Family Matters: Lexical Aspects of Japanese Rendaku

Jeroen van de Weijer
Shanghai Int'l Studies University

Clemens Poppe
The University of Tokyo

Marjoleine Sloos
University of Groningen

Abstract: Some roots in Japanese compounds always undergo the rule of rendaku, others never undergo the rule, and still others vacillate. In this paper, we investigate this kind of lexical variation from the perspective of the frequency of such roots. Different types of frequency are considered, such as that of roots in isolation, and the roots' frequency of occurrence as a left- or right-hand member of compounds. We show that frequency is related to the likelihood of a root undergoing rendaku. Since rendaku also clearly involves phonological factors, we argue that this result should be interpreted in a model which integrates usage-based factors with phonological grammar.

1. Introduction

In this paper we discuss the Japanese rule of *rendaku* (lit. "sequential voicing") from a combined generative and usage-based perspective. Rendaku voices the first segment of a compound if a number of conditions are satisfied. The history of this rule, as well as its (ir)regularity and the conditions on its application in present-day Japanese, have attracted a great deal of attention in the morphological and phonological literature. One aspect has been especially problematic for past analyses: even if all conditions on its application are taken into account, rendaku appears to have a considerable number of lexical exceptions. That is, some lexical items are more prone to undergo the rule than others, on a seemingly idiosyncratic basis. In this paper, we show that the propensity of these roots for undergoing rendaku can be related to the frequency with which these roots occur in isolation as well as in compounds. Apart from contributing to the solution of a long-standing puzzle in Japanese morphophonology, a general conclusion that can be drawn from this analysis is that the frequency with which words are used (both in compounds and in isolation), should be included in the explanation for their phonological behaviour. Since rendaku is also governed by purely phonological constraints (such as Lyman's Law; see

below), a satisfactory solution must combine aspects of generative linguistics and of usage-based approaches.

This paper is organized as follows: in section 2, we illustrate the basic mechanisms of rendaku and point out some of the relevant conditions on the application of the rule that have been uncovered so far in the literature. In section 3, we introduce the lexical variation which forms the main topic of this paper and formulate a number of hypotheses to explain this behaviour. Section 4 describes the methodology and the data set we investigated. Section 5 provides the statistical results of our investigation, and section 6 discusses an outline of the architecture of the grammar and the lexicon in which these results can be interpreted.

2. Rendaku

The rule of sequential voicing in Japanese, normally referred to as *rendaku*, has been a topic of discussion in the Japanese and general literature for a very long time (see e.g. Haraguchi (2001), Irwin (2005; 2009), Ito & Mester (1986; 2003); Kubozono (2005); McCawley (1968); Otsu (1980); Rosen (2001); Vance (1987; 2005; 2007) and references cited there, also to the extensive literature in Japanese). The basics of the rule are straightforward: the initial voiceless obstruent of a root becomes voiced when the root appears as the right-hand member of a compound (subject to a number of conditions; see below). Examples, taken from various standard sources, are given in (1):

(1)	shima	-	kuni	→	shima-guni	'island country'
	'island'		'country'			
	roten	-	huro	→	roten-buro	'outdoor bath'
	'outdoor'		'bath'			
	maki	-	sushi	→	maki-zushi	'rolled sushi'
	'roll'		'sushi'			
	isi	-	tooroo	→	isi-dooroo	'stone lantern'
	stone		lantern			

There are two things to note: the voiced counterpart of [s] is [dz] (conventionally transcribed as z), and the voiced counterpart of [h] is [b] (the source for this latter alternation is historical: /h/ derives historically from /p/). See Frellesvig (2010: Ch. 2), Vance (1987: Ch. 10), and many of the contributions to van de Weijer et al. (2005) and references cited there, for discussion of the history of this phenomenon and other relevant information.

There are a number of conditions on the application of rendaku, some of which are regular and well-known, and some of which are the object of debate. First, rendaku is almost exceptionlessly blocked if there is a voiced obstruent in the second part of the compound itself. This condition on the application of rendaku, which is illustrated in (2), is usually referred to as Lyman's Law (Lyman, 1894) (see Vance (1987: 136ff.) and many other sources for discussion).

(2) Lyman's Law: two voiced obstruents in one root are not permitted:

kami	-	ka**z**e	→	kami-ka**z**e	'kamikaze'
'divine'		'wind'		(kami-*gaze)	

There is also an exception to this generalization, viz. the single lexical item in (3), in which the [g] in the second member does not block voicing of [h] to [b].

(3) Rendaku is, unexpectedly, not blocked (Vance, 1987: 137):

nawa	-	hasi**g**o	→	nawa-basi**g**o	'rope ladder'
'rope'		'ladder'			

Second, rendaku applies almost exclusively to native Japanese words, and less often to words borrowed from Chinese (which are referred to as Sino-Japanese) or other languages (i.e. older or more recent loanwords). In fact, Martin (1952) claimed that the process is frequent only in native Japanese words, but research since then has shown that there are a considerable number of examples involving Sino-Japanese words, older loanwords (e.g. from Portuguese), and even some more recent loanwords that undergo rendaku (see Takayama (2005), Irwin (2011: 151f)). This suggests that rendaku is a productive rule: see Vance (1987: 140-41) for examples and discussion, also of the extent to which rendaku applies in mimetics (reduplicated vocabulary) and in other parts of the morphology such as derivation and inflection. In this paper we will focus on the status of rendaku in native Japanese compounding.

Apart from Lyman's Law and the restriction to loanwords, other conditions on the application of rendaku that have been discussed are related to prosodic size (see e.g. Irwin (2009)), vowel length (Horton & Minami, 2011), syntactic branching in longer compounds (Otsu (1980), Kubozono (2005)), semantic inclusion/exclusion relationships (Shibatani, 1990: 174), and the location of accent (see e.g. Zamma (2005), Ohta (this volume), Yamaguchi & Tanaka (this volume)). Some of these conditions are (almost) exceptionless (such as

Lyman's Law), and others represent tendencies that may block rendaku in a certain percentage of cases, i.e. they are variable. It is also possible that other categories influence the application of rendaku, such as the initial consonant or the initial mora of the root that undergoes the rule: some consonants or moras might be more susceptible for rendaku than others, e.g. because an alternation between [h] and [b] might be less transparent than an alternation between [t] and [d]. Second, the number of syllables of the left-hand side of the compound might matter, or whether the left-hand side of the compound contains a voiced consonant or not (cf. the "extended Lyman's Law", see e.g. Vance (2007)). There is isolated discussion of these topics in the literature cited above.

In our investigation, we will focus on roots where rendaku would be expected, e.g. words in which Lyman's Law is not relevant. We will only take into consideration compounds consisting of two roots, to abstract away from phrasal effects. Similarly, we exclude words with long vowels, to avoid any interference of vowel length. Finally, we will not consider semantic aspects of compounds that do or do not undergo rendaku. The properties that were taken into consideration will be spelled out in detail below (§ 4).

First, however, we need to discuss the fact that, even if all of the constraints on rendaku outlined above are taken into account, some lexical items do and other lexical items do not undergo the rule, in a seemingly unpredictable fashion.

3. Lexical variation

Individual roots behave differently with respect to rendaku in Japanese. That is, some roots that would at first glance be expected to undergo rendaku do not undergo the process at all, or they show rendaku in some compounds but not in others. Vance (1987: 146ff.) refers to this situation as the "fundamental irregularity" of rendaku, and Miller (1967: 195) describes the (non-)application of the rule as "completely bewildering" (see also Ohno (2000), Rosen (2001)). Vance (1987: 147) gives a number of words that, unexpectedly, resist rendaku, which are reproduced in (4):

(4) *Exceptions to rendaku ("never")*

soko	-	tuti	'subsoil'	(*soko-duti)
bottom		soil		
kutu	-	himo	'shoelace'	(*kutu-bimo)
shoe		lace		

suna	-	kemuri	'sand cloud'	(*suna-gemuri)
sand		cloud		
yubi	-	saki	'fingertip'	(*yubi-zaki)
finger		tip		
asa	-	sio	'morning tide'	(*asa-zio)
morning		tide		

Note that the right-hand members in (4) have no voiced obstruents and all forms involve native Japanese roots. Thus, they are phonologically fully comparable to the examples in (1). In other words, the fact that they form exceptions to rendaku cannot be attributed to any of the known conditions on rendaku. The question is therefore how these forms should be dealt with: have we missed a condition that is relevant? Are they exceptions that must somehow be marked in the lexicon? Or is the exceptional status of these words due to factors outside the grammar? If rendaku is a productive rule in Japanese, and if no conditions can be found that explains the "exceptions" in (4), then such compounds must be marked as lexical exceptions.

The question of whether rendaku is a productive rule or not in contemporary Japanese is relevant here. The issue is discussed in detail by Kubozono (2005). After a consideration of the evidence, partly on the basis of experiments with SLI-speakers (Fukuda & Fukuda, 2000), he concludes that rendaku voicing is a productive rule in Japanese, although some cases of voicing (and, presumably, cases where voicing does not take place) may be lexicalized (see also Ito & Mester (2003: 124)). Hence, the question how the non-application of rendaku in forms such as those in (4) should be accounted for is an important one.

Rosen (2001: Appendix D) lists 19 further examples of roots like those in (4) which he characterizes as "never" undergoing rendaku, besides roots which always undergo rendaku ("rendaku-lovers"). He also distinguishes one other category: roots that vacillate between the two options ("rendaku-haters") (citing p.c. from Haruo Kubozono who used similar terms). He also draws attention to the lexical nature of the distinction, i.e. whether a word is a "lover" or a "hater" is unpredictable from any of the (known) conditions on rendaku.

Rosen provides 113 examples of "rendaku-lovers", i.e. words that always or almost always undergo the rule. The relatively large number of "lovers" also suggests that rendaku should be considered a productive rule in Japanese. Examples of rendaku "lovers" are given in (5):

(5) *Rendaku "lovers"*

huro	'bath',	e.g.	uti-buro	'inside bath'
			soto-buro	'outside bath'
			mizu-buro	'water-bath'
kiwa	'brink'	e.g.	yama-giwa	'mountain brink'
			hae-giwa	'receding hairline'
			te-giwa	'skill' [hand brink]
sake	'sake'	e.g.	nama-zake	'raw sake'
			taru-zake	'barrel sake'
			sio-zake	'salt salmon'

(from Rosen (2001: Appendix E))

Both Vance and Rosen note that a number of words vary in their status, i.e. they show rendaku in some compounds but not in others. Rosen refers to such roots as "haters", but we propose to refer to them as "doubters". Vance provides the example in (6a), and Rosen (2001: Appendix F) lists examples of compounds with four of such doubter roots, such as those in (6b):

(6) *Rendaku "haters/doubters"*

a. ki	'tree'	cf.	niwa - ki	'garden tree'
		vs.	yama - gi	'mountain tree'
b. hara	'field'	cf.	sino - hara	'bamboo field'
			sasa - hara	'bamboo grass field'
		vs.	una - bara	'ocean field'
			kuwa - bara	'mulberry field'
kusa	'grass'	cf.	ira - kusa	'nettle' [thorn grass]
			natu - kusa	'summer grass'
		vs.	hituzi - gusa	'sheep grass'
			no - gusa	'wild grass'

At first sight, it would seem to be impossible to predict in which category a root might fall, i.e. whether a given compound shows rendaku or not (recall the quotes from Vance and Miller above). However, we suspect that if the notion of *frequency* is taken into account, it is possible to predict whether roots belong in the category of "never" (4), lovers (5), or doubters (6). Frequency of use has been shown to be an important factor both in historical change and in synchronic language behaviour in much recent work (e.g. Bybee & Hopper (2001), Bybee (2006), Phillips (2006), Gahl (2008)). This is not surprising, be-

cause if a word is more frequently used it is more easily activated, more easily recognized, etc. Frameworks like Exemplar Theory argue that differences in frequency may have repercussions for linguistic structure (see references just cited, and also see the discussion in section 6 below). From this perspective, it would make sense that roots which are extremely frequent, or which appear in particular positions in a compound particularly frequently, show different linguistic behaviour.

We need to define exactly what types of frequency might be relevant in this respect. First, however, it is necessary to take into consideration that in rendaku, a root typically has two allomorphs, one of which is the 'basic' (non-rendaku) form and the other is the rendaku form. These terms are illustrated in (7) (cf. the examples in (5) above):

(7) root: /huro/ 'bath'

allomorphs:	[huro]	<u>non-rendaku form</u> (i) in isolation; (ii) as left-hand member in compounds
	[buro]	<u>rendaku form</u> (i) as right-hand member in rendaku compounds

Below we will investigate whether the frequency of occurrence of both allomorphs of a root is responsible for the variation in rendaku. Consider a root which appears extremely frequently in its *non*-rendaku form (i.e. in isolation or on the left side in compounds) but almost never in its rendaku form. The non-rendaku allomorph will be more strongly represented in the lexicon ("entrenched" in Exemplar Theory terms) and it is thus possible that it is more likely to impose its shape on the rendaku allomorph (the effect might be compared to a "majority rule" or "paradigm uniformity" effect, see e.g. Kenstowicz (2005)). This would make such roots "rendaku-haters". If roots appear in their rendaku form relatively frequently, we may expect the rendaku allomorph to be sufficiently well-entrenched in order to surface when the phonological conditions for rendaku are satisfied. This would make the root a "rendaku lover". If the rendaku forms and non-rendaku forms surface in a "reasonable" frequency proportion (where what is "reasonable" should of course be determined on the basis of a statistical investigation), we would expect it to be a "doubter".

What types of frequency are relevant here? The frequency of a root in isolation can be determined rather straightforwardly. This can be measured as the

number of tokens in a given corpus. The number of compounds with a particular root as its left- or right-hand member could be counted in two ways, viz. either as *family frequency* or as *family size*. Family *frequency* refers to the summed token frequencies of compounds in which a particular root appears (regardless of how many different compounds the root appears in). Family size, on the other hand, refers to the *number* of compounds formed with a particular root, i.e. it is a type frequency variable. If there is a large number of compounds with a particular root in the left-hand side of a compound (regardless of their token frequency), then the shape of that root might also be imposed on the root when it appears on the right-hand side of a compound (again, following the same rationale as for the previous hypothesis). The literature shows that this "family size" aspect may have an effect on (processing of) morphological structure (e.g. de Jong et al. (2000)). We hypothesize that both higher family size as well as higher family frequency of a root as the *left*-hand member of a compound (i.e. in its non-rendaku form) correlate with avoidance of rendaku ("hatred"). Conversely, we expect that both higher family size as well as higher family frequency of a root as the *right*-hand member of a compound (in its rendaku form) correlate with likelihood of rendaku occurring ("love"). Note, finally, that family size and family frequency are logically independent types of frequency: both, either, or neither could be found to play a role in rendaku.

So, three types of frequency are hypothesized to interact with the propensity for rendaku, in the sense that higher frequency scores are expected to block or reduce rendaku. In (8), we summarize the relevant hypotheses, together with the types of frequency on which they are based:

(8) a. Token frequency of the root in isolation
Hypothesis 1: The frequency of a root in isolation is negatively correlated with its propensity for rendaku.
b. Family size of the root as the left-hand member in compounds
Hypothesis 2: The family size of the root as the left-hand member in compounds is negatively correlated with its propensity for rendaku.
c. Family frequency of the root as the left-hand member in compounds
Hypothesis 3: The family frequency of a root as the left-hand member in compounds is negatively correlated with its propensity for rendaku.

On the other hand, two types of frequency are hypothesized to interact with rendaku in the sense that higher frequencies are expected to enhance rendaku:

(9) a. Family size of the root as the right-hand member in compounds
Hypothesis 4: The family size of the root as the right-hand member in compounds is positively correlated with its propensity for rendaku.
b. Family frequency of the root as the right-hand member of the compound
Hypothesis 5: The family frequency of the root as the right-hand member in compounds is positively correlated with its propensity for rendaku.

Apart from word frequency, we also took into account three other factors which might play a role but about which we had no strong *a priori* expectations. These were, first, the frequency of the different initial moras involved in rendaku, e.g. [hu] and [bu] in *huro* and *buro*, respectively. Perhaps some moras are favoured (i.e., more frequent) than others in Japanese and therefore preferred (cf. Pierrehumbert (2006: 526)). With respect to this factor, other considerations could also come into play, e.g. the preference for a voiced allophone in intervocalic position, or the general preference for voiceless obstruents over voiced ones. The second factor was the number of syllables on the left-hand side of the compound: is rendaku more likely to occur after shorter or longer roots? Thirdly, we included in the database whether another voiced segment was contained in the left-hand member of the compound (cf. extended Lyman's Law, see above). These three factors were treated as random variables; we will return to them in section 5.

In the next section, we will describe the methodology which we used to test the hypotheses in (8) and (9).

4. Methodology

To investigate the relevance of frequency for the lexical variation in rendaku, we selected a number of roots based on the appendices in Rosen (2001) (see section 4.1 below). Subsequently, we constructed a database of around 2,700 compounds in which these roots appear as either a left-hand member or a right-hand member (section 4.2). Finally, we added frequency information for the compounds and for the roots in isolation (section 4.3).

4.1. Selection of the roots

Appendices D-F in Rosen (2001) form the starting-point of our investigation. These appendices contain 113 *lovers* (which (almost) always undergo rendaku),

19 *haters* (which (almost) never undergo rendaku – Rosen refers to these as the "never" category), and 4 *doubters* (which show variation, and which Rosen refers to as the "hater" category). Since our investigation focuses on the latter two categories, we intended to include all compounds with the roots that are marked as "haters". However, we were forced to leave out two of the 19 "doubters" (viz. *kera* 'mole cricket' and *kasa* 'shade'), because there were no compounds with these roots in the database created by Ogawa et al (2005), which was consulted in the second phase of the construction of our database (see the next subsection). The final sets of haters and doubters are given in (10a) and (10b), respectively:

(10) a. Rendaku haters (words that almost never undergo rendaku) in the database

sio	'tide'	saki	'tip'	himo	'string'
kata	'shoulder'	kase	'shackles'	kami	'above'
hime	'princess'	tami	'people'	take	'mushroom'
take	'measure'	siro	'materials'	tuti	'earth'
kuso	'dung'	kasu	'dregs'	tuyu	'dew'
suso	'cliff'	koi	'love'		

b. Rendaku doubters (words that show rendaku variation)

hara	'field'	kusa	'grass'	kuse	'habit'
kawa	'skin'				

The large number of rendaku "lovers" in Rosen (2001) made it necessary to select "lovers" that satisfied a number of conditions relevant for our purposes. A number of these conditions are connected to the segmental and prosodic composition of these roots. In order to control for any influence of the segmental material of the first mora of these roots, we tried to include the same number (viz. 2) of roots for each initial consonant-vowel combination. Given the fact that there are four consonants that may undergo rendaku (/t s k h/), and a total of five vowels in Japanese, our ideal set of lovers consisted of $2 \times 4 \times 5 = 40$ roots. Unfortunately, some CV-combinations are not included in Rosen (2001)'s appendix (e.g. roots starting with /ti/), or represented by only one item (e.g. roots starting with /hi/). Therefore, seven forms in our "lovers" database were randomly selected from the remaining lovers, without paying attention to the initial consonant. This resulted in the following set of lovers:

(11) *Rendaku lovers in the database*

	/t/		/s/		/k/		/h/	
/i/	—		sima siri siru	'stripe' 'buttocks' 'soup'	kimo kiri	'kidney' 'mist'	hire	'fin'
/e/	tera	'temple'	semi	'locust'	keta	'beam'	hera heri	'spatula' 'rim'
/a/	tama tana tane	'ball' 'shelf' 'seed'	sake sato	'alc. drink' 'village'	kai kami	'shell' 'paper'	hai hako hato hana hana	'ash' 'box' 'pigeon' 'flower' 'nose'
/o/	toki tori	'time' 'bird'	soko sono sora	'bottom' 'garden' 'sky'	koke kosi	'moss' 'hips'	hone hosi	'bone' 'star'
/u/	tuna tuti tutu	'rope' 'hammer' 'pipe'	sumi	'ink'	kuma kutu	'bear' 'shoes'	hue huta	'flute' 'lid'

To check whether the frequency of the initial mora might have an effect on rendaku variation, we included corpus-based frequency information for every mora, provided by Tamaoka & Makioka (2004).

In all, our database therefore consisted of 61 roots in total (17 haters, 4 doubters and 40 lovers). Having established the set of roots which we investigated, let us now turn to the compounds in which these roots appear.

4.2. Selection of the compounds

As a second step, we collected as many compounds as possible in which the roots in our database actually occur and that were relevant for our investigation. For this we used the Ogawa et al. (2005) database, which contains all 78,426 two-kanji words extracted from the fourth edition of the Kōjien dictionary (Kōjien, 1991). All characters in these compounds belong to the set of 2,965 standard characters (Japanese Industrial Standard Level 1). By restricting the investigation to two-kanji roots, we obtained only compounds of two members, thus avoiding the effect of branching in longer compounds (see section 2). In order to control for the other factors which might influence variation (see again section 2), we excluded all compounds with a Sino-Japanese morpheme as the left member. Similarly, all compounds with left-hand members containing a long vowel were excluded. Compounds ending in the moraic nasal /N/ were discarded as well, since post-nasal voicing make it impossible to determine whether voicing of an initial consonant on the right-hand side is due to rendaku or to post-nasal voicing.

This resulted in a data set of 2,702 compounds in which the 61 roots appear. In the next section, we explain how we collected frequency information on these roots and compounds.

4.3. Frequency

We included the following compound-dependent frequency types in the database (cf. the hypotheses in (8) and (9) above).

(12) a. Token frequency of the root in isolation
b. Token frequency of each compound
c. Family size left: The number of different compounds in which a particular root appears as the left-hand member
d. Family frequency left: The sum of all token frequencies of all compounds in which a particular root appears as the left-hand member
e. Family size right: The number of different compounds in which a particular root appears as the right-hand member
f. Family frequency right: The sum of all token frequencies of all compounds in which a particular root appears as the right-hand member

All token frequencies of roots and compounds were computed on the basis of the Chunagon database of the Balanced Corpus of Contemporary Written Japanese (BCCWJ, 2011). This is an online version of a database which is based on a corpus of written Japanese containing texts written between 1971 and 2008, taken from all kinds of genres: newspapers, magazines, textbooks, poetry, PR brochures, legal texts, recordings of meetings of the Japanese Diet, and the internet (such as blogs). All possible different spellings (if applicable) were included in the frequency count. The frequencies of compounds in the database ranged from 0 (e.g. *imogai* 'cone shell') to 7,147 (*tegami* 'letter'). It should be noted that, since the database is ultimately based on a general dictionary, it contains many infrequent words. We return to this point in the discussion (section 6).

All frequency counts were log-transformed, in order to correct for skewness in the frequency distribution (Baayen, 2008: 31). The different log-transformed frequencies vary as follows:

(13)

Root in isolation	0-5.2
Compounds	0-3.9
Family Frequency Right	0-3.9
Family Frequency Left	0-3.9
Family Size Right	0-2.1
Family Size Left	0-1.8

These figures mean that the frequency range of roots in isolation was, of course, largest, and the frequency range for Family Size Left was smallest.

5. Results

In this section we report on the tests of our hypotheses and the logistic regression test based on the database described in the previous section. These tests were carried out using the statistical package R (R-Team (2005), Baayen (2008)).

We found that the family frequency of the root as the left-hand member in compounds is the strongest predictor for occurrence or non-occurrence of rendaku. This *family frequency effect* has a *negative* impact on the propensity of undergoing rendaku of the stem ($z = -6.98$ $p < 0.001$), i.e. if a root occurs more often as a left-hand member (in terms of token frequency), then rendaku is less likely to occur. This confirms hypothesis 3 in (8c). On the other hand, the log-transformed *family size* of the left hand member has a *positive* effect on rendaku ($z = 4.93$ $p < 0.001$), i.e. if there are more compounds with a particular root on the left-hand side (type frequency), then rendaku is more likely to occur. This is the opposite of the outcome we initially expected (hypothesis 2 in (8b)). Similarly, the frequency of the root in isolation also has a positive effect on rendaku ($z = 5.06$ $p < 0.001$), which indicates that roots that are more frequent in isolation tend to favour rendaku. This is also a result that is counter to our initial hypothesis (hypothesis 1 in (8a)). We did not find any effect of the frequency of occurrence of the stem as a right-hand member of a compound, neither in terms of family frequency nor of family size (hypotheses 4 and 5, see (9)).

In order to test for interaction effects, we also performed a logistic regression test. The same variables are significant (the frequency of the root in isolation, and family frequency and family size of the root as a left member of the compound). Again, the frequency and family size of the root as a right-hand member of the compound is not significant. No significant interaction between the factors was attested. We also checked the random variables of initial mora and its frequency, the number of syllables on the left-hand side of the com-

pound, and the presence of another voiced segment in the left-hand member of the compound. We found that the initial phoneme and the number of syllables correlate with the propensity of the root of undergoing rendaku. The strongest effect of the mora, however, is its frequency. Mora frequency is negatively correlated with rendaku, that is, the higher the mora frequency, the less likely it is to undergo rendaku ($z = -4.549$ $p < 0.001$). Finally, roots which are monosyllabic or disyllabic are more likely to undergo rendaku than longer roots ($z = -6.096$ $p < 0.001$). All results are summarized in Table 1.

Table 1. Logistic regression on rendaku

	Estimate	S.E.	z-value	p-value
(Intercept)	-2.609	0.456	-2.029	0.043
LogFreqIsol	0.816	0.128	6.499	< 0.001
FamFreqL*1000	-0.481	0.047	-10.17	< 0.001
LogFamSizeL	0.909	0.184	4.884	< 0.001
Number of syllables	0.179	0.105	-6.096	< 0.001
FreqMora*1000	-0.004	-0.001	-4.549	< 0.001

To summarize, we found a number of factors that have a negative effect on rendaku: family frequency of the root as a left-hand member of the compound, the frequency of the initial mora of the root, and the number of syllables of the left side of the root. We also found two factors that are positively related to rendaku: the frequency of the root in isolation and the family size of the root when it occurs as a left-hand member of the compound. In the following section we will discuss these results.

6. Discussion

Although rendaku is of course concerned with nouns on the right-hand side of compounds, the results reported in the previous section suggest that we should also take into account the occurrence of such roots when they appear on the *left-hand* side of compounds if we wish to understand the behaviour of "lovers", "haters" and "doubters" in rendaku. The results show that the root frequency of the left part of the compound has a statistically significant negative effect on rendaku. In other words, such roots will tend to be haters or doubters. In an Exemplar Theory approach this makes sense, because this non-rendaku allomorph will have a mental representation that is relatively strong

compared to the rendaku form (with a voiced consonant). This is an important finding.

However, the other hypotheses initially spelt out in (8) and (9) are not confirmed: we did not find any effect of the frequency of the root as a right-hand member of the compound (either in terms of family size or of family frequency). Moreover, our hypotheses about the negative effect of the root in isolation and as of family size a left-hand member of a compound, which would lead to less rendaku in right-hand position, are even disconfirmed. In addition, we also found an unexpected effect of the frequency of the initial mora: moras with higher frequency tend to show less rendaku. How can these results be interpreted?

As briefly pointed out above, a large part of our database, based on observations of rendaku "hatred" and rendaku "love" by Rosen (2001) and Vance (1987), contains compounds that are very infrequent: no fewer than 1,775 out of 2,700 compounds have a frequency of zero in the corpus that we used. Some of these words are rare and/or archaic. Although our database is not based on the compounds in Rosen (2001) (but on the roots identified in his appendices), it should be pointed out, following Tanaka (2003), that some of the compounds listed by Rosen no longer form part of active native speaker knowledge. The same is probably true for our final selection, made on the basis of the Kōjien dictionary (Kōjien, 1991). In order to test whether the compounds with zero frequency are familiar or not to Japanese speakers, we carried out a small-scale experiment. We randomly selected sixteen words with a frequency of zero and added five words with higher frequency as controls. Sixteen native speakers of Japanese (eight men and eight women) scored these words on a seven-point Likert scale indicating their familiarity with these words. It turned out that a majority of these compounds were in fact recognized by native Japanese speakers, although, of course, our test does not establish whether these words also form part of their active usage. We leave this matter for further research and hope that future corpus study (or, better yet: a study based on actual spoken Japanese) can improve on our database.

As for the low-frequency words, we tentatively suggest that the morphological status of extremely low-frequency compounds may be unclear for some speakers. Some of these words might not be recognized as compounds and therefore it might be unclear whether rendaku should be applied or not, in which case competition between neighbouring exemplars may arise. The same is actually true for high-frequency words: it is questionable whether words like *tegami* 'letter' or *mabuta* 'eyelid' are still recognized as compounds for many con-

temporary speakers, especially when we take into account the fact that the meanings of these compounds cannot be predicated from their parts.

Let us now turn to a concrete example. In Exemplar Theory, the different parts of a morphologically complex word are all connected with the same morphemes in other forms (e.g. Bybee (2001)). All plurals in English, for instance, share a phonological (final *–s*) and a semantic relation (*plurality*), which together express the morphological relation of plural. In the Japanese compound *sinobue* 'flute', both *sino* and *bue* are connected to all other instances of these roots, either as independent lexical items or as parts of compounds. In our data, we found that the frequency of *bue* in isolation is relatively high (2.8, log-transformed) and its family frequency-left is zero. A higher frequency of the root in isolation is likely to increase the recognition of the root as a separate morpheme. This is confirmed in this case, since *bue* is considered to be a "rendaku lover" (Rosen (2001)). The situation is different for haters like *siro* 'substitution, price': this root seldom occurs in isolation in our data (0.6), and has a very small family size (0.8) as a left-hand member in compounds. With such low frequencies, *tumisiro* may not be treated as a compound by the language system (although of course it may well be acknowledged as such by native speakers). On the contrary, the log family frequency-left is relatively high (2.7), which thus provides the strongest cue for pronunciation, which is the voiceless variant. A typical case of a doubter is *kusa* 'grass' (see (6)), which has a relatively high frequency of the root in isolation (3.5), a high family frequency left frequency (2.1), as well as high family size left frequency (2.7). Which form will win is not predictable on the basis of these frequencies, and hence the different variants vacillate.

Note that the root frequencies presuppose that the root can be recognized as a separate morpheme. If there are enough exemplars in the lexicon in which a root occurs, it can be morphologically categorized and will be recognized as a morpheme. If there are not enough exemplars in the lexicon, it cannot be morphologically categorized and will not be recognized as a separate morpheme.

7. Conclusion

Rendaku voices the first consonant of the second part of a compound, which can be characterized as a phonological rule. This process can for a large part be captured either by a (number of) (morpho-)phonological rules or as a result of constraint interaction (see, among others, Ito & Mester (2003)). That is, the process is partly subject to purely phonological conditions, some of which are

near-categorical (e.g. Lyman's Law, see (2)). The variation to which rendaku is subject should, however, not be ignored. In order to explain this kind of lexical variation, i.e. whether Japanese roots behave as "rendaku-lovers" or "rendaku-haters", frequency of use and morphological family size are relevant. This holds especially true for cases in which rendaku fails to apply. We should emphasize that variation occurs only in a small subset of the Japanese lexicon: it affects mostly roots that occur in infrequent compounds.

In other words, in order to provide a full account for rendaku we need to make reference to phonological rules, i.e. grammar, as well as to frequency. Thus, *both* the frequency of particular items *and* phonological rules (or constraints) play a role in the application (or non-application) of this process. This result supports an integrated approach combining formal approaches to grammar and usage-based approaches (as advocated in Jackendoff (2007); see van de Weijer (2009; 2012), Sloos (2013), among others).

Moreover, the results show in what kinds of situation grammar and usage may interact, viz. in situations where words of extreme low frequency are involved. Due to the low frequency, uncertainty arises regarding the preferred output form, which gives room for competing forms. The frequency of these competing forms will eventually determine which candidate will be selected for production. In frequency studies, it is well known that high-frequency word often behave in an idiosyncratic way, due to an extremely strong mental representation, which prevents (analogical) change throughout time (e.g. Bybee (2002)). However, our results show that low-frequency words may also behave differently. Very recently, similar results were found for sound change in German (Sloos, 2012) and Dutch loanword integration in Indonesian (Sloos, 2013).

The exceptional behaviour of extremely infrequent words should be further investigated. Also, the rendaku puzzle is far from solved: location of accent is known to play a role as well in rendaku, which we did not investigate. The interaction between the location of accent and frequency seems to be a logical next step for further investigation.

References

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

BCCWJ. 2011. *Balanced Corpus of Contemporary Written Japanese*. Tokyo: NINJAL.

Bybee, Joan L. 2001. *Phonology and Language Use*. Cambridge and New York: Cambridge University Press.

Bybee, Joan L. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3): 261-290.

Bybee, Joan L. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4): 711-733.

Bybee, Joan L. & Hopper, Paul. 2001. Introduction to Frequency and the Emergence of Linguistic Structure. In Bybee, J. L. & Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure*. Amsterdam and Philadelphia: John Benjamins. 1-24.

Frellesvig, Bjarke. 2010. *A History of the Japanese Language*. Cambridge: Cambridge University Press.

Fukuda, Suzy E. & Fukuda, Shinji. 2000. The Operation of Rendaku in the Japanese Specifically Language-Impaired: A Preliminary Investigation. *Folia Phoniatrica et Logopaedica* 51(1-2): 36-54.

Gahl, Susanne. 2008. Time and Thyme are not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech. *Language* 84(3): 474-496.

Haraguchi, Shosuke. 2001. On Rendaku. *On'in Kenkyuu [Phonological Studies]* 4: 9-32.

Horton, Bruce & Minami, Sayoko. 2011. Vowel Length and Rendaku. *The Journal of Kanda University of International Studies* 23: 113-135.

Irwin, Mark. 2005. Rendaku-Based Lexical Hierarchies in Japanese: The Behaviour of Sino-Japanese Mononoms in Hybrid Noun Compounds. *Journal of East Asian Linguistics* 14(2): 121-153.

Irwin, Mark. 2009. Prosodic Size and Rendaku Immunity. *Journal of East Asian Linguistics* 18(3): 179-196.

Irwin, Mark. 2011. *Loanwords in Japanese*. Amsterdam and Philadelphia: John Benjamins.

Ito, Junko & Mester, R. Armin. 1986. The Phonology of Voicing in Japanese: Theoretical Consequences for Morphological Accessibility. *Linguistic Inquiry* 17(1): 49-73.

Ito, Junko & Mester, R. Armin. 2003. *Japanese Morphophonemics: Markedness and Word Structure*. Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 2007. A Whole Lot of Challenges for Linguistics. *Journal of English Linguistics* 35: 253-262.

Jong, Nivja H. de, Schreuder, Robert & Baayen, R. Harald. 2000. The Morphological Family Size Effect and Morphology. *Language and Cognitive Processes* 15(4-5): 329-365.

Kenstowicz, Michael. 2005. Paradigmatic Uniformity and Contrast. In Downing, L. J., Hall, T. A. & Raffelsiefen, R. (eds.), *Paradigms in Phonological Theory*. Berlin: Mouton. 145-169.

Kōjien. 1991. *Kojien Dictionary*. Tokyo: Iwanami Shoten.

Kubozono, Haruo. 2005. Rendaku: Its Domain and Linguistic Conditions. In van de Weijer, J. M., Nanjo, K. & Nishihara, T. (eds.), *Voicing in Japanese*. Berlin and New York: Mouton de Gruyter. 5-24.

Lyman, Benjamin S. 1894. *Change from Surd to Sonant in Japanese Compounds*. Philadelphia: Oriental Club of Philadelphia.

Martin, Samuel E. 1952. *Morphophonemics of Standard Colloquial Japanese*. Language Dissertations 47. Baltimore: Linguistic Society of America.

McCawley, James D. 1968. *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.

Miller, Roy A. 1967. *The Japanese Language*. Chicago: University of Chicago Press.

Ogawa, Taeko, Saito, Hirofumi & Yanase, Yoshinobu. 2005. Combinability of 2965 Kanji Characters of Japanese Industrial Standard on Formation of two-Kanji Compound Words. *Shinrigaku kenkyu [The Japanese Journal of Psychology]* 76(3): 269-275.

Ohno, Kazutoshi. 2000. The Lexical Nature of Rendaku in Japanese. *Japanese/Korean Linguistics* 9: 151-164.

Ohta, Satoshi. This volume. On the Relationship between Rendaku and Accent.

Otsu, Yukio. 1980. Some Aspects of Rendaku in Japanese and Related Problems. *Theoretical Issues in Japanese Linguistics. MIT Working Papers in Linguistics* 2: 207-227.

Phillips, Betty S. 2006. *Word Frequency and Lexical Diffusion*. New York: Palgrave.

Pierrehumbert, Janet B. 2006. The Next Toolkit. *Journal of Phonetics* 34(4): 516-530.

R, Core Development Team. 2005. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rosen, Eric R. 2001. Phonological Processes Interacting with the Lexicon: Variable and Non-Regular Effects in Japanese Phonology. PhD Thesis, University of British Columbia.

Shibatani, Masayoshi. 1990. *The Languages of Japan*. Cambridge: Cambridge University Press.

Sloos, Marjoleine. 2012. Chasing Shadows in the Alemannic Variety of Standard German. Ms, University of Groningen.

Sloos, Marjoleine. 2013. Phonological Grammar and Frequency: An Integrated Approach. Evidence from German, Indonesian, and Japanese. PhD Thesis, University of Groningen.

Takayama, Tomoaki. 2005. A Survey of Rendaku in Loanwords. In van de Weijer, J. M., Nanjo, K. & Nishihara, T. (eds.), *Voicing in Japanese*. Berlin and New York: Mouton de Gruyter. 177-190.

Tamaoka, Katsuo & Makioka, Shogo. 2004. Frequency of Occurrence for Units of Phonemes, Morae, and Syllables Appearing in a Lexical Corpus of a Japanese Newspaper. *Behavior Research Methods, Instruments and Computers* 36(3): 531-547.

Tanaka, Shin-ichi. 2003. Review of Rosen (2001). *Glot International* 7(5): 1-7.

Vance, Timothy J. 1987. *An Introduction to Japanese Phonology*. Albany, NY: State University of New York Press.

Vance, Timothy J. 2005. Rendaku in Inflected Words. In van de Weijer, J. M., Nanjo, K. & Nishihara, T. (eds.), *Voicing in Japanese*. Berlin and New York: Mouton de Gruyter. 89-103.

Vance, Timothy J. 2007. Have we Learned Anything about Rendaku that Lyman didn't Already Know? In Frellesvig, B., Shibatani, M. & Smith, J. C. (eds.), *Current Issues in the History and Structure of Japanese*. Tokyo: Kurosio. 153-185.

van de Weijer, Jeroen M. 2009. Optimality Theory and Exemplar Theory. *Phonological Studies* 12: 117-124.

van de Weijer, Jeroen M. 2012. *Grammar as Selection: Combining Optimality Theory and Exemplar Theory*. Nagoya: Kougaku Shuppan.

van de Weijer, Jeroen M., Nanjo, Kensuke & Nishihara, Tetsuo. 2005. *Voicing in Japanese*. Berlin and New York: Mouton de Gruyter.

Yamaguchi, Kyoko & Tanaka, Shin-ichi. This volume. Rendaku Variation in Deverbal Compounds.

Zamma, Hideki. 2005. The Correlation between Accentuation and Rendaku in Japanese Surnames: A Morphological Account. In van de Weijer, J. M., Nanjo, K. & Nishihara, T. (eds.), *Voicing in Japanese*. Berlin and New York: Mouton de Gruyter. 157-176.