A Corpus-Based Study of Phonological Variation: The Domain of OCP and Morphological Boundaries

Shin-ichiro Sano

1. Introduction

By studying the *Corpus of Spontaneous Japanese* (National Institute for Japanese Language and Linguistics, 2008, 2012, hereafter CSJ), this project offers new findings on the patterns of phonological processes in actual usage. In particular, this study examines the effect of the Obligatory Contour Principle (OCP) (voice), which bans multiple occurrences of [+voice] in proximity to each other, also known as Lyman's Law (Itô & Mester, 1986), on the patterns of two phonological processes in Japanese: 1) devoicing of voiced obstruent geminates (geminate devoicing) and 2) rendaku. This study also examined the application domain of OCP (voice) with reference to the presence/absence of a morphological boundary. Consequently, the effect of OCP (voice) on the application of geminate devoicing and rendaku was confirmed. However, its effect was suppressed by the presence of a morphological boundary between the trigger/blocker and the target, suggesting that OCP (voice) is stem/morpheme-bound (Itô & Mester, 1986, 2003).

2. Background

This section provides an overview of the background information that characterizes this project. Section 2.1 outlines two phonological processes: geminate devoicing and rendaku. Section 2.2 introduces OCP (voice), and its effects on the application of geminate devoicing and rendaku. Section 2.3 presents the issues to be addressed and identifies the goals of this study.

2.1. Phonological processes
2.1.1. Geminate devoicing

Geminate devoicing is the process where voiced obstruent geminates (hereafter voiced geminates) in Japanese loanwords, such as [bb], [dd], and [gg] devoice, resulting in [ff], [tt], and [kk] (Kuroda, 1965; Itô & Mester, 1986, 2003; Nishimura, 2003; Kawahara, 2006). When loanwords with word-final voiced obstruents are borrowed in Japanese, the original voiced singletons are realized as voiced geminates followed by an epenthetic vowel (e.g., 'dog' => /doggu/). However, voiced geminates are traditionally prohibited in Japanese phonology. Therefore, these segments undergo some phonological processes to fit into the well-formed segmental configurations in Japanese. The most common phonological process involved is devoicing, as exemplified in (1).

(1) a. /beddo/ => [betto] 'bed'
    b. /baggu/ => [bakku] 'bag'
    c. /doggu/ => [dokku] 'dog'

[*] Shin-ichiro Sano, Keio University, shinichirosano@gmail.com. I would like to thank Shigeto Kawahara, Yu Tanaka, the members of the Keio phonetics-phonology study group and the participants of WCCFL 34 for their invaluable comments. I am also grateful to the members of the linguistics department at the University of Utah for their help organizing the conference. This project was supported by JSPS KAKENHI Grant #16K16831, and #26284059. Any remaining faults are, of course, my own.

 Proceedings of the 34th West Coast Conference on Formal Linguistics, ed. Aaron Kaplan et al., 439-446. Somerville, MA: Cascadilla Proceedings Project.

As with other repair strategies to avoid voiced geminates, geminate devoicing is optionally applied. This produces the variable phenomenon that consists of voiced geminates and voiceless geminates, as in [beddo]~[betto] and [doggu]~[dokku]. The identification of factors affecting the choice of voiced/voiceless geminates has long been the main objective of phonological research in Japanese (Itô & Mester, 1986, 2003; Nishimura, 2003; Kawahara, 2006, 2011a, 2011b, 2012; Kawahara & Sano, 2013; Sano, 2013; Sano & Kawahara, 2013a, 2013b).

2.1.2. Rendaku

Rendaku (i.e., sequential voicing), one of the best known and most studied morphophonological processes in Japanese, refers to the voicing of initial voiceless obstruents of the second member of morphologically derived words (most likely compounds) (Vance, 1979, 1980). As exemplified in (2a), when the two nouns /hosi/ and /sora/ are morphologically concatenated, the initial consonant /s/ of the second noun becomes voiced, producing /hosizora/. Similarly, rendaku makes /t/ of /tako/ in (2b) and /k/ of /kutu/ in (2c) voiced, resulting in [oodako] and [undoogutu].

(2)	a.	/hosi/	'star'	+ /sora/	'sky'	=>	[hosizora]	'starry sky'
	b.	/oo/	'big'	+ /tako/	'octopus'	=>	[oodako]	'big octopus'
	c.	/undoo/	'exercise'	+ /kutu/	'shoe'	=>	[undoogutu]	'sneaker'

Not all compounds, however, undergo this voicing process; namely, rendaku shows variability. Previous research has identified various kinds of factors that affect the likelihood of rendaku application (Vance, 1979, 1980; Itô & Mester, 1986, 2003; Sano, 2014; Kawahara & Sano, 2014).

2.2. OCP(voice)

As mentioned above, not all loanwords/compounds undergo geminate devoicing/rendaku; that is, these processes show variability. Prior work has identified various kinds of lexical, phonological, and morphological factors that affect the applicability of the processes. The representative factor that affects the applicability of two phonological processes is the OCP (voice) (i.e., Lyman's Law) that has long been subject to phonological studies in Japanese. In short, OCP (voice) has dissimilatory effects specific to the voice feature; namely, OCP (voice) bans multiple occurrences of [+voice] in proximity to each other. The patterns of geminate devoicing due to the OCP (voice) effects are as follows.

(3)	a.	/**d**oggu/	=>	[**d**okku]		'dog'
	b.	/eggu/	=>	[eggu]	(*[ekku])	'egg'
	c.	/**g**ibu/	=>	[**g**ibu]	(*[**g**ipu])	'give'

If voiced geminates co-occur with other voiced obstruent(s), devoicing is triggered, as /**d**oggu/ in (3a) becomes [**d**okku]. If there is no other voiced obstruent, the OCP does not have an effect, and devoicing is not applied, as /eggu/ in (3b) remains intact without undergoing devoicing. Voiced singletons never devoice, even if there is another voiced obstruent, as in /gibu/ in (3c) is faithfully realized as [gibu]. Thus, neither voicing nor geminacy alone causes devoicing, but their combination can trigger devoicing (Nishimura, 2003). Similarly, OCP (voice) also affects the patterns of rendaku.

(4)	a.	/hosi/	'star'	+ /ku**z**u/	'dust'	=>	[hosiku**z**u] (*[hosigu**z**u])	'stardust'
	b.	/oo/	'big'	+ /toka**g**e/	'octopus'	=>	[ootoka**g**e] (*[oodoka**g**e])	'big lizard'

In short, rendaku is blocked when the second noun of a compound already contains a voiced obstruent. As exemplified in (4a) and (4b), /hosi/ + /ku**z**u/, as well as /oo/ + /toka**g**e/, do not become *[hosigu**z**u] and *[oodoka**g**e] and remain intact, because the second nouns already contain the voiced obstruents /z/ or /g/. This structure is subject to OCP (voice), and accordingly, rendaku that otherwise produce OCP (voice) violations are blocked (Vance, 1979; Itô & Mester, 1986).

2.3. Problems and goals

Previous studies on geminate devoicing and rendaku have been conducted from various perspectives, including formal approaches such as Optimality Theory (Itô & Mester, 2003) and experimental approaches. These studies have made contributions to the development of phonological theory and to many theoretical debates, and the aspects of perception are still being researched. However, the aspects of actual production, including the effects of OCP (voice) and its application domain, have yet to be explored. With this background, this study focused on the effect of OCP (voice) in spontaneous speech, and examined the domain of its application.

3. Method

This section explains the method adopted by this study. Section 3.1 introduces the corpus. Section 3.2 describes the data collection procedure. Section 3.3 presents the criteria for data filtering. Section 3.4 summarizes the dataset subjected to the analysis.

3.1. Corpus

The corpus employed is the *Corpus of Spontaneous Japanese* (henceforth CSJ, National Institute for Japanese Language and Linguistics, 2008). The CSJ is a large-scale spontaneous speech corpus of Japanese with rich annotations. In terms of its size, the CSJ consists of 3,302 speech samples, amounting to 662 hours of speech, and 7.5 million words. The richly annotated subcomponent of the CSJ is called the "Core," and consists of 201 speech samples, amounting to 45 hours of speech. The data in the Core is provided as the Corpus of Spontaneous Japanese – Relational Database (CSJ-RDB) (National Institute for Japanese Language and Linguistics, 2012), where various kinds of annotations, such as phonetic/phonological information and morphological information at different linguistic levels are linked together. This feature is especially important for the study of rendaku, because in order to retrieve rendaku data by batch processing, at minimum the following information must be specified: the segment (voiced/voiceless obstruents) in the phonetic transcription files, the position in each word (the initial position), and the position in each noun-noun sequence (the second noun) in the morphological transcription files.

Among its properties, the CSJ mainly consists of two types of speech samples: Academic Presentation Speech, which is stylistically formal, and Simulated Public Speaking, which is stylistically informal.

3.2. Data collection

For geminate devoicing, all subcorpora of the CSJ were targeted, as geminate devoicing is observed only in loanwords and their frequency is relatively low. I retrieved all potential voiced geminates in the CSJ. The segmentation is based on the annotated syllabic transcription called *Hatsuonkei,* which represents the surface form. If devoicing applies, the underlying form and the surface form are presented separately as in "バック; バッグ" (SF; UF) 'bag'; while if it does not apply, the underlying form and the surface form are represented by a single notation as in "バッグ."[1]

For rendaku, all subcorpora of the Core (CSJ-RDB) were examined. The data collection was conducted using Navicat Premium, software that implements the SQL programming language (http://jp.navicat.com/) with reference to the phonological and morphological information annotated in the CSJ-RDB. This study focused on compound nouns that consist of two or more nouns, where the initial consonant of the second noun is underlyingly voiceless. From the CSJ-RDB, I extracted tri-grams that included potential rendaku undergoers (the second noun of compounds) with the following structure: $[_{\text{compound}}$ 1st Noun + 2nd Noun] + following context. The first nouns correspond to the preceding contexts, and the second nouns to rendaku-undergoers. The instruction is given to the

[1] In the Japanese syllabary, the letter "ッ" represents gemination, and the diacritic "゛" represents the voicing of obstruents. These two were included in the data-mining search strings.

program by using the search formula with regular expressions. The segmentation is based on the alphabetic transcription of surface forms in the Core. If rendaku applies, the initial consonant of the second noun is voiced; otherwise, it is voiceless.

3.3. Filtering

The batch processing cannot exclude some tokens that are outside of our scope of investigation. Therefore, the retrieved data was subjected to filtering, and tokens that did not match the purpose of this study were excluded from the dataset. For geminate devoicing, tokens were excluded from the dataset 1) if there were phonological processes involved other than devoicing (e.g., degemination or deletion) and 2) if targeted segments were a part of filled pauses or word fragments.

Rendaku tokens were excluded from the dataset if 1) the initial consonants of the second nouns (rendaku-undergoers) were segments other than [h, t, k, s] or [b, d, g, z], 2) the initial consonants of the second nouns were underlyingly voiced, 3) there was a word boundary between the first noun and the second noun (as these are not compounds), and 4) rendaku was categorically blocked due to lexical restrictions.[2]

3.4. Dataset

After filtering the tokens from the CSJ, each token was classified according to the application/non-application of the processes. This primary classification constitutes the basis of the following analysis, and ratio of geminate devoicing/rendaku is calculated based on this variable. The overall distribution is summarized in Table 1 and Table 2.

Table 1: The distribution of (non-)geminate devoicing in the CSJ.

	Frequency	Ratio
+Devoicing	464	0.29
–Devoicing (faithful)	1,153	0.71
Total	1,617	1

Table 2: The distribution of (non-)rendaku in the CSJ (Core).

	Frequency	Ratio
+Rendaku	709	0.59
–Rendaku (faithful)	484	0.41
Total	1,193	1

An exhaustive search of the data in the CSJ resulted in 1) a total of 1,617 tokens of potential candidates of geminate devoicing, of which 464 (29%) showed devoicing, and 2) a total of 1,193 tokens of potential candidates of rendaku, of which 709 (59.4%) showed rendaku.

Each token was analyzed in terms of OCP (voice) and the presence/absence of a morphological boundary between the trigger/blocker and the target in preceding/following contexts. The results were analyzed by the logistic linear mixed model (Jaeger, 2008), using the `glmer` function in R (R Development Core Team, 2016), where (i) random factors were speakers and items and (ii) both random slope and random intercepts were included in the model to have maximal random effects structure (Barr, 2013; Barr et al., 2013).

4. Analysis

This section presents an overview of examinations of OCP (voice) and its application domain. Section 4.1 presents the analysis of the data of geminate devoicing. Section 4.2 presents the analysis of

---

[2] Unlike geminate devoicing, rendaku is subject to various kinds of lexical restrictions , and accordingly, in some contexts the application of rendaku is categorically blocked.

rendaku data. Each of the results serves as supporting evidence that the domain of OCP (voice) is stem/morpheme-bound.

4.1. Geminate devoicing

4.1.1. Presence/absence of trigger(s) and morphological boundaries

The previous literature reports that geminate devoicing is more likely to be triggered when a voiced obstruent other than a voiced geminate is present (Nishimura, 2003; Kawahara, 2012). If OCP (voice) is active, the probability of devoicing would be higher with the presence of other voiced obstruent(s) than with the absence thereof. The effect can be tested by comparing the ratio of devoicing with or without other voiced obstruent(s).

Figure 1 illustrates the likelihood of geminate devoicing for the presence of a voiced obstruent (e.g., /**d**oggu/) and for the absence of a voiced obstruent (e.g., /eggu/). We observed that the likelihood of devoicing was significantly higher with the presence of another voiced obstruent (presence: 39.4%, absence: 5%, z=6.468, *p*<0.01). Thus, we can argue that OCP (voice) is active. Moreover, the current production-based data replicates the findings of Nishimura (2003) and Kawahara (2006), that OCP is a crucial factor in inducing devoicing (Kawahara & Sano, 2013; Sano & Kawahara, 2013b).

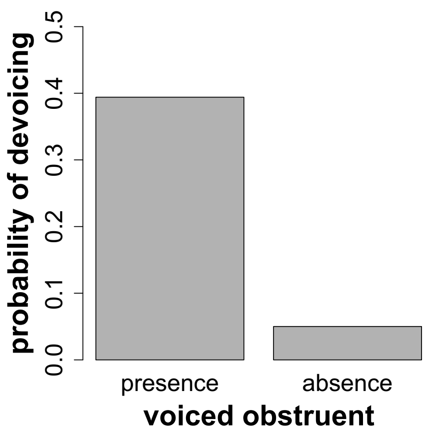

Figure 1: Probability of geminate devoicing by the presence/absence of a voiced obstruent.

With the effect of OCP (voice) being confirmed, the next step was to examine its effect in connection with the presence/absence of a morphological boundary between a trigger and a target. If a morphological boundary functions as a blocker of OCP (voice) effect, then the likelihood of devoicing would be higher with the absence of a morphological boundary than with the presence thereof. The effect can be tested by comparing the ratio with or without a morphological boundary.

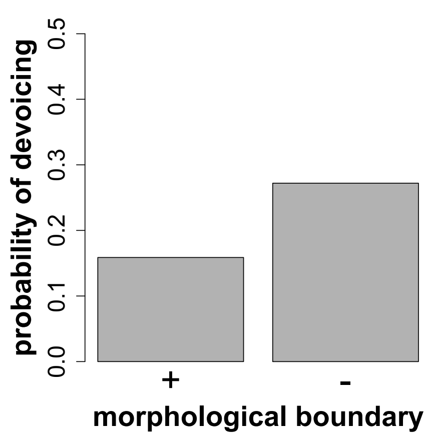

Figure 2: Probability of geminate devoicing by the presence/absence of a morphological boundary.

Figure 2 illustrates the likelihood of geminate devoicing for the presence of a morphological boundary and for the absence thereof. We observed that the probability of devoicing was lower when there was a morphological boundary between a trigger and a target (+boundary: 15.9%, –boundary: 27.2%, z=2.288, $p<0.05$). This result suggests that the effect of OCP (voice), which triggers geminate devoicing, is suppressed by the presence of a morphological boundary. This further supports the claim that OCP is stem/morpheme-bound (Itô & Mester, 1986, 2003).

4.1.2. Asymmetry between preceding and following contexts

Next, let us turn our attention to the asymmetry between preceding and following contexts. As voiced geminates occur in word-final position, a morphological boundary intervenes between the target and the trigger in the following context, while this is not the case for the preceding context. In other words, if a voiced obstruent precedes the target, they are in the same stem, while if a voiced obstruent follows the target, the morphological boundary intervenes between them as shown in example (5).

(5) a. trigger in the preceding context /**d**oggu/ 'dog'
b. trigger in the following context /heddo-**g**a/ 'head-Nom'

Thus, by comparing the distribution of geminate devoicing in the preceding context and in the following context, we can examine the effect of morphological boundaries. Assuming that a morphological boundary is at work in canceling out the effect of OCP, which induces devoicing, we can hypothesize that in the preceding context where the target and the trigger are in the same stem without an intervening morphological boundary, the likelihood of devoicing would be higher when the trigger is present than when it is absent. While in the following context, where the trigger and the target are in distinct stems across the morphological boundary, the likelihood of devoicing would be consistent regardless of the presence/absence of the trigger.

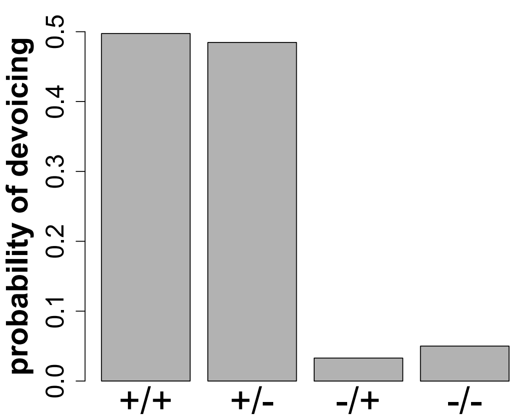

Figure 3: Probability of geminate devoicing by the presence/absence of a voiced obstruent in the preceding and following contexts.

In Figure 3, "+" represents the presence of the trigger, and "–" represents the absence thereof; for the preceding context, the presence/absence of the target is specified before the slash, while it is specified after the slash for the following context. Thus, the leftmost bar (+/+) represents "preceding presence and following absence," the second leftmost bar (+/–) represents "preceding presence and following absence," the second rightmost bar (–/+) represents "preceding absence and following presence," and the rightmost bar (–/–) represents "preceding absence and following absence."

We observed that the likelihood of devoicing significantly differs according to the presence/absence of the trigger in the preceding context (z=2.916, $p<0.01$): the presence of a voiced obstruent in the preceding context increases the devoicing likelihood (+/+ and +/–), while the absence lowers the likelihood (–/+ and –/–). Thus, we confirmed that OCP (voice) is active as long as no morphological boundary intervenes between the trigger and the target.

For the following context, however, we observed that the presence/absence of the trigger did not make any significant difference in the likelihood of devoicing (z=1.027, n.s.): there was no significant difference both in the two leftmost bars (+/+ and +/–), and in the two rightmost bars (–/+ and –/–). Furthermore, there was no significant interaction between the preceding context and the following context (z=0.316, n.s.), showing that the distributional difference is solely due to the preceding context. This suggests that the presence of a morphological boundary suppresses the effect of OCP (voice), which otherwise induces geminate devoicing. These results support the claim that the effect of OCP (voice) is stem/morpheme-bound (Itô & Mester, 1986, 2003).[3]

4.2. Rendaku

The final piece of evidence that demonstrates the domain of OCP (voice) was obtained by analyzing rendaku data. As described in Section 2.2, OCP (voice) plays a role in blocking rendaku: rendaku does not apply if there is another voiced obstruent. In rendaku, a morphological boundary intervenes between the first noun and the second noun. As exemplified in (6), if the blocker appears in the second noun, the blocker and the target are both in the same stem, and no morphological boundary intervenes between them. However, if the blocker appears in the first noun, the blocker and the target are in distinct stems across the morphological boundary.

(6) a. blocker in the second noun /hosi + kuzu/ 'star + dust'
b. blocker in the first noun /kage + hosi/ 'shade + dry'

Thus, by comparing the distribution of rendaku in the first noun and in the second noun, we can examine the effect of morphological boundaries. If the morphological boundary suppresses the effect of OCP (voice), the presence of the blocker in the second noun would lower the likelihood of rendaku application, while the blocker in the first noun would produce no such distributional skew.

As previous research has reported that the presence of voiced obstruent(s) in the second noun blocks rendaku (Vance, 1979; Ito & Mester, 2003), no examples of exceptional cases were found (that is, rendaku still occurred in spite of the presence of a voiced obstruent in the second noun). OCP (voice) has thus far been found to be active in several experimental studies, though its blocking effect is not categorical (Vance, 1979; Kawahara, 2012). This result suggests that OCP (voice) categorically blocks rendaku, and that it has a stronger impact in actual production than in experimental settings.

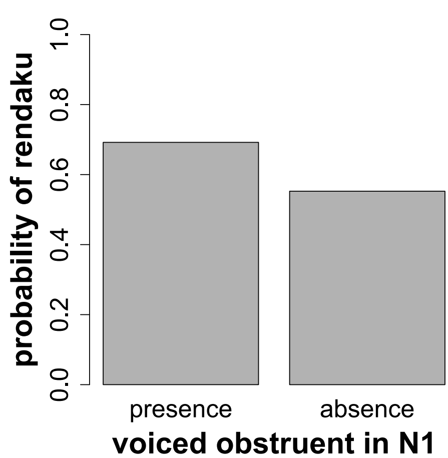

Figure 4: Probability of rendaku by the presence/absence of a voiced obstruent in the first noun.

[3] I extend my gratitude to Bruce Hayes and Rachel Walker for pointing out the directionality of OCP (voice) effect. This issue is worth pursuing, although the examination suggests that the apparent directionality in geminate devoicing can be reduced to the presence/absence of a morphological boundary.

On the other hand, if the blocker is in the first noun across the morpheme boundary, the presence/absence of the blocker does not produce a significant difference in the likelihood of rendaku (z=1.628, n.s.), as illustrated in Figure 4. Summarizing the results, we can argue that the presence of a morphological boundary suppresses the blocking effect of OCP (voice).

5. Conclusion

To summarize, the results of the examination of geminate devoicing and rendaku showed that 1) the likelihood of the application of the processes is higher/lower when there is a trigger/blocker (voiced obstruent), showing that OCP (voice) is active; and 2) the triggering/blocking effect of OCP (voice) is suppressed by the presence of a morphological boundary. Thus, this study confirmed the effect of OCP (voice) and its domain: the effect of OCP (voice) is stem/morpheme-bound. Furthermore, the corpus-based study revealed the gradient aspects of these effects, in contrast to intuition-based studies.

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

Itô, Junko & Armin Mester. (1986). The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry, 17*, 49–73.

Itô, Junko & Armin Mester. (2003). *Japanese morphophonemics.* Cambridge: MIT Press.

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.

Kawahara, Shigeto. (2006). A faithfulness ranking projected from a perceptibility scale: The case of voicing in Japanese. *Language, 82,* 536-574.

Kawahara, Shigeto. (2011a). Japanese loanword devoicing revisited: A rating study. *Natural language and Linguistic Theory, 29,* 705-723.

Kawahara, Shigeto. (2011b). Aspects of Japanese loanword devoicing. *Journal of East Asian Linguistics, 20*(2), 169–194.

Kawahara, Shigeto. (2012). Lyman's Law is active in loanwords and nonce words: Evidence from naturalness judgment studies. *Lingua, 122,* 1193-1206.

Kawahara, Shigeto & Shin-ichiro Sano. (2013). A corpus-based study of geminate devoicing in Japanese: Internal factors. *Language Sciences, 40*, 300-307.

Kawahara, Shigeto & Shin-ichiro Sano. (2014). Identity avoidance and Lyman's Law. *Lingua, 150*, 71-77.

National Institute for Japanese Language and Linguistics. (2008). The Corpus of Spontaneous Japanese, Ver. 2. Database.

National Institute for Japanese Language and Linguistics. (2012). CSJ-RDB. Database.

Kuroda, Shige-Yuki. (1965). *Generative grammatical studies in the Japanese language*. Doctoral dissertation, MIT. Retrieved from <http://hdl.handle.net/1721.1/13006>

Nishimura, Kohei. (2003). *Lyman's Law in loanwords*. MA Thesis, Nagoya University.

R Development Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sano, Shin-ichiro. (2013). Patterns in avoidance of marked segmental configurations in Japanese loanword phonology. In: *Proceedings of GLOW in Asia 2012: Main Session*, 245-260.

Sano, Shin-ichiro. (2014). Examining lexical and phonological factors on rendaku in spontaneous speech. *MIT Working Papers in Linguistics, 73*, 179-190.

Sano, Shin-ichiro & Shigeto Kawahara. (2013a). A corpus-based study of geminate devoicing in Japanese. RUMMIT 2013, at University of Massachusetts, Amherst, Massachusetts, USA, April 2013.

Sano, Shin-ichiro & Shigeto Kawahara. (2013b). A corpus-based study of geminate devoicing in Japanese: The role of the OCP and external factors. *Gengo Kenkyû, 144*, 103-118.

Vance, Timothy. (1979) *Nonsense word experiments in phonology and their application to rendaku in Japanese*. Doctoral dissertation, University of Chicago.

Vance, Timothy. (1980) The psychological status of a constraint on Japanese consonant alternation. *Linguistics, 18,* 245-267.

Proceedings of the 34th West Coast Conference on Formal Linguistics

edited by Aaron Kaplan, Abby Kaplan, Miranda K. McCarvel, and Edward J. Rubin

Cascadilla Proceedings Project Somerville, MA 2017

Copyright information

ISBN 978-1-57473-471-3 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Sano, Shin-ichiro. 2017. A Corpus-Based Study of Phonological Variation: The Domain of OCP and Morphological Boundaries. In *Proceedings of the 34th West Coast Conference on Formal Linguistics*, ed. Aaron Kaplan et al., 439-446. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3350.