Shin-Ichiro Sano

The Role of Exemplars and Lexical Frequency in Rendaku

Abstract: This paper examines the relationship between lexical frequency and phonological processes, focusing on rendaku in Japanese. Recently, the effect of lexical frequency on linguistic processes, either direct or indirect, has been confirmed in a growing body of studies. However, little attention has been paid to the potential effect of lexical frequency on rendaku. With this background, I examined the effect of lexical frequency on the applicability of rendaku, and developed an analogy-based model by incorporating lexical frequency. The results demonstrate (i) that lexical frequency affects the applicability of rendaku, (ii) less frequent compounds follow the existing patterns that the exemplar offers, and (iii) that rendaku is psychologically real; in other words, rendaku is productively applied to innovative forms, and such an application can be accounted for by the current model.

Keywords: rendaku, lexical frequency, Exemplar Theory, corpus, phonological process, voicing, analogy, similarity, psychological reality, productivity

1 Introduction

This paper examines the relationship between lexical frequency and phonological processes, focusing on rendaku (also known as sequential voicing) in Japanese. In recent years, the role of lexical frequency in shaping phonological patterns has been extensively discussed in a range of linguistic phenomena, specifically in the studies of language variation and change. The effect of lexical frequency on linguistic processes, either direct or indirect, has been confirmed in a growing body of studies (Bybee, 2001, 2002, 2010; Erker & Guy, 2012; Frisch, 2011; Johnson, 2007; Phillips, 2006; Pierrehumbert 2002, 2006). For example, in usage-based and exemplar theoretic models of phonology, where the lexicon is assumed to be a highly interconnected network of lexical exemplars (words or phrases) with rich information including frequency distribution (Bybee, 2001, 2002; Johnson, 2007; Pierrehumbert, 2002, 2006), highly frequent words have stronger mental representations and are more accessible during production; while infrequent words have weaker mental representations, and hence are more difficult to access. This difference in the strength of the mental representation is believed to drive variation and change, at least to some degree. In particular, highly frequent words are more likely to undergo phonological processes or to be the target of sound change; on the other hand, low-frequency words are more likely to be regularized, or to be the target of analogical change.

Thus, given these previous findings, it seems reasonable to assume that lexical frequency plays some role also in the distribution of rendaku, a morphophonological pattern in Japanese. However, little
attention has thus far been paid to the effect of lexical frequency on rendaku. Although Ohno (2000) proposes an analogy-based model of rendaku based on their phonological and/or semantic similarities, the mechanism of analogy, especially how the similarities are defined, is not specified and remains an open question.\(^2\) If an exemplar-theoretic model is a correct model of our phonological knowledge, then lexical frequency should play a crucial role in phonology of rendaku. Specifically, highly frequent forms should participate in phonological processes, whereas infrequent forms should show analogy to existing lexical patterns. Furthermore, to the extent that the analogy plays some role in the distribution of rendaku, one can hypothesize that the directionality of the analogy follows the prediction made by exemplar-theoretic models; i.e., low-frequency words (but not high-frequency words) rely on the existing exemplar patterns.

Even if the role of lexical frequency is identified, the modeling of productivity of novel forms still poses a problem for such a model. In particular, assuming that linguistic knowledge is not a system of abstract rules, but an inventory of concrete examples, we cannot account for the native speakers’ ability to correctly judge the well-formedness of novel compounds, and to correctly apply rendaku to such novel forms. Indeed, there are a number of studies that support the claim that rendaku is a productive phonological process, making the productivity of rendaku undeniable (Fukuda & Fukuda, 1999; Kawahara, 2015; Ito & Mester, 1986, 2003; Kubozono, 2005; Otsu, 1980; Vance, 2014).

With this background, this study presents an empirical examination of the effect of lexical frequency on the applicability of rendaku using the Balanced Corpus of Contemporary Japanese (henceforth BCCWJ, National Institute for Japanese Language and Linguistics, 2011). Specifically, building upon Ohno’s (2000) analogy-based model of rendaku, I propose a model that incorporates the role of lexical frequency. Furthermore, the validity of rendaku in terms of its productivity is also examined, using the Corpus of Spontaneous Japanese (henceforth CSJ, National Institute for Japanese Language and Linguistics, 2008). The results demonstrate that (i) the lexical frequency plays a role in determining whether or not the items undergo rendaku, (ii) less frequent compounds follow the existing patterns that the related exemplar offers, and (iii) that rendaku is psychologically real; in other words, rendaku is productively applied to innovative forms, and such an application can be accounted for by the current model.

The remainder of this paper is organized as follows. In Section 2, I briefly describe the properties of rendaku, and review the research on rendaku, followed by the goal and the research question of this study. Subsequently, I present two separate studies. In Section 3, I examine the effect of lexical frequency on the applicability of rendaku. Section 4 examines the psychological reality of rendaku. Finally, Section 5 concludes the discussion.

### 2 Background

In this section, I will firstly describe the properties of rendaku followed by a review of the research on rendaku. Finally, I will clarify the goal of the study and the research questions to be addressed in this study.

#### 2.1 Rendaku

Rendaku is one of the well-known and well-studied morphophonological processes in Japanese, where the initial voiceless obstruent of the second member of morphologically derived words (most often compounds) becomes voiced (Vance, 1979, 1980). As examples (1a) and (1b) show, when two nouns, such as /oo/ ‘big’ and /tako/ ‘octopus’ are morphologically concatenated, the initial consonant /t/ of the second noun becomes voiced.

Another work that addresses the effect of analogy on the applicability of rendaku is Asao (2007). Assuming the analogy-based model, Asao (2007) conducted a statistical examination of the effects of the following factors: item, segment (consonant/vowel), part of speech, etymological status, length (mora counted) for both the first and the second members, and demonstrates the validity of these factors in that the analogy-based model with correspondence in all of these factors significantly improves rendaku predictability. The effect of lexical frequency, however, is not included in the investigation.
voiced, producing /oodako/ 'big octopus.' Similarly, a compound consisting of /hosi/ 'star' and /sora/ 'sky' is realized as /hosizora/ 'starry sky,' with the initial consonant /s/ of the second member being voiced.

(1) Examples of rendaku

a. oo + tako ‘big + octopus’ \(\Rightarrow\) oodako ‘big octopus’

b. hosi + sora ‘star + sky’ \(\Rightarrow\) hosizora ‘starry sky’

Not all compounds, however, undergo this voicing process; that is, rendaku shows variability. Prior research has identified various kinds of lexical, phonological and morphological factors that affect the likelihood of rendaku application (Fukuda & Fukuda, 1999; Hamada, 1960; Ito & Mester, 1986, 2003; Kawahara, 2012; Lyman, 1894; Martin, 1952, 1975; Ohno, 2000; Otsu, 1980; Vance, 1979, 1980, 1987, 1996, 2008, among others). The factors identified in this research tradition include, for example, the segmental properties of the preceding and following contexts of a rendaku-undergoing segment, the length of the first member of the compound (Rosen, 2001, 2003), and the internal structure of the compound. I present some representative factors in (2).

(2) Representative factors affecting rendaku

a. Lyman’s Law (OCP(voice), Ito & Mester, 1986; Lyman, 1984)

Voiced obstruent(s) in the second member blocks rendaku.

e. g. hosi + kuzu ‘star dust’ \(\Rightarrow\) hosikuzu (*hosiguzu)

b. Right branch condition (Otsu, 1980)

Nouns on the right (*left) branch undergo rendaku.

e. g. nise + tanuki + siru ‘fake raccoon soup’

\(\Rightarrow\) [[nise [tanuki jiru]] ‘raccoon soup that is fake’

\(\Rightarrow\) [[nise danuki] jiru] ‘soup made from a fake raccoon’

c. Lexical Strata (Ito & Mester, 1986)

Rendaku applies only to Yamato words, not to Sino-Japanese, or foreign loanwords.5

e. g. gengo + kenkyuu ‘language study’ \(\Rightarrow\) gengokenkyuu (*gengogenkyuu)

waarudo + kappu ‘world cup’ \(\Rightarrow\) waarudokappu (*waarudogappu)

The most famous and presumably the most influential factor is Lyman’s Law (also known as OCP(voice)) in (2a), where rendaku is blocked when the second member of a compound already contains a voiced obstruent. /hosi/ + /kuzu/, for example, does not form */hosiguzu/; the voicing of /k/ remains unaltered, because the second noun contains a voiced obstruent /z/, which blocks the application of rendaku. The right branch condition in (2b) refers to the structure of the compound, and it specifies that only (the second) members on the right branch can undergo rendaku. Thus, even though both of the examples consist of exactly the same members /nise/ ‘fake,’ /tanuki/ ‘raccoon,’ and /siru/ ‘soup,’ only the compound with the structure [[nise danuki] jiru], where the second member is on the right branch, can undergo rendaku. The final example is Lexical strata in (2c). The Japanese lexicon consists of four kinds of strata organized according to the etymological status of the lexical items: Yamato (native) words, Sino-Japanese words (old borrowings from Chinese), foreign loanwords (recent borrowings mostly from English), and onomatopoeia or sound symbolic/mimetic words (Ito & Mester, 1986). Among these strata, rendaku applies mainly to Yamato words, and not to words of the other strata. Therefore, Sino-Japanese words, such as /gengo+kenkyuu/ and foreign

---

3 The transcription system adopted in this paper is the standard phonemic transcription of Japanese, rather than the narrow phonetic transcription (Vance, 1987, 2008), as phonetic details are outside the scope of this paper.

4 /siru/ is the third member of the compound also undergoes rendaku, resulting in /jiru/.

5 A number of counterexamples to this generalization are reported in the previous works (Kubozono, 1999; Takayama, 1999, 2005; Vance, 1987, 1996). In Section 4, I used test items with Sino-Japanese words as the second member, standing on the position that Sino-Japanese words also exhibit rendaku.
loanwords, such as /waarudo+kappu/ do not undergo rendaku. As these factors and others come into play, the distribution of rendaku is not uniform.

2.2 Research on rendaku

As described above, rendaku has long been studied in a variety of linguistic subfields. This research on rendaku is generally classified into two kinds of approaches. The first of these is the grammatical approach, where rendaku is regarded as a morphophonological rule (e.g. Ito & Mester, 1986, 2003; Kawahara & Sano, 2014; Kubozono, 2005; Otsu, 1980). The second of these is the lexical approach, where rendaku is regarded as a lexical property; in other words, the information about whether or not items undergo rendaku is stored in the lexicon on an item by item basis (e.g. Hamada 1960; Vance 1980, 1987, 1996). One of the major issues to be addressed in studies on rendaku, both in the grammatical approach and in the lexical approach, is how to account for the exceptions to rendaku. In this section, I will review the research on rendaku and the problems therein, and describe Ohno’s (2000) analogy-based account in detail.

In order to account for the exceptions to rendaku, the following rule-based analysis has thus far been proposed (Ito & Mester, 1986).

(3) Analysis of rendaku to account for exceptions
   a. Rule: [+rendaku] & rendaku rule
   b. Exception: lexicalized compound

This analysis assumes that each of the lexical items comprising compounds has a feature bearing on rendaku, as in [+rendaku] and [-rendaku], and the rendaku rule targets only the items with a [+rendaku] feature. As schematized in (3), if the second member of a compound has a [+rendaku] feature, then the rendaku rule applies, and the initial voiceless obstruent in the second member becomes voiced. There are, however, many exceptions to this rule, even though the conditions for rendaku identified in the previous literature are satisfied. The lexical approach proposes that the rendaku-immunity of these exceptional items is lexically specified, and is not a byproduct of the productive morphological processes. In other words, these items are treated as single lexical items rather than compounds in the strictest sense of the word. I will illustrate this point with examples in (4).

(4) Examples of rendaku (non-)application
   a. [+rendaku] (e.g. /kami/ ‘hair’)
      maegami ‘bangs’
      nihongami ‘Japanese hair’
      kurokami ‘black hair’ (lexicalized)
   b. [-rendaku] (e.g. /ti/ ‘blood’)
      ikiti ‘vital blood’
      kaeriti ‘spurt of blood (from wound)’
      hanadi ‘nosebleed’ (lexicalized)

The example (4a) shows that /-kami/ ‘hair’ has the [+rendaku] feature, and accordingly, the rendaku rule applies to compounds that involve /-kami/ as the second member. As a result, the items such as /maegami/ ‘bangs,’ and /nihongami/ ‘Japanese hair’ undergo rendaku. However, /kurokami/ ‘black hair’ does not undergo rendaku, even though the second member /kami/ has the [+rendaku] feature, because /kurokami/

---

6 There are works that support both lexical and productive aspects of rendaku. For example, Fukuda & Fukuda (1999), through their experiments on children with specific language impairment (SLI), suggest that some familiar compounds with rendaku are stored in the lexicon, but others (including novel compounds) are processed in a productive manner.

7 In the following, examples of lexical items that appear as the second member of compounds are presented with hyphenation, as in /-kami/.
is a lexicalized compound. Similarly in (4b), because /-ti/ ‘blood’ has the [-rendaku] feature, the rendaku rule does not apply to compounds that involve /-ti/ as the second member. It follows that the items such as /iki/ ‘vital blood,’ and /kaeriti/ ‘spurt of blood’ do not undergo rendaku. However, /hanadidi/ ‘nosebleed’ undergoes rendaku, because /hanadidi/ is a lexicalized compound.

Building upon the analysis above, Ohno (2000) conducted a word-formation task to address the following research question: if the rendaku rule is psychologically real, then novel compounds would follow the value of the [rendaku] feature of the second members (because such compounds should not be lexicalized). However, as shown in (5), the experiment did not provide evidence for the psychological reality of the rendaku rule.

(5) Result of Ohno’s (2000) experiment
a. [+rendaku]
   -kami ‘hair’ => sirokami (sirogami) ‘white hair’
   b. [-rendaku]
   -ti ‘blood’ => mimidi (mimiti) ‘ear bleed’

As shown in (5a), a novel compound /sirokami/ ‘white hair’ does not undergo rendaku, even though /kami/ has the [+rendaku] feature. On the other hand, /mimidi/ ‘ear bleed’ in (5b) undergoes rendaku, even though /ti/ has the [-rendaku] feature. Thus, Ohno (2000), on the basis of the result of his experiments, argues that the applicability of rendaku is a matter of lexical analogy, not of a phonological rule. To clarify the point in question, I will describe Ohno’s (2000) proposal in more detail.

(6) Analogy-based account of rendaku
a. Refer to existing compounds (semantically and/or phonologically similar).

   Follow the referents’ response to rendaku.\(^8\)
   e.g. kurokami ‘black hair’ => sirokami ‘white hair’
   hanadidi ‘nosebleed’ => mimidi ‘ear bleed’

b. Do not undergo rendaku, if there is no referent.
   -taka ‘hawk’ never undergoes rendaku.
   e.g. hagetaka ‘vulture,’ yotaka ‘night hawk’
   surume ‘dried squid’ never undergoes compounding.

Ohno’s (2000) analogy-based account of rendaku is twofold. Firstly, as shown in (6a) native speakers determine the rendaku/non-rendaku form of a compound with reference to existing compounds that have semantic and/or phonological similarities, and follow the referents’ response to rendaku. For example, native speakers determine the rendaku/non-rendaku form of /sirokami/ ‘white hair,’ a novel compound, with reference to /kurokami/ ‘black hair,’ an existing compound, because /kurokami/ is semantically and phonologically similar to /sirokami/. Since /kurokami/ does not undergo rendaku, native speakers follow the /kurokami/’s response to rendaku (non-rendaku), and apply it to /sirokami/. It follows that /sirokami/ also does not undergo rendaku. Similarly, the rendaku/non-rendaku of /mimidi/ ‘ear bleed,’ a novel compound, is determined with reference to /hanadidi/ ‘nosebleed,’ because these two compounds share the semantic and phonological similarities. Since /hanadidi/ undergoes rendaku, native speakers follow the /hanadidi/’s response to rendaku, and apply it to /mimidi/. Consequently, /mimidi/ also undergoes rendaku.

Secondly, if there is no possible referent, compounds do not undergo rendaku; in other words, if native speakers cannot refer to existing compounds with rendaku, they choose non-rendaku forms as a default. For example, /-taka/ ‘hawk’ as the second member of a compound never undergoes rendaku, no matter what the first member is. Therefore, such compounds as /X-taka/ do not provide any example of rendaku that otherwise

\(^8\) Throughout this paper, the referent means an existing compound that provides the pattern to follow, unlike the established meaning in other fields of linguistics.
would be a referent. Similarly, there are no compounds containing the word /surume/ ‘dried squid,’ and accordingly /surume/ also does not provide any example of rendaku. Thus, any novel compounds that take /-taka/ or /-surume/ as the second member cannot undergo rendaku, because there is no referent.

In the current paper, I expand on Ohno’s (2000) analogy-based account of rendaku by incorporating lexical frequency. In the next section, I will clarify the goal and the research questions to be addressed.

2.3 Research question and goal

In this section, I will clarify the research questions by introducing potential roles of lexical frequency in rendaku suggested by Ohno’s (2000) observations and by revising the analogy-based model of rendaku. Specifically, the research questions are based on three observations.

The first observation is related to the second portion of Ohno’s (2000) proposal. As shown in (6b), /-taka/ as the second member of a compound does not undergo rendaku, no matter what the first member is. Therefore, compounds such as /X-taka/ do not provide any example of rendaku. This can be interpreted as “the frequency of compounds with /-taka/ as the second member that undergo rendaku is zero.” Similarly, /surume/ also does not undergo compounding, and accordingly /surume/ also does not provide any example of rendaku. This can also be interpreted as “the frequency of compounds with /-surume/ as the second member that undergo rendaku is zero.” In other words, the frequency of forms without rendaku is higher than those with rendaku, and a novel compound should select a non-rendaku form. Thus, the examples of /-taka/ and /-surume/ imply that lexical frequency plays a role in the choice of rendaku/non-rendaku forms.

In addition, the directionality of the analogy is also suggested. That is, low-frequency words (including novel compounds) rely on the existing exemplar patterns, as predicted by exemplar-theoretic models (Bybee, 2001).

Secondly, in the analogy-based model of rendaku, native speakers determine the rendaku/non-rendaku form of a compound with reference to existing compounds, and the reference is established based on the semantic and/or phonological similarities. For instance, /sirokami/’s application of rendaku in (6a) is determined with reference to /kurokami/, the existing compound. For some compounds, however, there could be multiple potential candidates for the referent. For example, there are many compounds that involve /-kawa/ ‘river’ as the second member. In the CSJ, both rendaku forms of /-kawa/ and non-rendaku forms of /-kawa/ are attested, such as /tamagawa/, /edogawa/, /sakaigawa/, and /arakawa/ (river-names). That is, for any novel compound with /-kawa/ as the second member, there are multiple potential candidates for the referent. Thus, assuming the model proposed in Ohno (2000), when native speakers are presented with novel compounds with /-kawa/ as the second member, as in /nihon+kawa/ ‘Japan river,’ and /tokyoo+kawa/ ‘Tokyo river,’ they cannot unambiguously determine which referent to select. Then the question arises: “on what basis is the referent selected?” One might well assume that some additional index would be at work. Based on the assumption of exemplar-theoretic models, I assume in this study that the index in question is the lexical frequency, where more frequent examples are prominent, and these examples can be a referent; on the other hand, less frequent examples are not prominent, and these examples cannot be a referent.

With this in mind, I first develop the reference in the analogy-based model by incorporating lexical frequency into the model, as illustrated in Figure 1.

![Figure 1. Analogy-based model with lexical frequency](https://example.com)
Suppose the situation where /tamagawa/ has the highest frequency among many candidates that share semantic and phonological similarities. Novel compounds /nihon+kawa/ and /tookyoo+kawa/ refer to /tamagawa/, as this candidate has the highest frequency. Since /tamagawa/ undergoes rendaku, native speakers follow /tamagawa/’s response to rendaku, and apply it to /nihon+kawa/ and /tookyoo+kawa/. Consequently, /nihon+kawa/ and /tookyoo+kawa/ also undergo rendaku, producing /nihongawa/ and /tookyogawa/. Furthermore, taking the idea one step further, one can even argue that without the lexical frequency, native speakers cannot determine which is the existing compound (referent) and which is the novel compound, and the reference cannot be established.

The third observation is that in spontaneous speech, the distribution of rendaku is not categorical. That is, there are a non-negligible number of compounds that have both a rendaku variant and a non-rendaku variant. For example, in the CSJ /nan+ken/ ‘how many cases/doors’ is realized either as /nangen/ (with rendaku) or as /nankan/ (without rendaku). Similarly, /yon+kai/ ‘fourth floor’ is realized either as /yongai/ (with rendaku) or as /yonkai/ (without rendaku). If this is the case, however, the proposed model runs into a problem: in spite of the fact that the most frequent compound is selected as the referent, both the rendaku form and the non-rendaku form occur. Once again this raises the question: “on what basis is the referent selected?”

In order to solve this problem, I revise the reference in the analogy-based model as illustrated in Figure 2.

I assume that the choice of the rendaku/non-rendaku form is not based on the frequency distribution of existing compounds, but on the frequency distribution of variants (with rendaku or without rendaku) across every compound that involves the lexical item in question as the second member. Taking /-ken/ as an example, /-ken/ as the second member of a compound is realized either as a variant with rendaku or as a variant without rendaku. As shown above, exactly the same compound has both of these two variants, as in /nangen/ and /nankan/. Suppose that /nan+ken/ has the highest frequency among many compounds that involve /-ken/ as the second member, the rendaku (non-)application of a novel compound such as /ichioku+ken/ ‘one hundred million cases/doors’ is determined with reference to /nan+ken/. If the reference is established between a novel compound and the frequency distribution of the variants of the referent. For example, /-ken/ occurs in many compounds, such as /ikken/ ‘one case/door,’ /niken/ ‘two cases/doors,’ and /sangen/ ‘three cases/doors.’ Although /-ken/ has both rendaku and non-rendaku variants, the non-rendaku variant /-gen/ is more frequent than the rendaku variant /-gen/ in the entire frequency distribution of the compounds. Then, /ichioku+ken/’s application of rendaku is determined following the model presented by rendition of /-gen/ with the higher frequency, resulting...
in /ichiokuken/. In the CSJ, /-ken/ is more frequent (N=50) than /-gen/ (N=4), and /ichiokuken/ without rendaku is intuitively well-formed. Thus, my assumption is supported; that is, the choice of the rendaku/non-rendaku form is based on the frequency distribution of variants across every compound containing the lexical item in question as the second member.

The revised model can also account for the cases where there are multiple candidates for the referent. As shown above, /kawa/ also occurs in a wide range of compounds, such as /tama\(g\)awa/, /edo\(g\)awa/, /saka\(g\)awa/, and /ara\(k\)awa/ (river-names). In those compounds that contain /-kawa/ as the second member, within-compound variation is not observed. In the entire frequency distribution in compounding, the non-rendaku variant /-gawa/ is more frequent than the rendaku variant /-kawa/. Accordingly, /nihon+kawa/ and /tookyoo+kawa/’s application of rendaku is determined following the model presented by rendition of /-gawa/ with the higher frequency, resulting in /nihongawa/ and /tookyogawa/. In the CSJ, /-gawa/ is more frequent (N=22) than /-kawa/ (N=4), and /nihongawa/ and /tookyogawa/ with rendaku are intuitively well-formed. The revised model was again supported. Thus, in the revised model, the rendaku (non-)application of a single compound does not matter, rather what is crucial is the item-wide trend observed in compounding.

Putting these observations together, the general question is whether or not lexical frequency plays a role in the analogy and the reference. More specifically, I will examine the validity of the analogy-based model with lexical frequency in Figure 2 using Japanese corpora. I hypothesize that more frequent examples are prominent, and therefore more likely to be used as a referent. On the other hand, less frequent examples are not prominent, and therefore less likely to be used as a referent. Instead, less frequent words follow the existing patterns that the exemplar offers. In other words, rendaku (non-)application of compounds with low frequency (including novel compounds) follows the rendaku (non-)application of more frequent variants (to the extent that the conditions such as Lyman’s Law are met). In doing so, I will elaborate the mechanism of analogy based on the lexical frequency. In addition, what has been viewed as exceptions to rendaku can also be accounted for by lexical frequency. Furthermore, in order to solve the problem of productivity that the exemplar-theoretic models face, the validity of rendaku in terms of its productivity is also examined. To preview the results, these hypotheses are borne out: (i) rendaku (non-)application of compounds with low frequency follow the one of more frequent variants, (ii) rendaku is productively applied beyond a speaker’s lexicon, supporting the psychological reality of rendaku. However, novel applications can be accounted for by the analogy-based model with lexical frequency.

### 3 Effect of lexical frequency

In this section, I present the empirical examination of the effect of lexical frequency on the applicability of rendaku in an analogy-based model. Particularly, the question to be addressed here is whether or not the rendaku (non-)application of compounds with low frequency follow the same pattern as more frequent variants.

#### 3.1 Method

In order to examine the effect of lexical frequency, a large-scale quantitative analysis needs to be conducted. At the same time, the data used in such an analysis should reflect actual language use. Thus, the issue of representativeness needs to be kept in mind. In order to fulfill these requirements, large-scale corpora are a useful tool. For the purpose of this study, the BCCWJ and the CSJ are the most reliable candidates among such corpora of Japanese. The BCCWJ consists of 100 million words, and is the largest corpora of Japanese as of now. Additionally, the BCCWJ is well-balanced in terms of 1) the kinds of register/genre and 2) the proportion (amount) of each register/genre in the corpus. In other words, the BCCWJ best represents the (written) language use of modern Japanese, among other Japanese written corpora. Similarly, the CSJ consists of 7.5 million words. This is the largest among the speech corpora of Japanese where the representativeness is considered. While most of the speech samples are monologues, the CSJ does provide limited register-
level variation, namely “Academic Presentation Speech,” and “Simulated Public Speaking.” Furthermore, speakers were chosen from a wide range of ages, gender, and educational backgrounds.

For the first examination, I used the BCCWJ, because the CSJ is insufficient for the present purpose in terms of its scale. By using the BCCWJ, I first examined whether or not lexical frequency plays a role in the choice of rendaku application, in a way assumed in the analogy-based model with lexical frequency in Figure 2. In the current examination, I focused on some specific lexical items. The test items used in the first examination are presented in (7).

(7) Test items
a. siro+kami ‘white hair’  =>  sirokami (sirogami)
   b. mimi+ti ‘ear bleed’  =>  mimidi (mimiti)

The test items consisted of two compounds. These two examples were taken from Ohno (2000). In the previous experiment (Ohno, 2000), the novel compound that consists of /siro/ ‘white’ and /-kami/ ‘hair’ was shown to produce a voiceless variant, as in /sirokami/. Similarly, the combination of /mimi/ ‘ear’ and /-ti/ ‘bleed’ was shown to produce a voiced variant, as in /mimidi/.

I retrieved the target items from the BCCWJ in the following manner: (i) the online search system of the BCCWJ Chuuagon version 1.0.2 (National Institute for Japanese Language and Linguistics, 2011) was used; (ii) All sub-corpora of BCCWJ were targeted; (iii) Both phonetic and morphological information was employed; (iv) The targeted compounds consisted of two nouns, and the second member was either one of /-kami/ or /-ti/; (v) The search formulae with regular expressions were used, as shown in (8).

(8) Search formulae
a. Search formula for “X+kami”
   キー: 語彙素 LIKE “_髪” WITH OPTIONS unit=”1” AND tglWords=”10” AND tglKugiri=”|” AND tglFixVariable=”2”
   b. Search formula for “X+ti”
   キー: 語彙素 LIKE “_血” WITH OPTIONS unit=”1” AND tglWords=”10” AND tglKugiri=”|” AND tglFixVariable=”2”

The retrieved data was then subjected to data screening. Some Chinese characters have multiple readings, and accordingly other readings than /-kami/ and /-ti/ were removed. For example, the Chinese character “髪” that represents /-kami/ is also read as /hatu/, and any forms with such a reading were removed from consideration.9 I equate the number of tokens retrieved to the lexical frequency of that form. The analysis followed the procedure illustrated in Figure 3.

![Figure 3: Analysis procedure](image)

For both /-kami/ and /-ti/, I compared the frequency of the forms with rendaku against the frequency of the forms without rendaku, and in this way determined whether application or non-application of rendaku was more frequent for all compounds containing the words under study as a second member. I then counted the frequency of both the rendaku application and non-application forms for the test items listed in (7). If the hypothesis is correct, then the rendaku (non-)application of the test items should coincide with the general pattern. It follows for example that because the combination of /siro/ and /-kami/ produces the voiceless

---

9 Also “血” that represents /ti/ has /ketsu/ reading.
variant as in /sirokami/, /-kami/, the voiceless variant, should be more frequent than /-gami/, the voiced variant.

### 3.2 Results

An exhaustive search of the BCCWJ resulted in 1,037 tokens. The distribution of the tokens is shown in Table 1.

Table 1. Distribution of voiced/voiceless variants of the second noun in compounds in the corpus, and in the test items

<table>
<thead>
<tr>
<th>2nd noun</th>
<th>-kami ‘hair’ (N=811)</th>
<th>-ti ‘blood’ (N=226)</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>voiced / voiceless (382/429)</td>
<td>voiced / voiceless (202/24)</td>
</tr>
<tr>
<td>test item</td>
<td>sirokami ‘white hair’</td>
<td>mimidi ‘ear bleed’</td>
</tr>
</tbody>
</table>

As for /-kami/, the voiceless variant /-kami/ (N=429) is more frequent than the voiced variant /-gami/ (N=382). The test item /siro+kami/ does not undergo rendaku, and produces a voiceless variant, as in /sirokami/. Therefore, we can argue that the voicing value of /sirokami/ is consistent with the more frequent variant /-kami/. Similarly for /-ti/, the voiced variant /-di/ (N=202) is more frequent than the voiceless variant /-ti/ (N=24). The test item /mimi+ti/ undergoes rendaku, producing the voiceless variant /mimidi/. The voicing value of /mimidi/ is also consistent with the more frequent variant /-di/. Thus, the voicing value of the test items matches the value of more frequent variants.

### 3.3 Discussion

The results of the first analysis together with the preliminary investigation of the CSJ in Section 2.3 support the effect of lexical frequency on the (non-)application of rendaku and the assumption on which the analogy-based model with lexical frequency is grounded. That is, the choice of the rendaku/non-rendaku form is based on the frequency distribution of variants (with rendaku or without rendaku) of lexical items in every compound that involve those lexical items as the second member, and the rendaku (non-)application of compounds with low frequency follows the one of more frequent variants, as far as other conditions such as Lyman’s Law are met.

In addition, the result shows that what have been viewed as exceptions to rendaku can be accounted for by the analogy-based model with lexical frequency. In the literature, /sirokami/ and /mimidi/ have been viewed as exceptions to rendaku, because these items do not follow the trend observed in other compounds of similar kinds. For example, /sirokami/ does not undergo rendaku even though other compounds such as /maegami/ and /nihongami/ undergo rendaku. Likewise, /mimidi/ undergoes rendaku even though other compounds such as /ikiti/ and /kaeriti/ do not undergo rendaku. What causes the problem in the previous literature is that the rendaku (non-)application of a certain lexical item is assessed by looking at the number of compounds that involve the lexical item in question (type frequency), instead of the distribution in the actual language use (token frequency). Many compounds with /-kami/ such as /maegami/ and /nihongami/ undergo rendaku, and accordingly /-kami/ is a rendaku-undergoing item; nevertheless, /sirokami/ does not undergo rendaku in Ohno’s experiment, and hence it was considered to be an exception. But as the current examination shows, if we examine the token frequency, /-kami/ is more frequent than /-gami/, and we can

---

10 An anonymous reviewer pointed out that as one of the factors affecting the applicability of rendaku, Sato (1989) proposes the effect of morphological structure, i.e., compounds are unlikely to undergo rendaku if the first member is derived from adjective. If this is the case, /siro+kami/ and /kuro+kami/ derived from /siori (adjective) + kami/ and /kuroi (adjective) + kami/ are less likely to undergo rendaku than /nihon+gami/ derived from /nihon (noun) + kami/. Testing this alternative is beyond the scope of this study, and I will leave this issue for future research. The effect of lexical frequency, however, cannot be nullified solely because of the possible effect of morphological structure.
argue that compounds with low frequency, such as the novel compound /sirokami/, follows the (non-)application of more frequent variant /-kami/, resulting in /sirokami/. Thus, the non-application of rendaku to /sirokami/ is not exceptional. Likewise, the application of rendaku seen in /mimidi/ is accounted for in the same manner. This emphasizes the importance of token frequency, rather than type frequency, and of natural linguistic data in linguistic research.

The current result also emphasizes the importance of item-wide trends in the entire frequency distribution. This is consistent with what the exemplar-theoretic models assume, i.e., low-frequency words rely on the existing exemplar patterns. In this case, the novel (and hence low frequency) compounds /siro+kami/ and /mimi+ti/ rely on the existing patterns that the referents or exemplars offer.

In summary, the result of the first analysis confirms that the rendaku (non)-application of compounds with low frequency follow the one of more frequent variants, and also supports the analogy-based model with lexical frequency that is in accord with the assumption of the exemplar-theoretic models.

### 4 Psychological reality of rendaku

In the discussion so far, the roles of analogy and lexical frequency in rendaku have been confirmed. Then, the next step is to examine the productivity of rendaku with the question: “Is rendaku a psychological reality, or is it solely based on lexical properties?” As evidence for the psychological reality of rendaku, we need to demonstrate the productivity of rendaku. For this purpose, I focused on variability in rendaku. As shown in (9), the following existing compounds are variably produced either with rendaku (voiced variants) or without rendaku (voiceless variants). An increasing number of voiced variants that replace the traditional voiceless variants are also taken up by NHK Hoosoo Bunka Kenkyuujo (1998, 2005) that showcase the norm of language use for broadcasters.

(9) Variability in rendaku

<table>
<thead>
<tr>
<th>Compound</th>
<th>Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>ka + husoku ‘deficiency and excess’</td>
<td>kahu or kahu</td>
</tr>
<tr>
<td>i + son ‘dependence’</td>
<td>i or i</td>
</tr>
<tr>
<td>kyoo + son ‘coexistence’</td>
<td>kyoo or kyoo</td>
</tr>
<tr>
<td>ki + son ‘existing’</td>
<td>kiso or kiso</td>
</tr>
</tbody>
</table>

The combination of /ka/ and /-husoku/ in (9a) can produce either voiceless variants as in /kahu/ or voiced variants as in /kahusoku/ ‘deficiency and excess.’ Similarly, the combination of /i/ and /-son/ in (9b) can produce either voiceless variants as in /i/ or voiced variants as in /izon/ ‘dependence.’ /kyoo+son/ in (9c) is realized either as /kyoo/ or /kyoo/ ‘coexistence.’ Finally, /ki+son/ in (9d) is realized either as /kison/ or /kizoon/ ‘existing.’ These compounds all satisfy the conditions for rendaku application. However, in dictionary entries, non-rendaku is the norm (Shimmura, 2008).

Before moving on, the use of these words as test items needs some justification.11 Firstly, the second member of these four compounds, /husoku/ and /son/, are Sino-Japanese words that do not undergo rendaku as stated in (2c): rendaku applies only to Yamato words. However, previous research has shown that non-Yamato compound words also commonly exhibit rendaku, though not to the extent that Yamato words do. The application of rendaku in non-Yamato compound words is particularly prevalent among frequently-occurring words, such as /kaisha/ ‘company,’ and /shasin/ ‘photograph’ (Kubozono, 1999; Sano, 2014; Takayama, 1999, 2005; Vance, 1987, 1996). For example, Vance’s (1996) dictionary-based investigation shows that Sino-Japanese words undergo rendaku around 10% of the time. Sano’s (2014) corpus-based study using the CSJ shows that Sino-Japanese words undergo rendaku 44% of the time in spontaneous speech. These studies also show that other loanwords than Sino-Japanese words do not undergo rendaku unless the loanword has become completely absorbed into Japanese, as in /karuta/ ‘playing card,’ and /kappa/ ‘raincoat.’

11 I would like to thank two anonymous reviewers for making me aware of this point.
Secondly, the voicing of /son/ varies depending on its etymological status, either Kan-on (voiceless variant) or Go-on (voiced variant). Then, the following question arises as to whether /son/ of three test items are underlyingly voiced. If /son/ is underlyingly voiced (Go-on), voiced renditions of the test items would not be due to rendaku. According to NHK Hoosoo Bunka Kenkyuujo (1998, 2005), however, /son/ in /ki+son/ is voiceless (Go-on), and /son/ in /i+son/ and /kyoo+son/ is also underlyingly voiceless (Go-on), but reflecting recent trends the voicing of /son/ is variably described as voiceless or voiced. As far as the etymology is consistent throughout the history of the individual lexical items, it is fair to say that /son/ in three test items are underlyingly voiceless, and the voicing of /son/ is the result of the application of rendaku.

For these reasons, Sino-Japanese words, such as /i+son/, /kyoo+son/, and /ki+son/ along with /ka+husoku/ where rendaku seems to be productively applied beyond normative pronunciations best fulfill the purpose of this study, although these may not be the best candidates for testing the productivity of rendaku in general.

Using these examples, I will first demonstrate the productivity of rendaku. Then, after the productivity of rendaku is confirmed, I show that the analogy-based model with lexical frequency can account for the issue of productivity.

4.1 Method

For the second examination, I used the CSJ. For the purpose of studying variability in actual production, speech corpora have been shown to be an excellent source of data. Although the CSJ is much smaller than the BCCWJ in size, it is nevertheless still sufficient for our current purpose, because unlike the first examination, this time only the four compounds presented in (9) are examined.

I retrieved the target items from the CSJ in the following manner: (i) I used YokkaGrep (http://www.yokkasoft.net/), a general-purpose editor; (ii) All sub-corpora of the CSJ were targeted; (iii) I only employed the phonetic information, because we focus only on four compounds in (7), making it unnecessary to refer to the detailed morphological conditioning. As in the previous examination, the search results that the YokkaGrep returned include the number of retrieved examples, and this number corresponds to the lexical frequency.

For each of the four items in (9), I examined the lexical frequency of rendaku forms and non-rendaku forms. The hypothesis to be addressed is as follows. If these items follow the item-specific lexical information embodied in the norm described in dictionaries, then for each of the compounds, the voiceless variants should be more frequent than the voiced variants. On the other hand, if the lexical information is not correctly reflected in the language use, then the productivity of rendaku is supported.

4.2 Result

An exhaustive search of the CSJ resulted in 722 tokens. The distribution of the tokens is shown in Table 2.

<table>
<thead>
<tr>
<th>dictionary entry</th>
<th>ka+husoku ‘deficiency and excess’ (N=5)</th>
<th>i+son ‘dependence’ (N=538)</th>
<th>kyoo+son ‘coexistence’ (N=54)</th>
<th>ki+son ‘existing’ (N=125)</th>
</tr>
</thead>
<tbody>
<tr>
<td>spontaneous speech</td>
<td>voiced / voiceless (2/3)</td>
<td>voiced / voiceless (531/7)</td>
<td>voiced / voiceless (50/4)</td>
<td>voiced / voiceless (94/31)</td>
</tr>
</tbody>
</table>

This concern was raised by two anonymous reviewers.

Throughout this article, “lexical information” simply means information assigned to each lexical item, in contrast to the productive phonological process. Discussion on how the information in dictionaries (or the norm) and native speakers’ lexical knowledge are related with one another is beyond the scope of this article.
As for /ka+husoku/, voiced variants (N=2) and voiceless variants (N=3) show almost the same frequency. As for /i+son/, however, voiced variants (N=531) are predominant. Similarly, for /kyoo+son/, and /ki+son/, the voiced variants (N=50, 94, respectively) are predominant. Thus, voiced variants were predominant in three out of four cases in spontaneous speech. All in all, each of the compounds is more likely to be realized as voiced variants with rendaku, despite the specifications in the dictionary to the contrary. The result suggests that rendaku is productively applied beyond the lexical information. With the productivity of rendaku being confirmed, I will in turn show that an analogy-based model that incorporates lexical frequency can account for the issue of productivity.

4.3 Discussion

According to the dictionary entries, no-rendaku is the norm. In other words, compounds such as /ka+husoku/, /i+son/, /kyoo+son/, and /ki+son/ should be realized as voiceless variants, as in /kahusoku/, /ison/, /kyooson/, and /kison/, even though they satisfy the conditions for rendaku application. This suggests that for some compounds rendaku applicability is learned by Japanese speakers as a lexical property. If rendaku applicability is correctly learned as the lexical property, voiceless variants should be predominant. The result of the second examination, however, shows that voiced variants were predominant. This suggests that rendaku is productively applied beyond the lexical information, in support of the psychological reality of rendaku.

With the productivity of rendaku being confirmed, the next step is to show that the analogy-based model with lexical frequency can account for the productivity. Following the method in the first examination, I explored the CSJ to examine the frequency of voiced variant with rendaku and voiceless variant without rendaku, focusing on all compounds that involve either /-husoku/ or /-son/ as the second member. The CSJ provided the following distribution.

<table>
<thead>
<tr>
<th></th>
<th>X+husoku</th>
<th>X+son</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiced</td>
<td>-busoku (N=92)</td>
<td>-zon (N=757)</td>
</tr>
<tr>
<td>voiceless</td>
<td>-husoku (N=3)</td>
<td>-son (N=57)</td>
</tr>
</tbody>
</table>

As Table 3 shows, as for /X+husoku/, the voiced variant /-busoku/ (N=92) is more frequent than the voiceless variant /-husoku/ (N=3), and as for /X+son/, the voiced variant /-zon/ (N=757) is more frequent than the voiceless variant /-son/ (N=57). Thus, for each item, the second member is predominantly realized as voiced variants. The test items /i+son/, /kyoo+son/, and /ki+son/ also predominantly undergo rendaku, except for /ka+husoku/. The voicing value of test items is mostly consistent with the more frequent variant. This suggests that even the cases, where rendaku is productively applied beyond the lexical information, can be accounted for by the analogy-based model with lexical frequency. These compounds follow the existing patterns that the exemplar offers.

The accumulated examples are the exemplars that offer the patterns to follow. Native speakers apply these kinds of patterns especially to words with low frequency, and accordingly they can correctly judge the well-formedness of novel compounds, and correctly apply rendaku to novel compounds. In other words, what appears to be productivity is a manifestation of the role of exemplars.

The result further offers an implication to the influence of frequency. If the lexical information about the applicability of rendaku is listed in dictionary entries, then such information would have a stronger mental representation and be more accessible in the choice of rendaku application, and hence prominent enough to be a referent. In reality, however, reference to the lexical information is cancelled, and the rendaku (non-) application is determined with reference to the frequency-based exemplar. This suggests that the influence of frequency overrides the lexical properties of each item.
5 Conclusion

This paper examined the effect of lexical frequency on the applicability of rendaku, and developed an analogy-based model by incorporating lexical frequency. In particular, based on the assumption of exemplar-theoretic models (Bybee, 2001), I proposed that low-frequency words (including novel compounds) rely on existing exemplar patterns. That is, compounds with low frequency are not prominent, and these examples cannot be a referent. Therefore, the rendaku applicability of such compounds is determined with reference to the existing frequency of rendaku or non-rendaku variants, and they follow the pattern offered by the more frequent variants. The analogy-based model with lexical frequency emphasizes that it is not the type frequency, or a single existing compound, but the token frequency of variants of a certain lexical item as the second member in every compound that becomes a referent or an exemplar offering patterns for compounds with low frequency. The current model cannot account for every aspect of rendaku, but it can deal with at least the (exceptional) behavior of compounds with low frequency such as novel compounds.

I demonstrated the productivity of rendaku as evidence for the psychological reality of rendaku. Furthermore, I also showed that rendaku is productively applied to innovative forms, and such an application can be accounted for by the current model. In such cases, the accumulated examples provide the patterns that the innovative forms follow. In other words, what appears to be the productivity is a manifestation of the role of exemplars.

Of course, the model proposed in this paper is not free from shortcomings. The number of the examples subjected to the examinations is small, and thus the model requires further validation. As mentioned above, the analogy-based model with lexical frequency does not account for every aspect of rendaku. Specifically, the following issues remain open questions. The first issue is whether or not the rendaku (non-)application of high-frequency compounds is subject to the same process, i.e., it is determined with reference to the (non-)application of the most frequent variant across every compound that involves the lexical item in question as the second member. The second issue is whether or not the role of the semantic and/or phonological similarities are excluded from the model. In the current model, the reference is established between compounds with low frequency/novel compounds and variants of a certain lexical item as the second member in every compound, but not a single existing compound. Thus, the first member is irrelevant. The semantic and/or phonological similarities differ according to the first member, as in /kuro+kami/ ‘black hair’ is more similar to /siro+kami/ ‘white hair,’ than to /nihon+kami/ ‘Japanese hair.’ Without the differences in the first member, the similarities do not play any role. These issues are beyond the scope of this study, and remain open questions, but should be addressed in future research.

There is also no denying that a number of studies have confirmed the status of rendaku as a phonological rule. Furthermore, there are some attempts to incorporate frequency effects into rule-based grammatical models (Coetzee, 2009, Coetzee & Kawahara, 2013). These studies are successful in showing, for example, that the English t/d-deletion rule is more likely to apply to more frequent items (see Coetzee and Kawahara, 2013 and references cited therein). However, the fact remains that rendaku has both lexical and phonological (grammatical) aspects (Kawahara, 2015; Vance, 2014). The present study does not offer any definitive answers to either the lexical or the grammatical approaches to rendaku. This also needs to be addressed in the future.

Finally, the results of the examinations emphasize the importance of token frequency and of actual linguistic data in linguistic research. I conclude by noting that the findings of this corpus-based study could not have been obtained by more traditional (i.e., intuition-based and dictionary-based) studies.

References


Vance (2014) admits the predictability of rendaku, but is skeptical about viewing the predictability as a rule.


Vance, Timothy J. (2014) If rendaku isn’t a rule, what in the world is it? In Kaori Kabata, & Tsuyoshi Ono (Eds.), Usage-Based Approaches to Japanese Grammar: Towards the Understanding of Human Language (pp. 137-152). Amsterdam: John Benjamins.