Phonotactically-Driven Rendaku in Surnames: A Linguistic Study Using Social Media

Yu Tanaka

1. Introduction

Like regular compound words, Japanese surnames with a compound structure can undergo a voicing alternation known as *rendaku*. Rendaku application in surnames is somewhat different from that in regular compounds, however; the voicing alternation is governed by avoidance of consonant sequences that are similar in terms of voicing and place. Although the peculiarity of rendaku in surnames has been noted in the literature (see Sugito, 1965 among others), no study has provided a full description or explanation of the patterns. The first aim of this study is thus to better describe the data. A corpus of rendaku in surnames is created using social media, which enables us to assess comprehensively and quantitatively the validity of proposed generalizations. The second aim of the study is to explain why rendaku applies in a different manner in names and in normal words. I propose that compound names are treated as if they are single stems and that the peculiar rendaku patterns in surnames can be attributed to stem-internal phonotactics, such as dispreferences for the occurrence of multiple voiced obstruents or sequential homorganic voiceless obstruents within stems.

2. Data: Rendaku in surnames

2.1. Strong Lyman's Law?

Rendaku, or sequential voicing (Martin, 1952), is a morphophonological process in Japanese whereby the initial obstruent of the second element of a compound becomes voiced,[1] as shown in (1).[2]

(1)	a.	maki	+	susi	→	maki-**z**usi	'rolled-sushi'
	b.	ori	+	kami	→	ori-**g**ami	'folding-paper'

Application of rendaku is variable, being affected by a number of factors including lexical idiosyncrasy (see Vance, 2015 for a comprehensive overview). However, there is one phonological factor that categorically blocks it; the voicing alternation does not occur if the second element (henceforth E2) contains a voiced obstruent, as shown in (2). This is commonly known as Lyman's Law (Lyman, 1894).

(2)	a.	yama	+	kaze	→	yama-**k**aze	*yama-**g**aze	'mountain-wind'
	b.	oo	+	tokage	→	oo-**t**okage	*oo-**d**okage	'big-lizard'

The law can be viewed as a kind of Obligatory Contour Principle effects (Ito & Mester, 1986); it disallows the occurrence of multiple voiced obstruents in the domain of a single stem.[3] Rendaku is thus blocked when it would create a second voiced obstruent in E2.

---

* Yu Tanaka, University of California, Los Angeles, yutanaka@ucla.edu. I thank Adam Chong, Robert Daland, Eleanor Glewwe, Bruce Hayes, Junko Ito, Shigeto Kawahara, Kazutaka Kurisu, Michael Lefkowitz, Eric Rosen, Shin-ichiro Sano, Brian Smith, Megha Sundara, Kie Zuraw and the audiences at WCCFL 34, WAFL 12, FAJL 8, SCAMP and UCLA Phonology Seminar for their comments.

[1] Rendaku can also occur in morphologically-derived words where one of the elements is an affix. In words composed of more than two elements, a non-initial element other than the second may also undergo rendaku.

[2] Throughout the paper, transcriptions are phonemic and broad. Accent is not marked unless it is relevant.

[3] The statement is also true as a static generalization. Native Japanese stems do not usually contain multiple voiced obstruents except those that are onomatopoeic or those with pejorative connonations.

 Proceedings of the 34th West Coast Conference on Formal Linguistics, ed. Aaron Kaplan et al., 519-528. Somerville, MA: Cascadilla Proceedings Project.

Most Japanese surnames are compounds in terms of their morphological structure and they may undergo rendaku if E2 starts with a voiceless obstruent.[4] As in regular compounds, rendaku in surnames is variable. Certain names typically show the voicing change while others do not, and there are also some that show variation.[5] This is illustrated in (3) with surnames which have /ta/ 'rice field' as E2.

(3)	a.	yama	+	ta	→	yama-**d**a	山田	'mountain-rice field (surname)'
	b.	mura	+	ta	→	mura-**t**a	村田	'village-rice field (surname)'
	c.	naka	+	ta	→	naka-**t**a ∼ naka-**d**a	中田	'center-rice field (surname)'

It is worth noting here that rendaku is usually not reflected in orthography. On most occasions, surnames are written in Chinese characters or *kanji*, with each character representing one morpheme. In (3), the second element is always written with the same kanji 田 whether it is realized as [ta] or [da].

Sugito (1965) reveals that rendaku in surnames with /ta/ as E2 is largely predictable, showing that it is conditioned by the last consonant of the first element (E1). For example, if E1's last consonant is /s/, /t/, /k/, /m/ or /n/, rendaku is commonly observed, as in (4a). On the other hand, if the consonant is a voiced obstruent such as /b/, /d/, /g/ or /z/, rendaku never applies, as in (4b).[6]

(4)	a.	asa-**d**a	matu-**d**a	taka-**d**a	yama-**d**a	sana-**d**a
	b.	siba-**t**a	kado-**t**a	naga-**t**a	kazi-**t**a	

This blocking of rendaku by a voiced obstruent in E1 may, at first glance, seem to be the application of Lyman's Law. Note, however, that the law's effect is normally stem-bounded. Rendaku is blocked if there is /b/, /d/, /g/ or /z/ in E2 since voicing its initial consonant would create two voiced obstruents within one stem. Voiced obstruents in E1, on the other hand, do not affect rendaku application across a morpheme boundary in regular compounds (e.g. /kabe/ + /kami/ → kabe-**g**ami 'wall-paper'). Why, then, do surnames show the rendaku patterns illustrated in (4)?

Diachronically, the domain of Lyman's Law was the prosodic word (see Unger, 1977 among others). In Old Japanese, rendaku is blocked if either E1 or E2 contains a voiced obstruent (e.g. /mîdu/ + /töri/ → mîdu-**t**öri, *mîdu-**d**öri 'water-bird'; Vance, 2005; the transcription is his). This ban on an occurrence of multiple voiced obstruents within the whole compound is refered to as the strong version of Lyman's Law (Unger, 1977) or simply Strong Lyman's Law. The domain of the law has changed over time, and the effect of Strong Lyman's Law has become weaker in later stages of the language (see Endo, 1980; Sakurai, 1972; Toda, 1988). It has been proposed that the word-bounded law still has a gradient effect in present-day Japanese (see e.g. Kindaichi et al., 1988:264; Sato, 1989). Recent studies, however, show that such a tendency cannot be found in large corpora of rendaku in regular compounds (Irwin, 2014; Sano, 2015) and that its psychological reality is also questionable (Kawahara & Sano, 2014b).

Does this simply mean that surnames, which appear to be subject to Strong Lyman's Law, have retained old sound patterns? The diachronic explanation does not seem satisfactory for several reasons. First, Japanese surnames are relatively new. Many ordinary people did not have official surnames until the late 19th century when the Surname Law, which required all citizens to have a surname, was enacted (The Meiji Government, 1875). Since then, the number of surnames has increased from about 30,000 to 120,000 or more (Yasuda, 1983). That is to say, not all surnames existed when Old Japanese was spoken (before the late 8th century; Miyake, 2003). Second, the reading of a surname could be changed relatively easily until recently. We find records (official and unofficial) stating that some people changed their surname from a rendaku form to a non-rendaku form or vice versa.[7] All these facts show that the effect of Strong Lyman's Law found in surnames should not simply be attributed to history.

[4] About 96% of the 10,019 most common surnames are compound names (Shirooka & Murayama, 2011).

[5] "Variation" here means that, among people who have a surname composed of the same morphemes (e.g. /naka-ta/ 中田), some have a non-rendaku pronunciation ([nakata]) while others have a rendaku pronunciation ([nakada]).

[6] Sugito (1965) also includes the glide /y/ in the set of consonants that block rendaku. Her survey is only concerned with surnames with an E1 that is bimoraic. If E1 is longer than two moras, rendaku is generally more likely to apply regardless of the voicing of E1's last consonant (Zamma, 2005); e.g. /yanagi-ta/ [yanagida] ∼ [yanagita].

[7] The readings of surnames could be changed easily because the Japanese household register known as *koseki*, which is a type of civil registry for each family unit, only uses the kanji script to represent registered people's surnames and does not provide their readings. For example, the founder of the music instrument company Yamaha Corporation, Torakusu Yamaha (1851-1916), changed his surname from Yamaba to Yamaha (Miura, 2012).

2.2. Peculiarity of /k/

Rendaku in surnames shows further complications, one of which is the peculiarity of /k/. As we saw in (4), a surname with /ta/ often undergoes rendaku if E1's last consonant is /s/, /t/ or /k/. Kubozono (2005) and Zamma (2005) point out that, among the voiceless obstruents, it is actually only /k/ that allows non-rendaku forms. As shown in (5a), rendaku is common in surnames with /k/ in E1; however, there are also quite a few names showing no rendaku, as in (5b), or those showing variation, as in (5c).

(5) a. huka-**da** oka-**da** take-**da** ike-**da** taka-**da** toku-**da**
b. saka-**ta** oki-**ta** aki-**ta** iku-**ta** taki-**ta**
c. naka-**ta** ∼ naka-**da** huku-**da** ∼ huku-**ta**

This contrasts with surnames having /s/ or /t/ in E1. Table (6) shows the number of /ta/-surnames with and without rendaku ([**t**a] or [**d**a]) sorted by E1's last consonant (/s/, /t/ or /k/) in Sugito's (1965) data. As can be seen, /s/ and /t/ always trigger rendaku while /k/ allows both rendaku and non-rendaku forms. Kubozono (2005:10) states "[t]he reason for this peculiar behavior of /k/ remains unclear."

(6)

E1 cons.	**t**a	**d**a
/s/	0	35
/t/	0	26
/k/	13	31

To summarize, previous studies (Kubozono, 2005; Sugito, 1965; Zamma, 2005 among others) have shown that compound surnames exhibit peculiar rendaku patterns; the alternation is conditioned by the voicing and place of the last consonant of the first element. However, it remains to be answered why such patterns exist or why they are found only in surnames and not in regular compounds.

3. Proposal: Rendaku as stem-internal phonotactic restrictions

To explain the difference between rendaku patterns in surnames and those in regular compounds, I propose that compound surnames behave like single stems and that rendaku in surnames is driven by stem-internal phonotactics. It will be shown that the proposal gives a unified account of the effect of Strong Lyman's Law and the peculiarity of /k/ reported in the literature.

3.1. Morphological parsing of compound names

Although many Japanese surnames seem to have a compound structure, they differ from regular compounds in that their meanings are not compositional. For example, the surname /ta-naka/ 田中 is composed of /ta/ 'rice field' and /naka/ 'center'. Etymologically speaking, it presumably derived from a place where inhabitants were surrounded by paddies. The name itself, however, does not actually mean 'the center of rice fields'; rather, it simply refers to an individual who has the surname.

It is known that phonotactics, morphological parsing and semantics interact with each other. Phonotactics often serves as a cue to morpheme boundaries. Hay (2003) shows that if a morphologically complex word has phonotactically weak boundary cues, it is less decomposable and more likely to be stored as one unit in speakers' mental lexicons. This can in turn cause semantic drift, making the word acquire a non-compositional meaning.[8] I argue that the opposite is also true. If a compound has inherently no semantic compositionality as in the case of proper nouns, including surnames, it is parsed more as one unit or a stem. As a result, phonology imposes stem-internal phonotactic restrictions.

In order to formalize the idea, I propose that compound surnames and regular compounds have the following structures. As is shown in (7a), a compound surname is composed of two stems, but the outcome is also a stem which semantically functions as a single personal name. This contrasts with the structure of a regular compound shown in (7b). A regular compound is also composed of two stems, but the outcome is a compound of stems which does not form a stem itself.

[8] It also has a consequence in phonology; stored complex words often undergo phonetic reduction. Word frequency is also an important factor in her model. Frequent words are more likely to be stored as a unit.

(7) a. A compound surname: A stem composed of stems

$[\,A\,]_{stem} \;+\; [\,B\,]_{stem} \;\rightarrow\; [\,[\,A\,]_{stem}\text{ - }[\,B\,]_{stem}\,]_{stem}$

b. A regular compound: Two combined stems

$[\,A\,]_{stem} \;+\; [\,B\,]_{stem} \;\rightarrow\; [\,A\,]_{stem}\text{ - }[\,B\,]_{stem}$

These structural differences bring about consequences in phonology. As it is a stem itself, a compound surname is subject to stem-internal phonotactics, which affects rendaku application.

3.2. Rendaku blocker: Stem-bounded Lyman's Law

The proposed analysis explains why rendaku in surnames is blocked by a voiced obstruent in E1 unlike in regular compounds. It is not that the effect of Lyman's Law extends beyond the stem-level, as is implied by the oft-used term "Strong Lyman's Law". Compound surnames are treated as stems (with an internal compound structure). As a result, the normal stem-bounded law applies. Rendaku, which would create multiple voiced obstruents *within a stem*, is thus blocked, as exemplified in (8).

(8) [siba]$_{stem}$ + [ta]$_{stem}$ → [[siba]-[**t**a]]$_{stem}$ *[[siba]-[**d**a]]$_{stem}$

3.3. Rendaku trigger: Identity Avoidance

Stem-internal phonotactics can not only block rendaku but also trigger rendaku. Japanese has gradient root/stem consonant cooccurrence restrictions based on place and voicing identity, which will be referred to here as Identity Avoidance.[9] Using data compiled from a large dictionary, Kawahara et al. (2006) show that stems containing consonants of the same place in adjacent syllables are generally underrepresented in the native Japanese lexicon. Voicing also contributes to the restrictions. Stems with adjacent pairs of homorganic voiceless obstruents such as /s...s/, /t...t/ and /s...t/ are rarer compared to those with homorganic obstruents that disagree in voicing such as /s...z/, /t...d/ and /s...d/. If surnames are treated like stems, they should show the same kind of Identity Avoidance effects.

I argue that such dispreferences for sequences of homorganic voiceless obstruents within stems are indeed reflected in the way rendaku applies in surnames. Recall from Sugito's (1965) data in (6) that surnames with /ta/ as E2 always undergo rendaku if E1's last consonant is /s/ or /t/. Notice that combining a morpheme with /s/ or /t/ in its last mora (e.g. /asa/, /matu/) with /ta/ would create voiceless coronal obstruents in adjacent moras. Application of rendaku can dissimilate such *stem-internal* homorganic obstruent sequences in terms of voicing. In other words, rendaku is triggered by stem-internal phonotactics, specifically, Identity Avoidance.[10] This is illustrated in (9).

(9) a. [asa]$_{stem}$ + [ta]$_{stem}$ → [[asa]-[**d**a]]$_{stem}$ *[[asa]-[**t**a]]$_{stem}$

b. [matu]$_{stem}$ + [ta]$_{stem}$ → [[matu]-[**d**a]]$_{stem}$ *[[matu]-[**t**a]]$_{stem}$

Under this analysis, the behavior of E1-final /k/ in surnames with /ta/ discussed in previous studies (see Section 2.2) is no longer strange. Combining a morpheme with /k/ in its last mora with /ta/ does not result in a disfavored homorganic consonant sequence. Thus, rendaku applies less often in surnames having an underlying /k...t/ sequence than in surnames having an underlying /s...t/ or /t...t/ sequence.

Note that the proposed analysis makes new predictions about how rendaku should apply in surnames with E2 other than /ta/. If Identity Avoidance is at work, /k...k/ sequences should trigger rendaku as well as /s...t/ and /t...t/ sequences. That is, surnames having a /k/-initial morpheme such as /kawa/ 'river' as E2 are expected to undergo rendaku when E1's last consonant is also /k/. Impressionistically, we indeed find many examples of surnames with a /k...k/ sequence that commonly show rendaku (with some possible variation), as in (10a), and fewer examples that do not, as in (10b). Compare this with surnames with /s...k/ or /t...k/ sequences, which seem to prefer non-rendaku forms, as shown in (10c). In Section 4, I will show that these predictions are actually borne out in a larger scale corpus study.

(10) a. naka-**g**awa toku-**g**awa huka-**g**awa kake-**g**awa taki-**g**awa (∼ taki-**k**awa)

b. take-**k**awa seki-**k**awa yoko-**k**awa

c. isi-**k**awa hoso-**k**awa matu-**k**awa iti-**k**awa

[9] Avoidance of similar consonants is found in a number of languages. See Walter (2007) and references therein.

[10] See Kawahara & Sano (2014a) for experiments testing the role of Identity Avoidance as a rendaku trigger.

3.4. Evidence from prosody: Noun stem accent for names

Other evidence that compound surnames are treated like stems in phonology comes from their prosodic patterns. Japanese has a complex compound accent rule. Simply put, for compound nouns with a short E2 (one or two moras), accent generally falls near the boundary between E1 and E2. Thus, compound accent is either on the final syllable of E1 as in (11a) or on the initial syllable of E2 (with E2 preserving its original accent) as in (11b); otherwise, the compound is unaccented.[11]

(11)	a.	síro	+	mé	→	si**ró**-me	'white-eye'
	b.	migi	+	máe	→	migi-**má**e	'right-front'

Compound surnames, however, do not follow this general pattern (see Ito & Mester, 2016; Kawahara, 2015). For three-mora names with a monomoraic E2, accent falls on the syllable containing the antepenultimate mora of the entire name instead of on the final syllable of E1. This is shown in (12a). Similarly, even if E2 is a morpheme which usually preserves its original accent, the antepenultimate accent rule still applies, as in (12b).

(12)	a.	síro	+	tá	→	**sí**ro-ta	*si**ró**-ta	'white-rice field (surname)'
	b.	miya	+	máe	→	mi**yá**-mae	*miya-**má**e	'shrine-front (surname)'

I argue that, as proposed in (7), the whole compound surname is treated as a stem and thus receives antepenultimate accent, which is the default for noun stems (McCawley, 1968; also see Kawahara, 2015).

3.5. Summary of the proposal

To summarize, I have proposed that compound surnames are treated as stems and that rendaku in surnames is driven by stem-internal phonotactics; it is blocked by stem-bounded Lyman's Law and triggered by Identity Avoidance. The analysis captures why rendaku in surnames is conditioned by the voicing and place of E1's last consonant, unlike in regular compounds. The proposal is also compatible with the accent behaviors of compound surnames, which ressemble those of noun stems. In the following section, I discuss a corpus study whose results support the hypothesis.

4. A corpus study using social media

4.1. Aims of the study

In order to test the proposed hypothesis with large-scale data, I conduct a corpus study using social media. Previous studies on rendaku in surnames are, by and large, descriptively inadequate.[12] Their observations are mostly based on the intuitions of a limited number of speakers or only concerned with surnames with certain E2s such as /ta/. The most extensive study is Zamma (2005) who reports judgments on rendaku in surnames with 17 different E2s obtained from 5 speakers; however, the study does not report the details of rendaku application based on the place of E1's last consonant and E2's initial consonant, which would make it possible to address the effect of Identity Avoidance. Additionally, none of the studies attempts to describe the variability of rendaku in detail. As mentioned above, the surname /naka-ta/ can be pronounced either [nakata] or [nakada], but the non-rendaku form is more common than the rendaku form. On the other hand, /taki-kawa/ (with a /k...k/ sequence) surfaces more often as the rendaku form [takigawa] than as [takikawa]. The frequency of rendaku in these variable names could also tell us how E1's last consonant affects the likelihood of rendaku application. To fully address these issues, an objective and extensive examination of rendaku patterns in surnames is necessary.

4.2. Methods

As a first step, I created an original list of Japanese surnames by combining data obtained from two existing databases. Main data came from *the Database of Japanese Surnames and Their Rankings*

[11] For more details, see Kawahara (2015) and Kubozono (2008) and references therein. Accent patterns can also be variable in certain words. For example, some speakers accept (11b) as unaccented.

[12] Besides Kubozono (2005), Sugito (1965) and Zamma (2005), linguistic studies which discuss rendaku patterns in surnames include Asai (2014), Hirata (2010, 2011), Kindaichi (1976), Ohta (2013) and Sato (1989).

(Shirooka & Murayama, 2011) which gives 25,000 surnames in kanji and the number of households having each surname that are registered in telephone directories. From the database, all the surnames written with two kanji letters (i.e. bimorphemic surnames) were extracted. Since kanji can have multiple different readings, possible pronunciations of each surname were taken from another online database (Suzaki, 2013) that provides logically possible readings of surnames. From the combined database of bimorphemic surnames and their readings, the 1,000 most common surnames with E2s which have an initial /s/, /t/, /k/ or /h/ underlyingly (i.e. potential rendaku undergoers) were extracted.

Note that the created list still does not give information as to whether or how often each surname undergoes rendaku. As mentioned above, the kanji script does not reflect rendaku voicing, and the household numbers of surnames in the database are all based on kanji. Thus, for example, it shows the number of registered households having the surname written "中田" which can read [nakata] or [nakada], but the ratio of the rendaku form to the non-rendaku form is not available. To solve the issue, I further compiled data from *mixi*, which is a popular social networking website in Japan with about 13 million users as of 2012. Many mixi users make their names public so that other users can search for them. Some users also provide their names in kanji and the readings in other types of scripts such as hiragana, katakana and the Roman alphabet, which always indicate rendaku voicing. A friend search function allows us to search for users whose profiles contain a particular kind of information. For example, one can type someone's surname in kanji and its reading in hiragana in the search box as in "中田 nakata" (the hiragana script is replaced here by the Roman alphabet for expository purposes). The function then returns the number of hits, namely the number of users who have the search terms in their user profiles. A computer script was used which would search for the rendaku reading and the non-rendaku reading of each surname in the list. For example, for /naka-ta/ 中田, the script searched for "中田 nakata" and "中田 nakada". From the obtained numbers, the rendaku rate of the surname was calculated. As there were 85 users with the keyword "中田 nakata" and 22 users with the keyword "中田 nakada", the rendaku rate of this surname was deemed to be 22/(85+22) or about 20%. The procedure was repeated using other orthography types, and the final rendaku rate was obtained by averaging the results of all the methods.[13] Then, based on the rendaku rate and the total number of households for each surname, the numbers of those with rendaku and those without rendaku were calculated. In the case of /naka-ta/ 中田, which has the total households of 37,099 and the rendaku rate of 18.578%, I concluded 30,207 of them have the non-rendaku reading [nakata] and 6,892 of them have the rendaku reading [nakada].[14]

4.3. Results

Since the main interest of the current study is the effect of the place and voicing of E1's last obstruent on rendaku application, surnames without an obstruent in E1's last mora (e.g. /nari-ta/ 成田; /oo-ta/ 大田) were excluded from analysis. Also, since non-native morphemes are generally less likely to undergo rendaku, surnames with Sino-Japanese E2s (e.g. /sugi-too/ 杉藤) were excluded. As a result, 322 surnames with 50 different E2s were analyzed.[15]

Let us first examine the effect of (Strong) Lyman's Law. I have proposed that a compound surname is treated as a stem and thus stem-bounded Lyman's Law applies to the whole compound. That is to say, a voiced obstruent in E1 blocks application of rendaku on E2's initial segment across the boundary between E1 and E2 (e.g. /huzi/ + /ta/ → huzi-**t**a, *huzi-**d**a), as reported in the previous literature on surnames with /ta/ as E2 (Sugito, 1965) and some other common surnames (Zamma, 2005). We expect such patterns to hold systematically in a large corpus of surnames.

[13] Six different searching methods were used: "kanji & hiragana", "kanji & katakana", "kanji & Roman alphabet", "hiragana" only, "katakana" only and "Roman alphabet" only.

[14] The original databases actually give 39,637 as the number of households with the surname written 中田 and also give [tyuuta] and [tyuuda] as other possible readings. ([tyuu] is a Sino-Japanese reading of the E1 kanji 中 'center'.) Based on the search results, [tyuuta] and [tyuuda] accounted for about 6.4% of the population with the surname 中田. Thus, I considered 2,538 (6.4% of 39,637) to be the total household number of the underlying /tyuu-ta/ and 37,099 (93.6%) to be the total household number of the underlying /naka-ta/, and treated them as different surnames.

[15] Sonorants in E1's last mora may also affect rendaku application. The liquid /r/ tends to block rendaku while nasals trigger it (e.g. [hira**t**a]; [ima**d**a]; see Kubozono, 2005; Zamma, 2005). Also, some Sino-Japanese morphemes do undergo rendaku commonly after a moraic nasal (e.g. /aN-too/ [an**d**oo]). I leave these issues for future research.

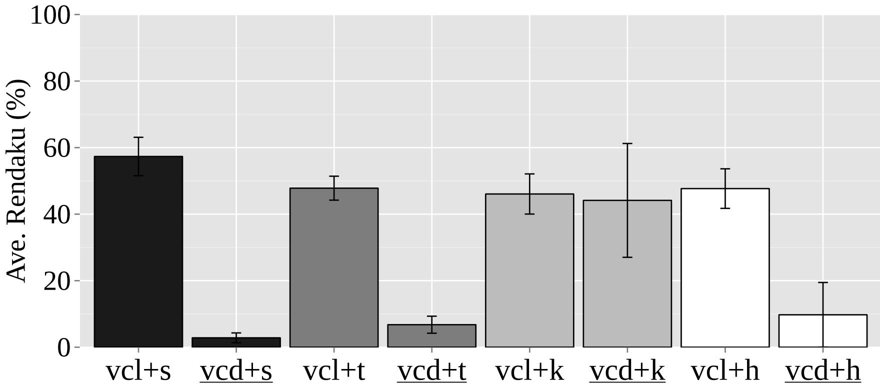

Figure 1: Average rendaku rates by surname type based on the voicing of E1's last consonant and the place of E2's initial consonant. Error bars represent ±1 standard errors.

Figure 1 plots the average rendaku application rates (%) by surname type based on the voicing of E1's last consonant ("vcl" for voiceless and "vcd" for voiced obstruents) and the place of E2's initial consonant (/s/, /t/, /k/ or /h/).[16] For example, the bar labelled "vcl+s" includes surnames such as /naka+sima/ 'center-island' and that labelled "vcd+s" includes /naga+sima/ 'long-island'. As can be seen, surnames with a voiced obstruent in E1 (which are underlined and predicted to not undergo rendaku) generally have lower rendaku rates than those with a voiceless obstruent in E1. /vcd+k/ appears to have a relatively high rendaku rate. This is mainly due to the fact that there is a particularly small number of surnames in this condition (9 out of the total 322 surnames) and one frequent E2 /kuti/ 'entrance' (appearing in 4 out of the 9 surnames) always undergoes rendaku regardless of the voicing of E1's last consonant (e.g. [mizoguti]; [deguti]).[17] Without surnames with /kuti/, the average rendaku rate of /vcd+k/ would be 0.88%.

To test the effects of the boundary-spanning Lyman's Law statistically, a mixed-effects logistic regression model was run using the glmer() function of the lmerTest package (Kuznetsova et al., 2013) built on lme4 (Bates et al., 2011) in R (R Development Core Team, 1993-2015). Rendaku application (either rendaku or no rendaku) based on the calculated household numbers were entered into the model as the dependent variable.[18] The independent variables were the voicing of E1's last obstruent (voiced or voiceless), the place of E2's initial obstruent (/s/, /t/, /k/ or /h/) and their interactions. Surnames, E1s and E2s were included in the model as random intercepts. The model shows that the effects of E1-voiced obstruents are significant in that their presence lowers the probability of rendaku application (z=–4.358, $p<0.001$).[19] The effects of the place of E2's initial obstruent are not significant, indicating that general rendaku application rate is not affected by the place of the consonant undergoing the voicing change. Also, the interactions of the voicing of E1's last consonant and the place of E2's initial consonant did not turn out to be significant. This indicates that the effects of E1-voiced obstruents mentioned above do not change significantly depending on the place of E2-initial consonants.

I conclude from these results that Lyman's Law applies to the whole compound in the case of a surname and a voiced obstruent in E1 generally blocks application of rendaku.

[16] It has been reported that in regular compounds the rate of rendaku application differs depending on the place of E2's initial consonant (i.e. the segment that undergoes voicing). See Ihara et al. (2011); Sano (2015).

[17] Rendaku application is affected by E2's idiosyncratic properties; certain morphemes are more likely to undergo rendaku than others (see Rosen, 2001 among others).

[18] For example, the surname /naka+ta/ 中田 discussed above is considered to show rendaku 30,207 times and no rendaku 6,892 times.

[19] The intercept of the model here refers to the condition where surnames have a voiceless obstruent in E1 and E2's initial consonant is /t/, which essentially corresponds to /vcl+t/ in the graph. Different models were run with the other conditions /vcl+s/, /vcl+k/ and /vcl+h/ as their intercepts; the effects of E1-voiced obstruents turned out to be significant in any model in the direction of lowering the probability of rendaku application (z=–6.931, $p<0.001$, z=–3.574, $p<0.05$ and z=–3.706, $p<0.01$, respectively). This suggests that the effect of Lyman's Law is robust regardless of the place of E2's initial consonant.

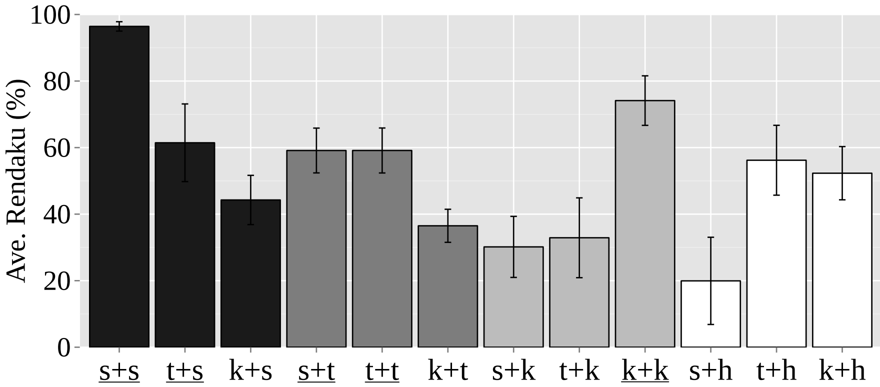

Figure 2: Average rendaku rates by surname type based on the place of E1's last consonant and the place of E2's initial consonant. Error bars represent ±1 standard errors.

Let us now turn to the effects of Identity Avoidance. I have proposed that a compound surname is treated as a stem and thus stem-internal restrictions on similar consonant sequences apply to the whole compound. That is to say, rendaku is triggered when a voiceless obstruent in E1's last mora and a voiceless obstruent in E2's initial mora share place (e.g. /asa/ + /ta/ → asa-**d**a, *asa-**t**a). Descriptions given by previous studies (Kubozono, 2005; Zamma, 2005) seem to suggest that surnames with /ta/ as E2 follow such patterns; E2-initial /t/ mostly undergoes rendaku if E1's last obstruent is /s/ or /t/, which is homorganic, but not as much if it is /k/, which is non-homorganic. We expect similar patterns to be found in surnames with E2s other than /ta/. Specifically, surnames with /k/-initial E2s should undergo rendaku when E1's last obstruent is also /k/ (e.g. /naka/ + /kawa/ → naka-**g**awa, *naka-**k**awa).[20]

Figure 2 plots the average rendaku application rates (%) by surname type based on the place of E1's last consonant (/s/, /t/ or /k/) and the place of E2's initial consonant (/s/, /t/, /k/ or /h/). For example, the bar labelled "s+s" includes surnames such as /nisi+sima/ 'West-island'. As can be seen, surnames that have a sequence of homorganic voiceless obstruents (which are underlined and predicted to undergo rendaku) generally have higher rendaku rates than those that do not. Notice that /k+k/ has a higher rendaku rate than /s+k/ and /t+k/, as predicted. Rendaku rates also seem to differ among surnames with homorganic consonant sequences. /s+s/ has a higher rendaku rate than /t+s/. This suggests that Identity Avoidance effects are stronger when the involved voiceless obstruents have not only the same place but also the same manner (i.e. when they are completely identical).[21]

A mixed-effects logistic regression model was run with rendaku application as the dependent variable. The independent variables were homorganicity (whether E1's last voiceless obstruent and E2's initial voiceless obstruent share place or not), the place of E2's initial obstruent (/s/, /t/, /k/ or /h/) and their interactions. Surnames, E1s and E2s were included as random intercepts. The model shows that the effects of homorganicity are significant in that the presence of a homorganic obstruent in E1 raises the probability of rendaku application (z=1.985, p<0.05). The interaction of homorganicity and E2-initial /k/ is significant; a homorganic obstruent in E1 further raises the probability of rendaku if E2-initial obstruent is /k/ (z=3.573, p<0.05). This indicates that avoidance of homorganic voiceless obstruents generally triggers rendaku and the effects are even stronger when the involved consonants are /k/.

To examine the effects of Identity Avoidance in more detail, total identity (the sequential obstruents share both place and manner) was entered into the model. Once both total identity and homorganicity are included in the model, the effects of homorganicity are no longer significant (z=1.288, p=0.160) and the

[20] Although the hypothesis predicts that surnames with a /h...h/ sequence should undergo rendaku, the prediction is hard to test. /h/ does not appear in intervocalic position for historical reasons, and there are very few surnames which have /h/ in both E1's last mora and E2's initial mora. Also due to history, /h/ becomes [b] when it undergoes rendaku (see Ueda, 1898), and thus rendaku may create a labial-labial sequence if E1's last consonant is /m/ (e.g. /ima/ + /hara/ → ima-**h**ara, *?ima-**b**ara). The hypothesis predicts that Identity Avoidance blocks rendaku in such a case. Due to the limit of space, I will not address this issue here.

[21] The hypothesis did not have particular predictions about surnames with /h/-initial E2s although some differences were found among them. More investigation is needed to figure out what factors are driving such differences.

effects of total identity are also not significant when E2's initial consonant is /t/ (z=0.6278, p=0.557). However, the interaction of complete identity and E2-initial /s/ and the interaction of complete identity and E2-initial /k/ are both significant in the direction of raising the probability of rendaku application (z=4.300, $p<0.05$ and z=3.231, $p<0.05$, respectively). This indicates that rendaku is triggered if a surname has a sequence of voiceless obstruents of the same place and manner, such as /s...s/ and /k...k/; however, the effects are weaker if the involved obstruents are /t/, as in /t...t/, or they share only place and not manner, as in /s...t/ and /t...s/.

I conclude from these results that, as has been proposed, rendaku in compound surnames is indeed driven by gradient restrictions on the occurrence of similar consonant sequences, which usually hold within stems (Kawahara et al., 2006). It remains unanswered, however, why the Identity Avoidance effects in surnames appear to be somewhat weaker than those in normal stems (for example, homorganicity alone is not strong enough to trigger rendaku) or why they are sensitive to the place of the involved consonants (even total identity is not strong enough to trigger rendaku in the case of /t...t/). Further investigation is needed to settle these issues.

To summarize, the overall results are compatible with the hypothesis that compound surnames are subject to stem-internal phonotactics. Stem-bounded Lyman's Law and Identity Avoidance apply to the whole compound and rendaku applies accordingly.

5. Conclusion

I have proposed that compound surnames are treated like stems and that rendaku in surnames is governed (blocked and triggered) by stem-internal phonotactics. The results of a corpus study using social media show that stem-bounded Lyman's Law and Identity Avoidance both play a role in rendaku application in surnames, supporting the hypothesis. The current research not only reveals the existence of such systematic patterns but also accounts for why compound surnames show rendaku patterns which are not found in regular compounds. The findings further our understanding of rendaku and the phonology of proper names in general.

References

Asai, Atsushi (2014). Rendaku seiki no keikō to teichaku-ka [A tendency in the occurrence of rendaku and lexicalization]. *NINJAL Research Papers* 7, 27–44.

Bates, Douglas, Martin Maechler & Ben Bolker (2011). lme4: Linear mixed-effects models using S4 classes. R package.

Endo, Kunimoto (1980). Hi-rendaku no hōsoku no shōchō to sono imi: Daku-shiin to bion to no kankei kara [A decay of the rendaku-blocking law and its consequence: An examination through the relationship between voiced obstruents and nasals]. *Kokugo Kokubun* 50:3, 38–54.

Hay, Jennifer (2003). *Causes and consequences of word structure*. [A revised version of the author's Ph.D. thesis (2000). Northwestern University], Routledge, New York.

Hirata, Junko (2010). Rendaku shiron [An essay on rendaku]. *Kobe Kaisei Joshi Gakuin Daigaku Kenkyū Kiyō* 49, 31–37.

Hirata, Junko (2011). Rendaku shiron (sono 2) [An essay on rendaku 2]. *Kobe Kaisei Joshi Gakuin Daigaku Kenkyū Kiyō* 50, 89–93.

Ihara, Mutsuko, Katsuo Tamaoka & Hyunjung Lim (2011). Rendaku and markedness: Phonetic and phonological effects. Talk presented at Tokyo Circle of Phonologists (TCP), July 24th.

Irwin, Mark (2014). Rendaku across duplicate moras. *NINJAL Research Papers* 7, 93–109.

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

Ito, Junko & Armin Mester (2016). Unaccentedness in Japanese. *Linguistic Inquiry* 47:3.

Kawahara, Shigeto (2015). The phonology of Japanese accent. Kubozono, Haruo (ed.), *The Handbook of Japanese Phonetics and Phonology*, De Gruyter Mouton, Berlin, 445–492.

Kawahara, Shigeto & Shin-ichiro Sano (2014a). Identity avoidance and rendaku. Kingston, John, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.), *Proceedings of Phonology 2013*.

Kawahara, Shigeto & Shin-ichiro Sano (2014b). Testing Rosen's Rule and Strong Lyman's Law. *NINJAL Research Papers* 7, 111–120.

Kawahara, Shigeto, Hajime Ono & Kiyoshi Sudo (2006). Consonant co-occurrence restrictions in Yamato Japanese. Vance, Timothy J. & Kimberley Jones (eds.), *Japanese/Korean Linguistics*, CSLI Publications, Stanford, vol. 14, 27–38.

Kindaichi, Haruhiko (1976). Rendaku no kai [An account of rendaku]. *Sophia Linguistica* 2, 1–22.

Kindaichi, Haruhiko, Ōki Hayashi & Takeshi Shibata (eds.) (1988). *Nihongo hyakka daijiten [An Encyclopaedia of the Japanese Language]*. Taishūkan, Tokyo.

Kubozono, Haruo (2005). Rendaku: Its domain and linguistic conditions. van de Weijer, Jeroen, Kensuke Nanjo & Tetsuo Nishihara (eds.), *Voicing in Japanese*, Mouton de Gruyter, Berlin and New York, 5–24.

Kubozono, Haruo (2008). Japanese accent. Miyagawa, Shigeru & Mamoru Saito (eds.), *The Oxford Handbook of Japanese Linguistics*, Oxford University Press, Oxford, 165–191.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojesen Christensen (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package.

Lyman, Benjamin (1894). The change from surd to sonant in Japanese compounds. The Oriental Club in Philadelphia (ed.), *Oriental studies: A selection of the papers read before the Oriental Club in Philadelphia 1888-1894*, Ginn and company, Boston, 160–176.

Martin, Samuel E. (1952). *Morphophonemics of Standard Colloquical Japanese*. No. 47 in Language dissertation, The Linguistic Society of America, Baltimore.

McCawley, James (1968). *The phonological component of a grammar of Japanese*. Mouton, The Hague.

Miura, Keiichi (2012). *Yamaha sōsōfu [The history of Yamaha]*. Ankasha, Hamamatsu.

Miyake, Hideo Marc (2003). *Old Japanese: A Phonetic Reconstruction*. Routledge, London.

Ohta, Satoshi (2013). On the relationship between rendaku and accent: Evidence from -kawa/-gawa alternation in Japanese surnames. van de Weijer, Jeroen & Tetsuo Nishihara (eds.), *Current issues in Japanese phonology: Segmental variation in Japanese*, Kaitakusha, Tokyo, 63–87.

R Development Core Team (1993-2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Rosen, Eric (2001). *Phonological processes interacting with the lexicon: Variable and non-regular effects in Japanese phonology*. Ph.D. thesis, University of British Columbia.

Sakurai, Shigeharu (1972). Heian Insei jidai ni okeru wago no rendaku ni tsuite [On rendaku in native words in Japanese of the Heian Insei era]. *Kokugo Kokubun* 41:6, 1–19.

Sano, Shin-ichiro (2015). Universal markedness reflected in the patterns of voicing process. Bui, Thuy & Deniz Özyıldız (eds.), *Proceedings of the North East Linguistic Society (NELS)*, vol. 45-3, 49–58.

Sato, Hirokazu (1989). Fukugō-go ni okeru akusento kisoku to rendaku kisoku [Accent rules and rendaku rules in compounds]. *Kōza nihongo to nihongo kyōiku 2: Nihongo no onsei, on'in (jō) [Japanese and Japanese teaching 2: Japanese phonetics, phonology 1]*, Meiji Shoin, Tokyo, 233–265.

Shirooka, Keiji & Tadashige Murayama (2011). Nihon no sei no zenkoku jun'i database [The database of Japanese surnames and their rankings]. Available at: http://www.ipc.shizuoka.ac.jp/˜jjksiro/kensaku.html.

Sugito, Miyoko (1965). Shibata-san to Imada-san: Tango no chōkakuteki benbetsu ni tsuite no ichi kōsatsu [Shibata-san and Imada-san: An examination of auditory distinction of words]. *Gengo Seikatsu* 165, 64–72.

Suzaki, Haruo (2013). A private on-line corpus of 111,711 Japanese family names. http://www2s.biglobe.ne.jp/˜suzakihp/index40.html.

The Meiji Government (1875). *Heimin myōji hisshō gimu rei [Act on the obligatory use of surnames by commoners]*. Hōrei Zensho. [Published by the Meiji Cabinet in 1887]. Available in the National Diet Library Digital Collections: http://dl.ndl.go.jp/info:ndljp/pid/787955/71, Naikaku Kanpō-kyoku, Tokyo.

Toda, Ayako (1988). Wago no hi-rendaku kisoku to rendaku keikō [The rendaku-blocking rules and rendaku-inhibiting factors in native Japanese words]. *Dōshisha Kokubungaku* 30, 80–98.

Ueda, Kazutoshi (1898). Gogaku sōken: p-on-kō [A new linguistic perspective: On the sound /p/]. *Teikoku Bungaku* 4:1, 41–46.

Unger, Marshall (1977). *Studies in Early Japanese morphophonemics*. [Originally presented as the author's Ph.D. thesis (1975). Yale University], Indiana University Linguistics Club, Bloomington.

Vance, Timothy J. (2005). Sequential voicing and Lyman's Law in Old Japanese. Mufwene, Salikoko S., Elaine J. Francis & Rebecca S. Wheeler (eds.), *Polymorphous linguistics: Jim McCawley's legacy*, MIT Press, Cambridge, MA, 27–43.

Vance, Timothy J. (2015). Rendaku. Kubozono, Haruo (ed.), *The Handbook of Japanese Phonetics and Phonology*, Handbooks of Japanese Language and Linguistics 2, Mouton de Gruyter, Berlin and New York, 397–441.

Walter, Mary Ann (2007). *Repetition Avoidance in Human Language*. Ph.D. thesis, Massachusetts Institute of Technology.

Yasuda, Nobufumi (1983). Studies of isonymy and inbreeding in Japan. *Human Biology* 55-2, 263–76.

Zamma, Hideki (2005). The correlation between accentuation and rendaku in Japanese surnames: A morphological account. van de Weijer, Jeroen, Kensuke Nanjo & Tetsuo Nishihara (eds.), *Voicing in Japanese*, Mouton de Gruyter, Berlin and New York, 157–176.

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:

Tanaka, Yu. 2017. Phonotactically-Driven Rendaku in Surnames: A Linguistic Study Using Social Media. In *Proceedings of the 34th West Coast Conference on Formal Linguistics*, ed. Aaron Kaplan et al., 519-528. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3359.