orthography consisting of Chinese characters and the kana syllabary. The transcriptions are morphologically parsed and tagged using the parser *MeCab* and the dictionary *UniDic*.


日本語日常会話コーパス (モニター公開版) CEJC

中納言 2.4.5 データバージョン 2018.12

現在のサーバ負荷状況：

現在 6 人ログイン中

短単位検索

キー ーー

書字形出現形 が

検索動作

文脈中の区切り記号 |

文脈中の発話単位区切り記号 #

前後文脈の語数 20

共起条件の範囲 発話単位境界をまたがない

Figure 1. Example of *Chuunagon* search (CEJC)

As the CSSJ is not implemented on *Chuunagon*, I collected the data in the CSSJ using *Himawari* (version 1.6) (Yamaguchi 2014) that is an offline full-text search system, where the morphological information was analyzed and annotated by the parser *MeCab* (version 0.996). In either case, the search results were exported and downloaded into Microsoft® Excel for Mac (version 16.9) and processed there.

This study focused on every sample in the CSSJ, the CSJ, the MCC, and the CEJC and in the core component of the BCCWJ. In particular, I targeted utterances involving the full form (*rare*) or the reduced form (*re*). On both *Chuunagon* and *Himawari*, the targeted utterances are searchable by selecting from the items listed in the pull-down menu and specifying the targeted letters in the required fields. In *Chuunagon*, to search for the full form, the lexeme was specified as られる (*rareru*); to search for the reduced form, the type of inflection was specified as 下一段-ラ行 (*ra*-line lower unigrade). While in *Himawari*, to search for the full form, “key” was specified as れ (*re*), and the lexeme was specified as れる|られる (*reru* or *rareru* in regular expression); to search for the reduced form, “key” was specified as れ (*re*), the type of inflection was specified as 一段で始まる (unigrade and subcategories of it), and a subclassification of the part of speech was specified as 自立 (free morpheme).

From the retrieved data, I filtered out certain irrelevant information. For the full form, the specification れる/られる (*reru/rareru*) covers four meanings: potential, passive, honorific, and spontaneous. Thus, I excluded tokens with meanings other than potential. Likewise, for the reduced form, the specification 下一段-ラ行 (*ra*-line lower unigrade) or 一段 (unigrade) in combination with 自立 (free morpheme) can still false-hit other verbal categories (e.g., *re*-final stem verbs, potential forms of consonant-final stem verbs); thus, I excluded categories other than the reduced form. Additionally, tokens were excluded from the dataset if the targeted segments were a part of filled pauses or word fragments, while some tokens were included if the targeted segments contained non-standard or non-prescriptive pronunciation.

3.3 *Dataset and quantitative method* The exhaustive search for the data in the five corpora and filtering resulted in 10,365 tokens, of which 9,328 were full forms and 1,037 were reduced forms. The frequency distribution of the full form and the reduced form is summarized in Table 3.

Table 3. The distribution of full form and reduced form in the five corpora.

Full/reduced form	BCCWJ	CSSJ	CSJ	MCC	CEJC
full form	787	341	7,615	387	198
reduced form	65	14	543	269	146

The variable consists of the full form and the reduced form. The primary measure in the quantitative analysis is the ratio of *ranuki* (reduced form) in all potential forms, which is based on token frequency count. In the analysis, the ratio of *ranuki* was calculated for each context and compared. The distributional skews were tested with a linear mixed-effects model (Jaeger 2008; Barr 2013; Barr et al. 2013) using the glmer function in R (R development Core Team 1993–2019). The dependent variable was the choice of full form or reduced form. The predictor was the corpus, genre (for the analysis of the BCCWJ), and speakers' age and gender (for the analysis of the CEJC). Random effects (grouping variables) were speakers/writers and items. (The CSSJ does not come with the speaker ID. I, therefore, referred to the title ID of speech samples instead. Additionally, for the BCCWJ, I referred to the publishers' information, since the writers' information was not presented in many tokens.) For the post-hoc test, multiple comparisons were run with Steel-Dwass's method.

4 Analysis

4.1 *Cross-corpus comparison* This section presents the cross-corpus comparison of the distribution of *ranuki*, in which the properties of *ranuki* were estimated by linking the distribution to the characteristics of each corpus regarding time period, register, and style. Figure 2 shows the ratio of *ranuki* (reduced form) in potential forms observed in the five corpora.

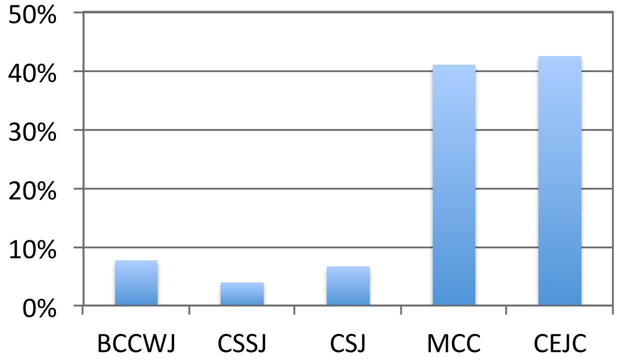


Figure 2. The ratio of *ranuki* (reduced form) in the five corpora

As Figure 2 shows, the ratio of *ranuki* significantly differs depending on the kind of corpus ($z = 9.957$, $p < 0.01$). In particular, a major difference was observed between two groups: the ratio of *ranuki* was higher in the MCC and the CEJC than in the BCCWJ, the CSSJ, and the CSJ (BCCWJ–MCC: $t = 15.47$, $p < 0.01$; BCCWJ–CEJC: $t = 14.29$, $p < 0.01$; CSSJ–MCC: $t = 12.52$, $p < 0.01$; CSSJ–CEJC: $t = 12.1$, $p < 0.01$; CSJ–MCC: $t = 29.27$, $p < 0.01$; CSJ–CEJC: $t = 23.82$, $p < 0.01$). No significant difference was observed within each group (BCCWJ–CSSJ: $t = 2.36$, $p = 0.13$; BCCWJ–CSJ: $t = 1.08$, $p = 0.82$; CSSJ–CSJ: $t = 2.02$, $t = 2.36$, $p = 0.25$; MCC–CEJC: $t = 0.44$, $p = 0.99$). As we saw in Section 3.1, the five corpora differ in their characteristics. The next step is to consider the distribution in terms of the characteristics of each corpus. Table 4 summarizes the distribution of *ranuki* and the major characteristics of the five corpora (based on Table 2).

Table 4. The distribution of *ranuki* and characteristics of the five corpora.

	ratio of *ranuki*	time period	type of speech	formality	spoken/written
BCCWJ	0.076 (65/852)	intermediate	NA	NA	written
CSSJ	0.039 (14/355)	old	conversation	intermediate	spoken
CSJ	0.066 (543/8,158)	recent	monolog	formal	spoken
MCC	0.410 (269/656)	recent	conversation	informal	spoken
CEJC	0.424 (146/344)	very recent	conversation	informal	spoken

Based on the result that the ratio of *ranuki* is higher in the MCC and the CEJC than in the BCCWJ, the CSSJ, and the CSJ, we can argue the following points with respect to the sociolinguistic aspects of *ranuki*. In terms of the time period, *ranuki* is more likely to be observed in older data (BCCWJ: intermediate and CSSJ: old) than in more recent data (MCC: recent and CEJC: very recent). Since the MCC and the CEJC do not differ in terms of other characteristics than time period, the result that there was no significant difference between the MCC (recent) and the CEJC (very recent) suggests that the *ranuki* variable did not show further diffusion in the recent 10 to 15 years. In other words, the absence of difference between the BCCWJ (intermediate), the CSSJ (old), and the CSJ (recent), and the (unexpected) difference between the CSJ (recent) and the MCC (recent) can be attributed to the effect of other characteristics.

Next, we consider the distribution with reference to the type of speech and formality. With respect to the type of speech, *ranuki* is more compatible with conversations (MCC and CEJC) than with monologs (CSJ). Similarly, as for formality, *ranuki* is more likely to be observed in informal speech (MCC and CEJC) than in formal speech (CSJ). In both cases, the CSSJ patterned differently from other corpora with same/similar specifications. Even though the CSSJ, the MCC, and the CEJC are grouped together in terms of type of speech, only the CSSJ patterned with the CSJ (monolog), in that these two corpora showed a lower ratio of *ranuki*. Additionally, the CSSJ is intermediate in formality, however it patterned with the CSJ (formal) that showed a lower ratio of *ranuki*. Thus, the absence of statistical difference in the ratio of *ranuki* between the CSSJ (conversation, intermediate) and the CSJ (monolog, formal) that are different in terms of type of speech and formality suggests that this is due to the time period.

Old conversation-style and less formal speeches on the one hand, and recent monolog-style and formal speeches on the other, share similar characteristics as a ground of linguistic variation and change. Linguistic change proceeds gradually, producing a continuum where innovative/in-coming forms are more diffused in more recent speech than in older speech. Given that the ratio of *ranuki* in recent speeches (CSSJ), in which the change is supposed to be more advanced, was as low as in old speeches (CSJ), we can argue that monolog-style formal speech is more resistant to the change, blocking the diffusion of the innovative forms. In other words, because monolog-style formal speech resists to the change, even recent speech in this context show a lower ratio of innovative forms (Sano 2019a). Furthermore, the result that the ratio of *ranuki* was higher in informal speech (MCC and CEJC) than in more formal speech (CSJ), and thus that the change is more advanced in informal speech suggests that *ranuki* is an example of change from below (Labov 1966, 1990, 1994) where the change is assumed to proceed from informal style to formal style.

The last item is the spoken/written distinction. The difference between spoken language and written language in the context of linguistic variation and change is that, due to its nature, written language is less likely to be subject to variation and change, and the change in written language, if any, is slower than in spoken language. This has been demonstrated in a variety of work (e.g., corpus based study in Japanese: Sano 2019 a, b, to appear). In the present data, however, written language (BCCWJ) was shown to include *ranuki* to the same extent that it was observed in spoken language (CSSJ and CSJ). This can be reduced to the following fact. The BCCWJ consists of several register categories: in addition to typical written language such as newspapers, books, magazines, the BCCWJ includes SNS/blogs called *Yahoo!blog* and *Yahoo!chiebukuro* (Q & A site, the Japanese counterpart of *Yahoo Answers*). As Figure 3 shows, the register difference produced the following gap in the ratio of *ranuki*: in the BCCWJ (t = 8.71, p < 0.01) , *ranuki* was predominant in two registers: *Yahoo!blog* (0.422), *Yahoo!chiebukuro* (0.367); while in other registers, the likelihood of *ranuki* was 0.007 (*Yahoo!blog*–*Yahoo!chiebukuro*: t = 0.68, p = 0.78; *Yahoo!blog*–others: t = 16.35, p < 0.01; *Yahoo!chiebukuro*–others: t = 14.45, p < 0.01).

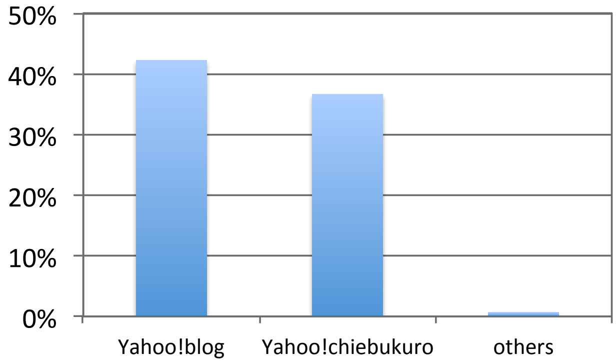


Figure 3. The ratio of *ranuki* (reduced form) in the BCCWJ

Considering the fact that the communication in social media, such as blogs, tends to have properties that are similar to spoken language even though the output is written texts, the skewed distribution by register within the BCCWJ gives further support for the resistance of written language against linguistic change. To reflect this into our model, I excluded both the full forms and the reduced forms observed in *Yahoo!blog* and *Yahoo!chiebukuro* from the dataset, and recalculated the distribution. This process resulted in the revised distribution shown in Figure 4.

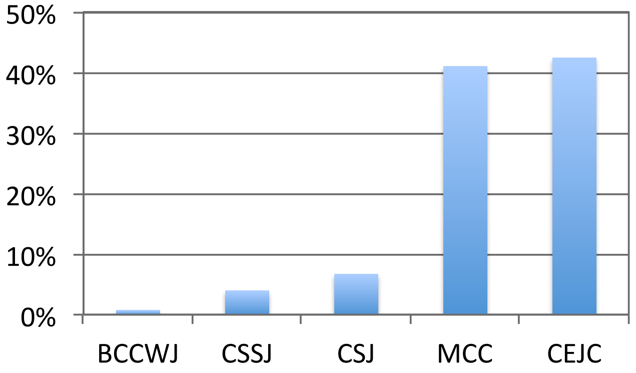


Figure 4. The ratio of *ranuki* (reduced form) in the five corpora (*Yahoo!blog* and *Yahoo!chiebukuro* are excluded from the BCCWJ data.)

In the revised distribution, there is still a significant difference due to the type of corpus (z = 9.795, p < 0.01). The major difference between the distributions in Figure 2 and Figure 4 is that in the latter the BCCWJ is different from both the CSSJ and the CSJ (BCCWJ–CSSJ: t = 3.73, p < 0.02; BCCWJ–CSJ: t = 6.27, p < 0.01), showing the lowest *ranuki* ratio. This supports the assumption that written language is resistant to linguistic variation and change. Except for this, the difference between two groups remains intact: the ratio of *ranuki* was higher in the MCC and the CEJC than in the BCCWJ, the CSSJ, and the CSJ (BCCWJ–MCC: t = 18.48, p < 0.01; BCCWJ–CEJC: t = 18.03, p < 0.01; CSSJ–MCC: t = 12.52, p < 0.01; CSSJ–CEJC: t = 12.1, p < 0.01; CSJ–MCC: t = 29.27, p < 0.01; CSJ–CEJC: t = 23.82, p < 0.01). Also, the absence of difference between the CSSJ and the CSJ, and between the MCC and the CEJC was kept constant (CSSJ–CSJ: t = 2.02, t = 2.36, p = 0.25; MCC–CEJC: t = 0.44, p = 0.99). Thus, the resistance to the change in formal speech is maintained. Although how to treat the status of the language in social media is debatable, we can argue at least that it is different from written language in a traditional sense. The generalizations drawn from the patterns of *ranuki* observed in the five corpora are summarized in Table 5.

Table 5. The patterns of *ranuki* observed in the five corpora

factors	patterns
time period	old < recent
type of speech	conversation > monolog
formality	less formal > formal
spoken/written	spoken > written

In terms of time period, *ranuki* is more likely to be observed in recent speech than in old speech, reflecting the progress of linguistic change. Regarding the type of speech, *ranuki* is more compatible with conversations than with monologs. Similarly, as for formality, *ranuki* is more likely to be observed in less formal speech than in formal speech. In terms of the spoken/written distinction, *ranuki* is more compatible with spoken language than with written language. These patterns are all compatible with previous observations in the research on *ranuki* (e.g., Sano 2011 et seq.).

4.2 *Properties of ranuki as a linguistic change* This section presents a prediction of the properties of *ranuki* as a linguistic change, based on the chronological trend, and a gender-biased pattern in the distribution of *ranuki* observed in the CEJC.

Firstly, I will present the analysis based on the chronological trend of the distribution of *ranuki*. Since the CEJC comes with speakers' information including their age at the time of recording, I analyzed the chronological transition of the distribution of *ranuki* by comparing the speech of individuals of different ages following the apparent-time method (Labov 1963, Cukor-Avila and Bailey 2013). Figure 5 plots the ratio of *ranuki* in six age cohorts with a fitted curve in the CEJC.

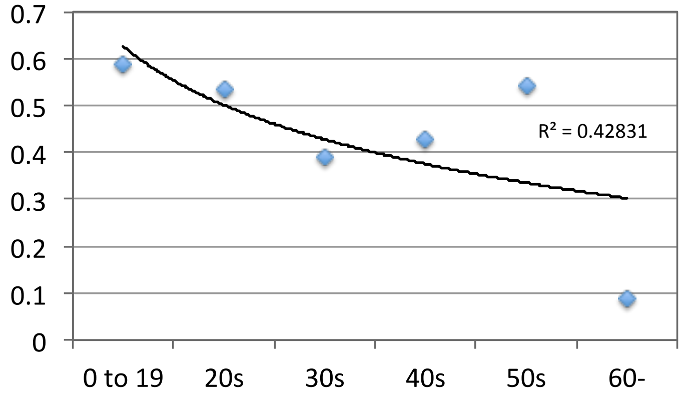


Figure 5. The distribution of *ranuki* by speakers' age in the CEJC

As Figure 5 illustrates, the ratio of *ranuki* significantly differs depending on speakers' age (z = -2.838, p < 0.01). Specifically, the overall trend is that *ranuki* ratio is inversely proportional to speakers' age. This suggests that *ranuki* is more likely to be used by younger speakers than by older speakers. This pattern is compatible with a symptom of linguistic change that innovative forms are more preferred by younger generations. The result supports the previous observation that *ranuki* is an example of linguistic change in progress (e.g., Matsuda 1993; Sano 2011).

In addition to that, there are several points worth considering. Except for the sharp decrease in the age group of 60 years old and upward, we can observe the V-shaped curve where the age groups of 0 to 19 years old and the 50s (heads of the curve) constitute the peaks of the ratio, and the age group of the 30s (center of the curve) shows the lowest ratio. This corresponds to the patterns observed in age grading. Age grading (Labov 1994; Sankoff 2005; Cheshire 2006; Wagner 2012) refers to linguistic change across individual's lifespan (contra linguistic change at the community level). Age grading is identified as a repetition of age-appropriate linguistic behavior in each generation regardless of the stability of the linguistic variable in the community (Sankoff 2005). The primary motivation for this is individuals' increasing awareness of the standard language (Eckert 1997). Linguistic behaviors viewed as age grading can include the following pattern: individuals start their linguistic life with vernacular/non-standard forms. As they get older, start school, and become working adults, individuals gradually increase the ratio of standard variety (decrease the ratio of vernaculars), adjusting their language use to the "age-appropriate" manner. After that, individuals re-increase the ratio of vernaculars due to such factors as change in their role or position. Given that *ranuki* shows an age-graded pattern, we can argue that the *ranuki* variable is socially identified with the specific context/setting in which it is used, such as informal and relaxed situations, and has established its role as a vernacular.

The next point concerns child language acquisition. In the present data, *ranuki* was never observed in the age group between 0 to 4, suggesting that the speakers aged between 0 to 4 (born in the 2010s) have a grammar where potential forms of vowel-final stem verbs are exclusively reduced forms, instead of full forms. The possible scenario is that recent generations naturally acquire the reduced form as a (sole) potential form of vowel-final stem verbs; after starting school, they acquire full forms through explicit learning. This sharply contrasts with the following observation.

Ranuki was never observed among speakers in their 80s and above (born between 1932 and 1938). If these speakers experienced parenting, they should not have used *ranuki* during this process. It follows that their children never heard *ranuki* as a linguistic input in their acquisition period, and thus they did not acquire this variant naturally; while they explicitly learned it afterward. This may have been reflected in the distribution of *ranuki*: there is a significant gap in the ratio of *ranuki* in the age groups before 50s and 60s: the speakers aged 60 years old and upward are less likely to use *ranuki*; while the speakers aged 59 and below use *ranuki* more frequently. Further investigation is required considering the possible causes such as social change, however this possibility is worth pursuing.

Furthermore, as we observed only the age group of 60 years old and upward did not follow the pattern of age-grading. This can be attributed to the process with which *ranuki* is acquired (naturally acquired vs. explicitly learned): for the same reason as above, *ranuki* did not diffuse to this age cohort to the extent that the speakers in their 60s (those who were born in 1958 and before) incorporated it into their repertoire, and thus it is impossible for these speakers to consciously arrange their linguistic behavior using *ranuki*, which should have been reflected in the age-graded pattern.

The last item to be examined is the gender gap in the distribution of *ranuki*. Figure 6 shows the distribution of *ranuki* by gender in the CEJC.

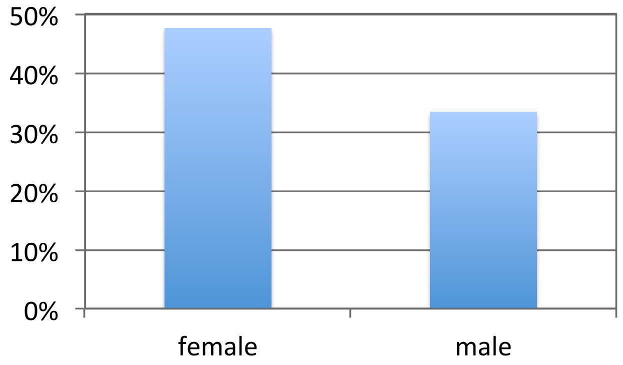


Figure 6. The distribution of *ranuki* by gender in the CEJC

The ratio of *ranuki* significantly differs depending on gender ($z = 2.155$, $p < 0.05$): female speakers are more likely to use *ranuki* than male speakers. The gender-biased pattern observed in the use of *ranuki* is consistent with the general tendency in linguistic variation and change formulated by Labov (1990):

Principle I: In stable sociolinguistic stratification, men use a higher frequency of nonstandard forms than women.
Principle Ia: In change from above, women favour the incoming prestige forms more than men.
Principle II: In change from below, women are most often the innovators.

Among three principles regarding gender and linguistic variation and change, Principles Ia and II are most relevant to linguistic change. In either cases, change from above or change from below, standard or non-standard, innovative forms are more likely in women's speech; in other words, women lead linguistic changes. This study suggests that in the case of *ranuki* too, women take the initiative in its development.

5 Summary

This study demonstrated a multi-corpus analysis of *ranuki*, and confirmed the aspects of *ranuki* (either already-known or novel) and some implications that the properties of *ranuki* offer. These findings are summarized as follows:

a. Time period: *ranuki* is more likely to be observed in recent speech than in old speech, reflecting the progress of linguistic change.
b. Type of speech: *ranuki* is more compatible with conversations than with monologs.
c. Formality: *ranuki* is more likely to be observed in less formal speech than in formal speech.
d. Spoken/written distinction: *ranuki* is more compatible with spoken language than with written language.
e. Formality: Formal style is more resistant to linguistic change than informal style.

Implications:

f. *Ranuki* is categorized as change from below.

g. Use of written language in social media is different from written language in a traditional sense.

h. *Ranuki* is a linguistic change at both the levels of the community and the individual (age-grading).

i. The distribution/diffusion of *ranuki* that is biased by age-cohort can have something to do with the manner with which *ranuki* is acquired (natural acquisition vs. explicit learning).

j. *Ranuki* is preferred by female speakers when compared to male speakers, reflecting the general role of women as innovators in the development of linguistic change.

Turning our attention to future prospects about *ranuki*, two things arise that are worth mentioning. Now that *ranuki* (full form vs. reduced form) has a variety of social meanings (non-referential meaning, Sano 2018), the function of *ranuki* has changed to a stylistic choice rather than just as a by-product appeared in the course of linguistic change that is characterized by innovative/traditional dichotomy. It can be predicted that in order to perform their stylistic functions, both the full and the reduced forms need to survive and will co-exist, contrary to the pattern of linguistic change where innovative forms replace traditional forms. In view of the recent trend in the use of potential forms, for example in the media where performers in TV shows utter reduced forms while subtitles transcribing the utterance are in full forms, it could even be the case that full forms will function as written language while reduced forms will function as spoken language. Follow-up research is required for this prediction to be borne out.

The other point concerns the bi-directional feedback: the properties of *ranuki* have been uncovered along with the development or paradigm shift in linguistic theory and research methodology. At the same time, *ranuki* has thus far been providing significant clues to linguistic research, including the solution to linguistic issues, a better characterization of Japanese grammar, and the demonstration of linguistic variation and change in progress. As theories and methods progress, unnoticed aspects of *ranuki* will further be made clear; while on the other hand, *ranuki* will continue to offer an effective and useful testing ground for linguistic hypotheses, and contribute to the development of linguistic research.

References

Barr, Dale J. (2013) Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology* 4, 328. doi: 10.3389/fpsyg.2013.00328.

Barr, Dale J., Roger Levy, Christoph Scheepers and Harry J. Tily. (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 255-278.

Cheshire, Jenny. (2006) Age- and generation-specific use of language. In Ammon, Ulrich, Norbert Dittmar, Klaus Mattheier, and Peter Trudgill (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*, 1552-63. Berlin: Walter de Gruyter.

Cukor-Avila, Patricia, and Guy Bailey. (2013) Real time and apparent time. In J. K. Chambers and N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, 237-262. Oxford: Wiley-Blackwell.

Eckert, Penelope. (1997) Age as a sociolinguistic variable. In Florian Coulmas (ed.), *The Handbook of Sociolinguistics*, 151-67. Oxford, UK: Blackwell.

Fujimura,Itsuko, Shoju Chiba, Mieko Ohso, (2012) Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, 393-398.

Inoue, Fumio. (1998) *Nihongo wocchingu* [Watching Japanese]. Tokyo: Iwanami Shoten.

Inoue, Fumio and Kanetaka Yarimizu. (2002) *Jiten: Atarashii nihongo* [Dictionary: New Japanese]. Tokyo: Toyo Shoten.

Ito, Junko and Armin Mester. (2004) Morphological contrast and merger: *Ranuki* in Japanese. *Journal of Japanese Linguistics* 20, 1-18.

Jaeger, Florian. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4), 434-446.

Kanda, Sumiko. (1964) Mireru, dereru: Kanoohyoogen-no ugoki [Mireru and dereru: Movement in potential forms]. In Kenji Morioka (ed.), *Koogo Bumpoo Kooza* 3 *Yureteiru Bumpoo* [*Lectures on Grammar of Spoken Language* 3 *Variable Grammar*], 81-91. Tokyo: Meiji Shoin.

Kindaichi, Haruhiko, Takeshi Shibata, and Oki Hayashi (eds.). (1995) *Nihongo Hyakka Daijiten* [Encyclopedia of the Japanese Language]. Tokyo: Taishukan Shoten.

Kinsui, Satoshi. (2003) Ranukikotoba-no rekishiteki kenkyuu [Historical study of *ranuki*]. *Gekkan Gengo* [*Monthly Magazine of Linguistics*] 32(4), 56-62.

Koiso, Hanae, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino, Yoshiko Kawabata, Yayoi Tanaka, Yasuharu Den, and Kenya Nishikawa. (2017) Nihongo nichijookaiwa koopasu-no koochiku [Compilation of the Corpus of Everyday Japanese Conversation]. *Gengoshori gakkai dai 23-kai nenjitaikai happyoo rombunshuu* [*Proceedings of the 23rd annual meeting of the Association for Natural Language Processing*], 775-778.

Labov, William. (1963) The social Motivation of a sound change. *Word* 19, 273-309.

Labov, William. (1966) *The social stratification of English in New York City*. Cambridge, UK: Cambridge University Press.

Labov, William. (1990) The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 205-54.

Labov, William. (1994) *Principles of linguistic change: internal factors*. Oxford, UK: Blackwell.

Maekawa, Kikuo. (2008) KOTONOHA gendainihongo kakikotoba kinkoo koopasu-no kaihatsu: Tokushuu shiryookenkyuu-no genzai [KOTONOHA the Balanced Corpus of Contemporary Written Japanese: Special issue on current state of the research on linguistic databases]. *Nihongo-no Kenkyuu* [Studies on Japanese], 4(1), 82-95.

Maekawa, Kikuo, Takayuki Kagomiya, Hanae Koiso, Hideki Ogura, and Hideaki Kikuchi (2000) Nihongo hanashikotoba koopasu-no sekkee [On the design of the Corpus of Spontaneous Japanese]. *Onsee kenkyuu* [*Phonetic Studies*] 4(2), 51-61.

Matsuda, Kenjiro. (1993) Dissecting analogical leveling quantitatively: The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change* 5(1), 1-34.

Matsuda, Kenjiro. (2008) Tookyooshusshin giin-no hatsuwa-ni miru ranukikotoba-no heni-to henka [variation and change observed in speeches of Diet members from Tokyo]. In Kenjiro Matsuda (ed.), *Kokkaikaigiroku-o tsukatta nihongokenkyuu* [*Studies on Japanese language using Diet database*], 111-134. Tokyo: Hitsuji Shoboo.

Maruyama, Takehiko. (2016) Shoowa hanashikotoba koopasu-no keekaku-to temboo: 1950-nendai-no hanashikotoba kenkyuu shooshi [Plans and prospects for the Corpus of Shoowa Spoken Japanese: A brief history of research on spoken language in the 1950s]. *Senshuudaigaku jimbunkagaku kenkyuusho geppoo* [*Monthly report of the Institute for the Humanities at Senshu University*] 282, 39-55.

Nakamura, Michio. (1953) Koreru mireru tabereru-nadotoiu iikata-nitsuite-no oboegaki [Notes on expressions such as koreru, mireru, and tabereru] *Kindaichihakushi kokikinen gengo-minzoku ronsoo* [*Linguistics and folkloristics: Festschrift for Doctor Kindaichi on the occasion of his seventieth birthday*], 579-594. Tokyo: Sanseido.

Ogiso, Toshinobu, and Takenori Nakamura. (2013) Chuunagon-no tsukaikata [How to use Chuunagon]. In Maekawa, Kikuo (ed.), *Kooza Nihongo Koopasu* 1 *Koopasu Nyuumon* [*Lectures on Japanese Corpus* 1 *Introduction to Corpus*], 159-169. Tokyo: Asakura Shoten.

Okazaki, Kazuo. (1980) Mireru, Tabererugata-no kanoohyoogen-nituite: Gendaitookyoo-no chuugakusee kookoosee-nituite okonatta hitotsu-no choosa-kara [On mireru-, tabereru-type potential forms: Based on a survey conducted on middle and high school students in present-day Tokyo]. *Language in Life* 340, 64-70.

R Development Core Team. 1993-2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sankoff, Gillian. (2005) Cross-sectional and longitudinal studies in sociolinguistics. In Ammon, Ulrich, Norbert Dittmar, Klaus Mattheier, and Peter Trudgill (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*, 1003-1013. Berlin: Mouton de Gruyter.

Sano, Shin-ichiro. (2011) Real-time demonstration of the interaction among internal and external factors in language change: A corpus study. *Gengo kenkyuu* 139, 1-27.

Sano, Shin-ichiro. (2012) Nihongo dooshi kanookee-no hensen-o tadoru [Tracing the change of potential forms in Japanese], In Junko Hibiya (ed.), Hajimete manabu shakaigengogaku [Introduction to sociolinguistics], 190-208. Tokyo: Minerva Shobo.

Sano, Shin-ichiro. (2013) Dynamic shift of word frequency effect in the course of linguistic change. *Working Papers from NWAV-AP2*. http://www.ninjal.ac.jp/socioling/nwavap02/Sano-NWAVAP2-2013.pdf

Sano, Shin-ichiro. (2015) Optimization of the verbal inflectional paradigm by the cyclic application of morphophonological processes: Evidence from potential forms in Japanese. *Open Linguistics* 1(1), 580-595.

Sano, Shin-ichiro. (2018) Productive use of indexicalized variable in social interaction: The case of *ranuki* in Japanese. In Shin Fukuda, Mary Shin Kim, and Mee-Jeong Park (eds.), *Japanese/Korean Linguistics* 25, 369-382. Stanford, CA: CSLI Publications.

Sano, Shin-ichiro. (2019a) Hanashikotoba-no keenenhenka-no teeryooteki kijutsu: "Shoowa hanashikotoba koopasu," "Nihongo nitijookaiwa koopasu," "Nihongo hanashikotoba koopasu"-o motiite." [Quantitative analysis of the chronological transition in spoken Japanese: Using the Showa Spoken Japanese Corpus, the

Corpus of Everyday Japanese Conversation, and the Corpus of Spontaneous Japanese] *Proceedings of the 158th Meeting of Linguistic Society of Japan*, 142-148.

Sano, Shin-ichiro. (2019b) Patterns of variable *ranuki* in Japanese: Identity avoidance and register. *Phonological Studies* 22, 75-82.

Sano, Shin-ichiro. (to appear) Ranuki kotoba [Ra-deletion]. In Takehiko Maruyama (ed.), *Koopasu-de manabu nihongogaku: Nihongo-no bumpoo-to onsee* [Learning Japanese linguistics using corpora: Grammar and sounds of Japanese]. Tokyo: Asakura Shoten.

Sherwood, Stacey. (2014) Social pressures condition *ranuki* in the potential form of Japanese verbs. Talk presented at NWAV-AP3 at Victoria University of Wellington.

Sherwood, Stacey. (2016) Indicating and perceiving social hierarchy through language variation: the case of *ranuki* in Japanese. BA thesis, Western Sydney University.

Shibuya, Katsumi. (1993) Nihongo kanoohyoogen-no shosoo-to hatten [Aspects and development of potential forms in Japanese]. *Oosakadaigaku bungakubu kiyoo* [Bulletin of the Faculty of Letters at Osaka University], 33(1), 1-262.

Yamaguchi, Masaya. (2014) Zembun kensakushisutemu himawari-o mochiita kisongengoshiryoo-no katsuyoohoohoo-no kentoo [Examination of the practical use of pre-existing linguistic resources using the full text search system Himawari]. *Dai 6-kai Nihongo koopasu waakushoppu yokooshuu* [*Proceedings of the 6th workshop on Japanese corpus*], 151-156.

Wagner, Suzanne Evans. (2012) Age grading in sociolinguistic theory. *Language and Linguistics Compass* 6(6), 371-382.