Stochastic phonological knowledge and word formation in Japanese

Abstract

The question of whether linguistic knowledge is binary (i.e. grammatical vs. ungrammatical) or stochastic is one of the most important questions in general linguistic inquiry. Much recent work in the last few decades argues that phonological knowledge is stochastic (e.g. Hayes & Londe 2006). Building on this body of research, we show that in Japanese, gradient phonological knowledge affects several word formation patterns in stochastic ways. Concretely, we show that identity avoidance effects hold at both the segmental and the CV-moraic levels. These identity avoidance effects stochastically affect two types of word formation patterns in Japanese: group name formation and rendaku. We show that Maximum Entropy Grammar (Goldwater & Johnson 2003), together with multiple OCP constraints (Coetzee & Pater 2008), successfully models both of the observed morphological word formation patterns, without any further stipulation. In addition to this theoretical contribution, one of the patterns discussed in this paper—group name formation—has not been analyzed from the perspective of formal phonological theories before, and hence this paper has descriptive novelty as well.

1. Introduction

Whether linguistic knowledge is dichotomous/binary (grammatical vs. ungrammatical) or gradient is one of the most important questions in the current linguistic inquiry. At the outset of the generative enterprise, sentences were divided into two distinct sets: those that could be generated by the posited grammar (i.e. "grammatical sentences"), and those that could not be generated by the grammar (i.e. "ungrammatical sentences") (Chomsky 1957). In reality, however,

acceptability judgment patterns in syntax often show gradient patterns, as indicated by the common use of a set of different prefixal symbols (?, ??, ???, ?*,*?,*) used in sentential judgments in the syntactic literature (see e.g. Chomsky 1965; Lasnik 2004; Lasnik & Saito 1984; Pullum 2013a,b; Schütze 1996; Sprouse 2015, among many others). However, it is still debated whether syntactic knowledge itself operates on a dichotomous grammatical vs. ungrammatical distinction or not; some researchers argue that grammar/competence makes only a binary distinction (yes grammatical vs. no grammatical), and it is other cognitive processes—i.e. performance—that yields graded judgments (e.g. Neelman 2013; Schütze 1996; Sprouse 2007). Other researchers, like Bresnan & Hay (2008), Kellar (2006), Lasnik (2000), Lasnik & Saito (1984), Pullum (2013a,b), and Sorace & Keller (2005), accept the thesis that syntactic knowledge itself is gradient, and maintain that linguistic models should be able to capture this gradiency. Some specific proposals have been put forward to capture the gradient nature of syntactic knowledge, such as Linear Optimality Theory (Keller 2006) and Model Theoretic Syntax (Pullum 2013a,b).

As with generative syntax, generative phonology began with the assumption that phonological knowledge is binary; a famous example is that whereas *brick* and *blick* are well-formed in English, *bnick* is not (Halle 1978). One of the fundamental tenets of early generative phonology is that phonological grammar should be able to capture this binary, grammatical vs. ungrammatical, distinction between possible words and impossible words (rather than existing words and non-existing words). However, it has become increasingly clear that phonological knowledge is stochastic, not a simple matter of possible vs. impossible (see also Pierrehumbert 1997 and Cohn 2006 for historical reviews). First, phonotactic judgment patterns have now long been known to be stochastic; i.e. the intuition about whether a particular string can be a word or not is usually not a matter of a yes/no dichotomy. This gradient nature of phonotactic judgments was shown, for example, by the word-likeliness judgment experiment reported in Greenberg &

Jenkins (1964). For instance, native speakers of English tend to judge [klæb] to be more natural—or more "English-sounding"—than [kleb], although both forms should be both "grammatical" in English. It is also known that consonant clusters with sonority plateau (e.g. [bdif]) are judged by English speakers to be better than clusters with falling sonority (e.g. [lbif]), despite the fact that both types of clusters should be "ungrammatical" in English (Berent et al. 2007 *et seq*.). See Daland et al. (2011) for recent extensive results showing gradient phonotactic judgment patterns in English and relevant discussion on the gradient nature of phonotactic knowledge.

Another type of well-known case of gradient phonotactics is the pattern of similarity avoidance, found in many Semitic languages, in which pairs of similar adjacent consonants are underrepresented in their lexicon. In the similarity avoidance pattern, the more similar two paired consonants are, the less likely it is that that pair exists in the lexicon (Frisch et al. 2004). These sorts of gradient phonotactic identity avoidance effects have been observed in many languages, beyond Semitic languages, including, English (Berkley 1994), Muna (Coetzee & Pater 2008), Russian (Padgett 1992), and the native words in Japanese (Kawahara et al. 2006), among others (see also Alderete & Frisch 2007; Yip 1998; Zuraw & Lu 2009 for other cases of identity avoidance). In short, phonotactic distribution patterns, as well as native speakers' judgments on word-likeliness, are undoubtedly gradient, which cannot be reduced to a yes/no dichotomy. This observation led to the recent development of theories with numerically weighted constraints, such as Harmonic Grammar (Coetzee & Pater 2008) and MaxEnt Grammar (Hayes & Wilson 2008; Goldwater & Johnson 2003). Hayes & Wilson (2008: 382) explicitly declare that they "consider the ability to model gradient intuitions to be an important criterion for evaluating phonotactic models". Gradiency in phonotactics is now generally considered to be an essential aspect of grammar that any grammatical theory is required to capture, at least in phonology.

What has been less clear is whether phonological *alternations* can show systematic stochastic variations. However, recent work again demonstrates that some phonological alternations show patterned, stochastic variations (e.g. Boersma & Hayes 2001; Hayes 2017, Hayes & Londe 2006; McPherson & Hayes 2016; Moore-Cantwell & Pater 2016; Zuraw 2000, 2010). For example, Hayes & Londe (2006), in a paper titled "Stochastic phonological knowledge", have demonstrated that the probabilities of suffixes undergoing vowel harmony in Hungarian are different for different suffixes, and their likelihood of undergoing vowel harmony is affected by various phonological considerations. Zuraw (2000, 2010) shows that in Tagalog, different segments undergo nasal substitution with different probabilities in the lexicon, and that native speakers are sensitive to these gradient—yet regular—patterns, when they are tested with nonce words. These phonological patterns are not only *optional*, but *systematic* in the sense that their patterns make phonological sense (see Hayes 2017 for recent discussion). Although the issue of whether or not phonological alternations can be systematically stochastic may be less well-established than the issue of the gradient nature of phonotactics, in the last few decades, we have witnessed a growing body of evidence which suggests the stochastic nature of phonological alternation patterns. One theoretical impetus that drove our research is to add more case studies to address this question of whether phonological alternations can be gradient or not.

Against this theoretical background, this paper offers two new pieces of evidence for stochastic phonological knowledge from Japanese, both of which affect word formation patterns. To the best of our knowledge, the issue of stochastic phonological knowledge has not been seriously tested using Japanese (perhaps except by a few works such as Kawahara 2013, Kilbourn-Ceron & Sonderegger to appear, and Tanaka 2017). Moreover, the current paper shows that such patterns can be successfully analyzed using Maximum Entropy (MaxEnt) Grammar (e.g. Colavin et al. 2014; Goldwater & Johnson 2003; Hayes 2017, Hayes & Wilson 2008; Hayes et al. 2009; Hayes et al. 2012; Jäger & Rosenbach 2006; Kumagai 2017; Martin 2011; McPherson & Hayes

2016; Shih 2016; Shih & Inkelas 2016; Tanaka 2017; White to appear; Wilson 2006; Zhang et al. 2011; Zuraw & Hayes to appear), by positing multiple OCP constraints (Coetzee & Pater 2008). Again, this paper is one of the first attempts to fit a MaxEnt grammar to Japanese data (see also Tanaka 2017).[1]

The first case study, developed in section 2, deals with the formation of names of a group consisting of two members, which are created by combining the name of each member. As far as we know, this paper is the first attempt to describe and analyze this word formation pattern in the formal linguistic literature. Japanese speakers often make up a group name for a pair of people. For example, a group consisting of two identical twin sister actresses, *mana* and *kana*, is called *mana-kana*. The current project started with a simple question of why the group name is *mana-kana*, instead of *kana-mana*. Our hypothesis is that phonological considerations affect the formation of such group names. For example, ***k*a*n*a-*m*a*n*a** is disfavored because of the consecutive three CV-moras[2] with nasal onset. This is a reminiscent of the blockage of "-ly" adverb formation in English, in which "-ly" cannot be attached to those roots that already end with "ly" (e.g. **friendly-ly* and **silly-ly*: Katamba 1993). Shih (2014) likewise shows through a corpus study that in English names, name pairs are subject to a similar phonotactic restriction, in such a way that for example, *Josh Smith* [ʃ-s] is less likely to occur than *Jack Smith* [k-s] as a full name (see also Yip 1998 for other similar cases). Shih & Zuraw (to appear) show that avoidance

[1] One alternative for modeling gradient patterns is noisy Harmonic Grammar (see e.g. Coetzee & Kawahara 2013), which was implemented in Praat (Boersma 2001) as early as 2006. We do not intend to engage ourselves in the comparison between a MaxEnt analysis and other related frameworks in this paper. See Hayes & Wilson (2008: section 9.1) and Hayes (2017) for extended comparison between MaxEnt Grammar and other related constraint-based approaches.

[2] A CV-mora is a unit that plays an active role in Japanese phonology, orthography, speech production and perception (see, e.g., Ito 1989; Kubozono 1989; Labrune 2012; Otake et al. 1993). Since all the stimuli discussed in this paper are light syllables, CV-moras can be considered to be identical to light syllables. We use the term "CV-mora", following Kawahara & Sano (2016).

against a sequence of nasals can affect adjective-noun word order in Tagalog, which can be variable (e.g., *magandá* "beautiful" + *babáe* "woman" + -ng (LINK) → *magandá-ng babáe*/ *babáe-ng magandá* "beautiful woman"). Their corpus study shows that, when the nasal-initial linker *-ng* or *na* is inserted between an adjective and a noun, the word that follows it is more likely to begin with a non-nasal; for example, the order like *manggá-ng diláw* "mango-LINK yellow" is more frequent than the order like *diláw* ***na*** ***m****anggá* "yellow-LINK mango".

The experiment reported below in section 2 is designed to test the hypothesis that identity avoidance constraints are at work in determining the order of two elements. The results show that identity avoidance restrictions do indeed affect the group name formation patterns, although it is not the case that those names that violate an identity avoidance constraint are categorically prohibited. To model the results, we develop a MaxEnt analysis. We demonstrate that having multiple OCP constraints, following Coetzee & Pater (2008), successfully models the results without any further stipulation.

The second type of word formation that this paper explores in depth in section 3 is rendaku, which is a well-studied morphophonological process. Rendaku is the phenomenon in which the initial voiceless obstruents of the second member of a compound appear as voiced (e.g., /nise+**t**anuki/ → /nise+**d**anuki/ 'fake raccoon') (McCawley 1968; Tanaka 2017; Vance 1980, 1987, 2015; Vance & Irwin 2016, among many others; see Irwin 2016 for an extended bibliography list). We build upon the results of Kawahara & Sano (2016), who show that identity avoidance restrictions apply stochastically to the application of rendaku in nonce words. Kawahara & Sano (2016) demonstrated with a nonce-word experiment that the more similar the pairs of segments are that rendaku creates, the less likely rendaku is to apply. In one condition of their experiment, two consonants across the word boundary were identical after rendaku applies (e.g., schematically, /i**g**a+**g**omoke/ from /i**g**a/+/**k**omoke/); in the other condition, two CV moras across the word boundary were identical after rendaku applies (e.g., schematically,

/i**ga**+**ga**niro/ from /i**ga**/+/**ka**niro/). The results show that rendaku was less likely to occur when it resulted in consecutive identical consonants, compared to the control condition in which no identity violations were involved (e.g. forms like /i**g**a+**g**omoke/ are avoided); furthermore, the applicability of rendaku was even more reduced when rendaku resulted in adjacent identical CV-moras (e.g. forms like /i**ga**+**ga**niro/ are even more strongly avoided). Importantly, it is not the case that either of the identity avoidance constraints blocks rendaku entirely; they stochastically reduce the application probability of rendaku. As is the case with the group name formation, these results can be modeled by multiple OCP constraints and a MaxEnt grammar. This analysis supports the generality of the analysis that we develop in section 2.

To summarize, in this paper we show that empirically, phonological knowledge can stochastically and systematically affect Japanese word formation patterns beyond a dichotomous grammatical vs. ungrammatical distinction, and that theoretically, a MaxEnt grammar is a useful tool with which to model that stochastic knowledge. We also emphasize the descriptive value of what we report in section 2, which has hitherto not been analyzed in the theoretical literature.

2. Group name formation in Japanese

2.1. Background

This section explores the compound formation pattern of group names in which two names are combined. As mentioned in section 1, the pair of Japanese identical twin sister actresses, *mana* and *kana*, is called *mana-kana*. Another example is a pair of two Japanese Ping-Pong players, *mima* and *miu*, which is *miu-mima, not *mima-miu*. In both of these examples, the possible-yet-unattested forms—*ka**n**a*#***m**a**n**a*** and ***m**i**m**a*#***m**iu*—contain three onset nasal consonants across the word boundary, whereas the attested examples—***m**a**n**a*#*ka**n**a*** and ***m**i*_*u*#***m**i**m**a***—contain no sequence of onset nasal consonants across the word boundary.[3] In the rest of the

[3] We assume that the vowel sequence of [iu], with no fall in sonority, is syllabified separately

section, the sequence of nasals is referred to as nasal clash (cf. "stress clash": Prince 1983). We experimentally examine whether nasal clash generally affects compound formation patterns in Japanese. We also examine whether degrees of similarity (e.g. /m/-/m/ vs. /m/-/n/) matter. The previous studies (e.g. Coetzee & Pater 2008; Frisch et al. 2004; Kawahara & Sano 2016) have shown that the more similar sequences are, the more strongly they are disfavored; hence it is predicted that the degrees of similarity should impact the Japanese group name formation pattern as well. On the other hand, in some languages, total identity has been found to provide "an escape hatch" for similarity avoidance restrictions (e.g. Berent & Shimron 1997; Frisch et al. 2004), and hence it may be the case that an /m/-/m/ pair may be favored over an /m/-/n/ pair. This is an empirical question, which remains unsettled in the phonology of Japanese (though see Kawahara et al. 2006 and Kawahara & Sano 2016 for some discussion).

Going beyond the segmental level, we also test the identity effects in the CV-mora. Recall that in Kawahara & Sano's (2016) experiment, rendaku was more likely to be blocked when it resulted in CV moraic identity (e.g. *[…**ga**-**ga**…]) than when it resulted in mere consonantal identity (e.g. *[**g**a…**g**o]). Therefore, Japanese speakers may disfavor a sequence of two identical CV moras in general, which may affect the group name formation as well.

Although an inquiry into the nasal clash effect—more generally, the effect of similarity avoidance—is the main focus of this paper, there is another phonological factor that went into the consideration of this experiment, which is sonority (e.g. Clements 1990; Kenstowicz 1994; Parker 2002, 2011)—in the general sonority hierarchy, although some details are debated, segments are ordered in the following order: stop < fricative < nasal < liquid < glide. In English, when two words are combined with *and*, the word with the more sonorous onset tends to come

as [i.u]. The onsetless nature of the second syllable is represented by "_" in the text. See Kubozono (2015) for extensive discussion on Japanese diphthong and hiatus.

first.[4] Some existing examples include, for example, ***l**ovey-**d**ovey*, ***w**alkie-**t**alkie*, and ***w**illy-**n**illy* (Parker 2002: 246). Parker (2002) further experimentally examined this tendency, by preparing several pairs of compounds such as *weeby-leeby* and *leeby-weeby.* The results showed that *weeby-leeby* was indeed preferred to *leeby-weeby*, which suggests that English speakers prefer to have the word with more sonorous consonant at the beginning of the derived word.[5] Given this observation, we needed to make sure that the preference for *mana-kana* over *kana-mana* does not (solely) come from some sonority-based preference, instead of avoidance of consecutive nasal onset consonants; it could be the case that Japanese speakers, just like English speakers, may order names in such a way that more sonorous consonants are placed word-initially, which prefers *mana-kana* over *kana-mana*, although this sonority-based theory cannot explain the *miu-mima* example.

To summarize, in this experiment, we examine whether various similarity-related factors affect word formation patterns in Japanese; in particular, (i) whether nasal clash is avoided, and if so, (ii) whether the number of nasal clashes matters, (iii) whether consonantal identity and moraic identity show different degrees of influence, and in addition, (iv) whether, as with English, sonority matters when speakers combine two words to make a larger word. In what follows, we express general nasal clash as the effects of OCP(nasal), nasal clash with identical nasal consonants as OCP(C), and nasal clash in identical CV moras as OCP(CV), respectively (where OCP=Obligatory Contour Principle: Leben 1973; McCarthy 1986).

[4] There are several studies of sonority effects on blend formation in other languages. Bat-El (1996) discusses the role of sonority in blend formation in Hebrew. Likewise, Labrune (2006) suggests that similar tendency may be observed in Japanese blending formation.

[5] While Parker (2002) has shown that sonority is one key fact that affects binomial ordering, it is not the only factor that affects word ordering in English binomials. See also Benor & Levy (2006), Mollin (2012), and Lohmann (2014) for recent corpus-based surveys on English binomial orderings.

2.2. Stimuli

The current experiment used disyllabic Japanese girls' names as stimuli. All of the names used were existing (or at least possible) names.[6] Sets 1 and 3 consisted of pairs that could result in two nasals in sequence, either non-identical (e.g., *ha**n**a-**m**oka*), or identical (e.g. *ha**n**a-**n**iko*). Sets 2 and 4 consisted of pairs that could result in three nasals in sequence (e.g., *ha**n**a-**m**o**n**a* and *ku**m**i-**m**i**n**a*).

Table 1: The overall stimulus structure

	Number of nasals	Non-nasal segment
Set 1	2	obs
Set 2	3	obs
Set 3	2	son
Set 4	3	son

The number of nasal consonants involved in nasal clash was included as a condition in the experiment, because in the *mana-kana* and *miu-mima* example, it may be three consecutive nasal onset consonants that make the unattested *ka**n**a-**m**a**n**a* and ***m**i**m**a-**m**iu* unviable options; we were interested in whether two consecutive nasal onset consonants were enough to affect group name formation patterns.

Sets 1 and 2 consisted of pairs in which one word begins with an obstruent and the other word begins with a nasal (e.g., *ha**n**a* and ***m**oka*),[7] and Sets 3 and 4 consisted of pairs in which one word begins with a liquid, and the other word begins with a nasal (e.g., *ri**n**a* and ***m**oka*). Recall that we wanted to tease apart the effects of identity avoidance and the effects of sonority.

[6] The disyllabic names used in the stimuli all have initial accent, and hence the stimuli are controlled in this respect. Whether Japanese accent, like English stress (Lohmann 2014), affects binomial ordering in Japanese is an interesting question for future research.

[7] Some consider /h/ to be a voiceless approximant; i.e., a voiceless sonorant (Chomsky & Halle 1968). We follow other work (Lass 1976: 64–68, Jaeger & Ohala 1984, Sagey 1986, Parker 2002) demonstrating that /h/ is a voiceless fricative.

Within each set, there were three conditions that were characterized in terms of different OCP violation profiles (i.e., OCP(nasal); OCP(nasal)+OCP(C); OCP(nasal)+OCP(CV)). In Set 1, shown in Table 2, the first syllable of one word had a nasal onset, and the second syllable of the other word had a nasal onset (e.g. ***m**oka* and *ha**n**a*). The word that did not begin with a nasal began with an obstruent (e.g. ***h**ana*). The condition in Table 2a was used to test whether the violation of OCP(nasal) is avoided. If *moka-hana* is preferred over *ha**n**a-**m**oka*, it would indicate that nasal clash (i.e., …***n**a-**m**o*…) is avoided. The condition in Table 2b was used to test the effects of identical consonants, in addition to the occurrence of two nasals; i.e. the effects of OCP(C). Given *niko* and *hana*, *ha**n**a-**n**iko* has a sequence of identical nasals (i.e., …***n**a-**n**i*…), thus violating OCP(C) in addition to OCP(nasal). The condition in Table 2c was used to test the OCP(CV), in addition to the OCP(C) and OCP(nasal). If *natu-hana* is favored over *ha**n**a-**n**atu*, an identical mora across the word boundary (i.e., ...***n**a-**n**a*...) may be being avoided. There are four possible combinations for each condition, and thus Set 1 consists of 12 combinations in total, as in Table 2.

Table 2: Set 1: 2 nasals (M = /m/; N = /n/; O = an obstruent; R = a sonorant)

	α	+	*β*	→	*α-β or β-α*
a.	*moka* (MO)	+	*hana* (ON)	→	***m**oka-ha**n**a* (MOON) or *ha**n**a-**m**oka* (ONMO)
	moka (NO)	+	*kana* (ON)	→	***m**oka-ka**n**a* (MOON) or *ka**n**a-**m**oka* (ONMO)
	natu (NO)	+	*kumi* (OM)	→	***n**atu-ku**m**i* (NOOM) or *ku**m**i-**n**atu* (OMNO)
	natu (NO)	+	*fumi* (OM)	→	***n**atu-fu**m**i* (NOOM) or *fu**m**i-**n**atu* (OMNO)
b.	*niko* (NO)	+	*hana* (ON)	→	***n**iko-ha**n**a* (NOON) or *ha**n**a-**n**iko* (ONNO)
	niko (NO)	+	*kana* (ON)	→	***n**iko-ka**n**a* (NOON) or *ka**n**a-**n**iko* (ONNO)
	moka (MO)	+	*kumi* (OM)	→	***m**oka-ku**m**i* (MOOM) or *ku**m**i-**m**oka* (OMMO)
	moka (MO)	+	*fumi* (OM)	→	***m**oka-fu**m**i* (MOOM) or *fu**m**i-**m**oka* (OMMO)
c.	*natu* (NaO)	+	*hana* (ONa)	→	***n**atu-ha**n**a* (NaOONa) or *ha**n**a-**n**atu* (ONaNaO)
	natu (NaO)	+	*kana* (ONa)	→	***n**atu-ka**n**a* (NaOONa) or *ka**n**a-**n**atu* (ONaNaO)
	mika (MiO)	+	*kumi* (OMi)	→	***m**ika-ku**m**i* (MiOOMi) or *ku**m**i-**m**ika* (OMiMiO)
	mika (MiO)	+	*fumi* (OMi)	→	***m**ika-fu**m**i* (MiOOMi) or *fu**m**i-**m**ika* (OMiMiO)

Set 2, shown in Table 3, was prepared to examine whether three consecutive nasals would be avoided more strongly than two consecutive nasals. Sequences with different OCP violation profiles were also examined, as in Set 1. The nasal clash in Table 3a violates only OCP(nasal),

the nasal clash in Table 3b violates OCP(nasal) and OCP(C), and the nasal clash in Table 3c violates all of OCP(nasal), OCP(C), and OCP(CV).

Table 3: Set 2: 3 nasals (M = /m/; N = /n/; O = an obstruent; R = a sonorant)

	α	+	*β*	→	*α-β or β-α*
a.	*mona* (MN)	+	*hana* (ON)	→	***mo**n**a**-ha**n**a* (MNON) or *ha**n**a-**mo**n**a* (ONMN)
	mona (MN)	+	*kana* (ON)	→	***mo**n**a**-ka**n**a* (MNON) or *ka**n**a-**mo**n**a* (ONMN)
	nami (NM)	+	*kumi* (OM)	→	***na**m**i**-ku**m**i* (NMOM) or *ku**m**i-**na**m**i* (OMNM)
	nami (NM)	+	*fumi* (OM)	→	***na**m**i**-fu**m**i* (NMOM) or *fu**m**i-**na**m**i* (OMNM)
b.	*nina* (NN)	+	*hana* (ON)	→	***ni**n**a**-ha**n**a* (NNON) or *ha**n**a-**ni**n**a* (ONNN)
	nina (NN)	+	*kana* (ON)	→	***ni**n**a**-ka**n**a* (NNON) or *ka**n**a-**ni**n**a* (ONNN)
	mona (MN)	+	*kumi* (OM)	→	***mo**n**a**-ku**m**i* (MNOM) or *ku**m**i-**mo**n**a* (OMMN)
	mona (MN)	+	*fumi* (OM)	→	***mo**n**a**-fu**m**i* (MNOM) or *fu**m**i-**mo**n**a* (OMMN)
c.	*nami* (NaM)	+	*hana* (ONa)	→	***na**m**i**-ha**n**a* (NaMONa) or *ha**n**a-**na**m**i* (ONaNaM)
	nami (NaM)	+	*kana* (ONa)	→	***na**m**i**-ka**n**a* (NaMONa) or *ka**n**a-**na**m**i* (ONaNaM)
	mina (MiN)	+	*kumi* (OMi)	→	***mi**n**a**-ku**m**i* (MiNOMi) or *ku**m**i-**mi**n**a* (OMiMiN)
	mina (MiN)	+	*fumi* (OMi)	→	***mi**n**a**-fu**m**i* (MiNOMi) or *fu**m**i-**mi**n**a* (OMiMiN)

In Sets 3 and 4, shown in Table 4 and Table 5, the word listed in β begins with a sonorant rather than an obstruent. If there is a sonority-driven word-ordering preference in Japanese, we would expect to observe different results between Sets 1 & 2 on the one hand, and Sets 3 & 4 on the other.

Table 4: Set 3: 2 nasals (M = /m/; N = /n/; O = an obstruent; R = a sonorant)

	α	+	*β*	→	*α-β or β-α*
a.	*moka* (MO)	+	*rina* (RN)	→	***mo**ka-ri**n**a* (MORN) or *ri**n**a-**mo**ka* (RNMO)
	moka (MO)	+	*rena* (RN)	→	***mo**ka-re**n**a* (MORN) or *re**n**a-**mo**ka* (RNMO)
	natu (NO)	+	*rumi* (RM)	→	***n**atu-ru**m**i* (NORM) or *ru**m**i-**n**atu* (RMNO)
	natu (NO)	+	*remi* (RM)	→	***n**atu-re**m**i* (NORM) or *re**m**i-**n**atu* (RMNO)
b.	*niko* (NO)	+	*rina* (RN)	→	***n**iko-ri**n**a* (NORN) or *ri**n**a-**n**iko* (RNNO)
	niko (NO)	+	*rena* (RN)	→	***n**iko-re**n**a* (NORN) or *re**n**a-**n**iko* (RNNO)
	moka (MO)	+	*rumi* (RM)	→	***mo**ka-ru**m**i* (MORM) or *ru**m**i-**mo**ka* (RMMO)
	moka (MO)	+	*remi* (RM)	→	***mo**ka-re**m**i* (MORM) or *re**m**i-**mo**ka* (RMMO)
c.	*natu* (NaO)	+	*rina* (RNa)	→	***n**atu-ri**n**a* (NaORNa) or *ri**n**a-**n**atu* (RNaNaO)
	natu (NaO)	+	*rena* (RNa)	→	***n**atu-re**n**a* (NaORNa) or *re**n**a-**n**atu* (RNaNaO)
	mika (MiO)	+	*rumi* (RMi)	→	***mi**ka-ru**m**i* (MiORMi) or *ru**m**i-**mi**ka* (RMiMiO)
	mika (MiO)	+	*remi* (RMi)	→	***mi**ka-re**m**i* (MiORMi) or *re**m**i-**mi**ka* (RMiMiO)

Table 5: Set 4: 3 nasals (M = /m/; N = /n/; O = an obstruent; R = a sonorant)

	α	+	*β*	→	*α-β or β-α*
a.	*mona* (MN)	+	*rina* (RN)	→	***mo**n**a*-*ri**n**a*** (MNRN) or ***ri**n**a*-*mo**n**a*** (RNMN)
	mona (MN)	+	*rena* (RN)	→	***mo**n**a*-*re**n**a*** (MNRN) or ***re**n**a*-*mo**n**a*** (RNMN)
	nami (NM)	+	*rumi* (RM)	→	***na**m**i*-*ru**m**i*** (NMRM) or ***ru**m**i*-*na**m**i*** (RMNM)
	nami (NM)	+	*remi* (RM)	→	***na**m**i*-*re**m**i*** (NMRM) or ***re**m**i*-*na**m**i*** (RMNM)
b.	*nina* (NN)	+	*rina* (RN)	→	***ni**n**a*-*re**n**a*** (NNRN) or ***re**n**a*-*ni**n**a*** (RNNN)
	nina (NN)	+	*rena* (RN)	→	***ni**n**a*-*re**n**a*** (NNRN) or ***re**n**a*-*ni**n**a*** (RNNN)
	mona (MN)	+	*rumi* (RM)	→	***mo**n**a*-*ru**m**i*** (MNRM) or ***ru**m**i*-*mo**n**a*** (RMMN)
	mona (MN)	+	*remi* (RM)	→	***mo**n**a*-*re**m**i*** (MNRM) or ***re**m**i*-*mo**n**a*** (RMMN)
c.	*nami* (NaM)	+	*rina* (RNa)	→	***na**m**i*-*ri**n**a*** (NaMRNa) or ***ri**n**a*-*na**m**i*** (RNaNaM)
	mina (MiN)	+	*remi* (RMi)	→	***mi**n**a*-*re**m**i*** (MiNRM) or ***re**m**i*-*mi**n**a*** (RMiMiN)
	mina (MiN)	+	*rumi* (RMi)	→	***mi**n**a*-*ru**m**i*** (MiNRM) or ***ru**m**i*-*mi**n**a*** (RMiMiN)
	mina (MiN)	+	*remi* (RMi)	→	***mi**n**a*-*re**m**i*** (MiNRM) or ***re**m**i*-*mi**n**a*** (RMiMiN)

2.3. Participants and Procedure

A total of 83 naive native speakers of Japanese participated in the experiment. All of the participants were undergraduate students at a Japanese university. There is no overlap of the participants between the current experiment and the one reported in section 3. In the instruction session, they were told that they would make up a group name for a pair of girls. In the test session, they were first given two names, and then were asked to choose one of the two combined forms (e.g., given two personal names, *mana* and *kana*, which order would you use to make up a group name, *mana-kana* or *kana-mana*?). All the names were written in the Japanese *katakana* orthography, which is commonly used to write personal names. There were a total of 48 questions (4 sets*12 combinations). The order of the questions was randomized.

2.4. Results

For statistical analysis, a generalized mixed-effects logistic regression was fit to the response using the `glmer` function in *R* (e.g., Baayen 2008). Subjects and items were coded as random effects. The first model included all the fixed factors (obs vs. son; two nasals vs. three nasals; OCP(C); OCP(CV)); follow-up specific comparisons were made based on contrast analyses using more specific logistic regression models. The resulting figures below show the ratios of the responses that contain nasal clash on the y-axis. The results of Set 1 and Set 2 are shown in Figure 1, and those of Set 3 and Set 4 are shown in Figure 2. Error bars represent 95% confidence intervals.

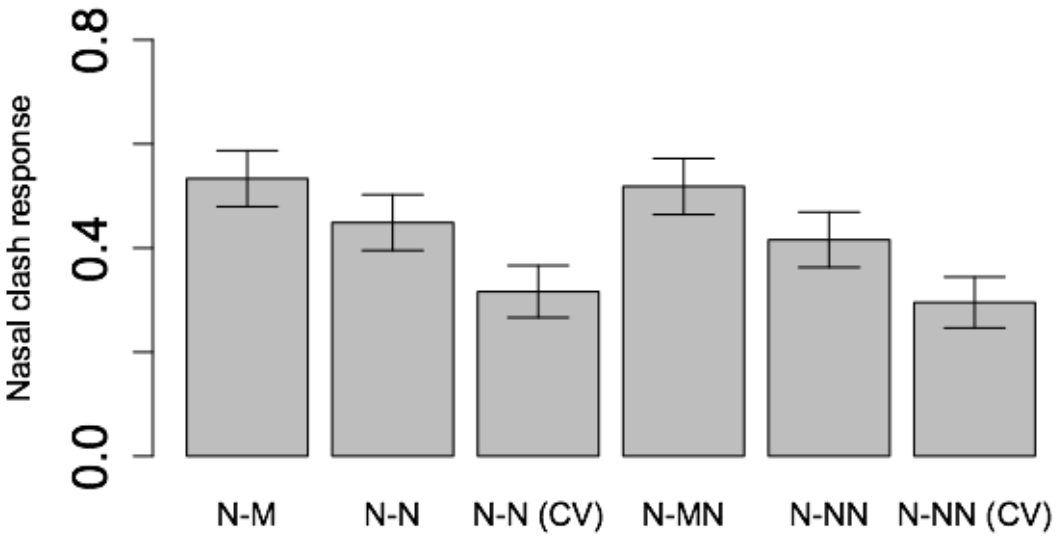

Figure 1: Nasal clash response ratio with 95% confidence intervals. Cases in which the words that do not begin with a nasal begin with an obstruent.

In Figure 1, the first three bars show cases in which *two* nasals are placed in adjacent syllables (e.g. *ha**n**a-**m**oka*), whereas the last three bars show cases in which *three* nasals are placed in proximity (*ha**n**a-**m**o**n**a*). Within each set, the three bars are ordered in the order of degrees of similarity (non-identical nasals (N-M), identical nasals (N-N), identical CV moras with a nasal

onset (N-N (CV)). The actual observed average values are: 0.53 vs. 0.45 vs. 0.32 for the first three bars and 0.52 vs. 0.42 vs. 0.30 for the last three bars. For the two-nasal condition (the leftmost three bars), there were significant differences between each condition: N-M vs. N-N, z = -2.366, p < .05; N-M vs. N-N (CV), z = -6.035, p < .001; N-N vs. N-N (CV), z = -3.874, p < .001. The same holds true of three-nasal condition (the rightmost three bars) (N-MN vs. N-NN, z = -2.885, p < .01; N-MN vs. N-NN (CV), z = -6.245, p < .001; N-NN vs. N-NN (CV), z = -3.618, p < .001). We thus observe a clear tendency for the more similar sequences to be avoided more strongly. It is important to note here that the effects are gradient; we see a three-way distinction, according to different violation profiles of OCP constraints. We maintain that this instantiates the effect of gradient phonological knowledge that affects the group name formation pattern.

There were no effects of the number of nasal consonants involved; i.e. there were no differences between the first three bars and the last three bars (z = *1.12*, *n.s.*). Finally, looking at the two N-M(N) conditions, the nasal clash response ratios are over 0.5 (i.e. 0.53 and 0.52); i.e. slightly higher than expected by chance. This may indicate that the avoidance of non-identical nasal consonants—OCP(nasal)—is not so strong as to show tangible effects in this experiment. The weak effect of OCP(nasal) will be made clearer in the MaxEnt analysis that is presented below, in which the weight of OCP(nasal) is low. As we will observe below, there may be a preference to put less sonorous consonants word-initially (Smith 2002), which would coerce nasal clash in this condition; i.e. ***h****ana-moka* is better than ***m****oka-hana* in that the former has an obstruent word-initially. This sonority-based effect may have "cancelled out" the effects of OCP(nasal).

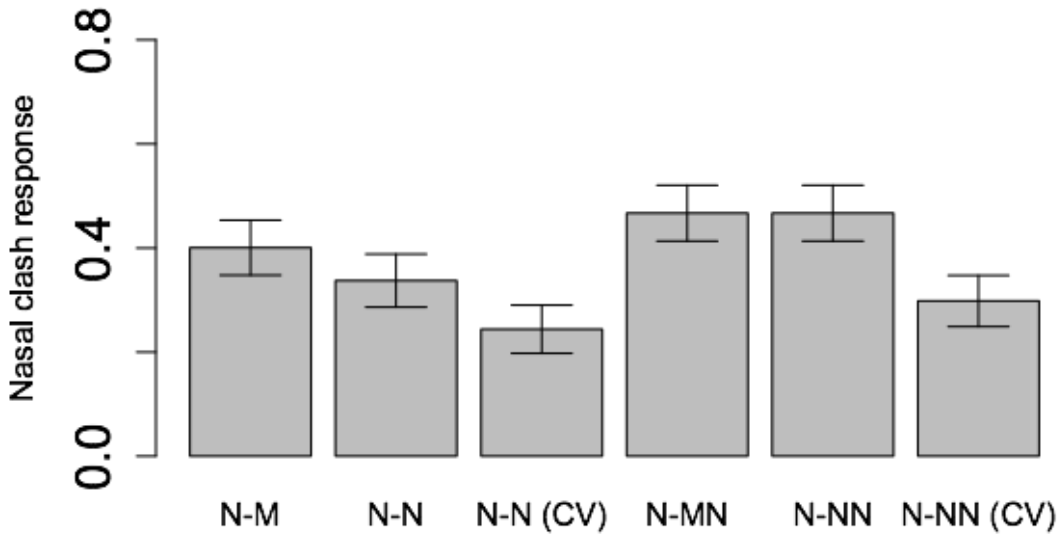

Figure 2: Nasal clash response ratios with 95% confidence intervals. Cases in which the words that do not begin with a nasal begin with a sonorant.

The first three bars in Figure 2 show the two-nasal condition, in which there were significant differences between N-M and N-N (CV) (z = -4.663, p < .001) and between N-N and N-N (CV) (z = -2.944, p < .01) (0.40 vs. 0.34 vs. 0.24). Though the difference between N-M and N-N did not reach significance (z = -1.852, *n.s.*), it is in the expected direction. For the three-nasal condition (the rightmost three bars), there were also significant differences between N-MN and N-NN (CV) (z = -4.919, p < .001) and between N-NN and N-NN (CV) (z = -4.956, p < .001). However, there were no significant differences between N-MN and N-NN (z = 0, *n.s.*) (0.47 vs. 0.47 vs. 0.30); there were no obvious effects of OCP(C) in this context. Surprisingly, there were slightly more nasal clash responses when there were three nasal consonants than when there were only two nasal consonants (z = 2.087, p < .05) (here but not in Figure 1). We do not have a clear explanation of these unexpected results.

346 Comparing Figure 1 (the obstruent condition) and Figure 2 (the sonorant condition), the pro-
347 portion of nasal clash is on average lower in the sonorant conditions than in the obstruent con-
348 dition (z = 3.189, p < .01). This difference shows that Japanese speakers are more likely to
349 tolerate nasal clash when it results in word-initial obstruents (e.g. ***h**ana-mona*) than when it
350 results in word-initial liquid (e.g. ***r**ina-mona*). There are two possible reasons for this difference.
351 The first possibility is that /r/ is avoided as a word-initial sound. This hypothesis is possible, as
352 there are few Japanese native words that begin with /r/ (e.g., Labrune 2014), and hence there
353 can be a constraint like *INITIAL-/r/ at work in Japanese phonology (Kawahara 2015). The se-
354 cond possibility is that, as was the case for Parker's (2002) experiments with English speakers,
355 the ordering of the two words was affected by sonority considerations: obstruent-initial words
356 were preferred to come before nasal-initial words, and nasal-initial words were preferred to
357 come before liquid-initial words (see Smith 2002 for related observations). Under this interpre-
358 tation, while English prefers more sonorous word-initial segments, Japanese prefers less sono-
359 rous word-initial segments. In the analysis that follows, we adapt the second explanation, be-
360 cause it explains why there were no clear effects of OCP(nasal) in Figure 1.[8] With the second
361 explanation, we can assume that the sonority preference and OCP(nasal) canceled each other
362 out, resulting in the near-chance performance.
363
364 To summarize, the results indicate that when Japanese speakers are asked to make a group name
365 based on two names, various factors affect the ordering; (i) sequences of two identical nasals
366 are avoided; (ii) sequences of identical CV-moras are avoided even more; (iii) a word with a
367 lower sonority consonant is preferred to be placed word-initially. As we will see, each of these
368 factors can be represented by phonological constraints, and a MaxEnt analysis is suitable to
369 model the overall results.

[8] Of course, it is possible to tease apart these two hypotheses empirically by using glide initial words in place of /r/-initial words. In our experiment, however, we found it hard to have an enough number of glide-initial, disyllabic, girls' names in Japanese.

370

371 **2.5. A MaxEnt analysis**

372 To model the stochastic nature of the Japanese name ordering patterns observed in the experi-
373 ment above, we used a MaxEnt grammar model (Hayes & Wilson 2006). MaxEnt is similar to
374 Optimality Theory (OT: Prince & Smolensky 2004) in that a set of candidates are evaluated
375 against a set of constraints. Unlike OT, however, the constraints are weighted (rather than
376 ranked), as in Harmonic Grammar (HG: Legendre et al. 1990, 2006; Pater 2009, 2016; Potts et
377 al. 2010). The probabilities of each candidate are assigned based on their constraint violation
378 profiles. More specifically, for each candidate, weighted constraint violations are summed,
379 which is its H(armonic)-score. H-score is mapped to probabilities in such a way that
380 $P(cand_i)=exp(-H(cand_i))$, relativized to all the other candidates so that all probabilities sum to
381 1.

382

383 The procedure of calculating probabilities is as follows (Hayes 2017; Hayes et al. 2009; Hayes
384 et al. 2012; Zuraw & Hayes to appear, and Hayes & Wilson 2008, in particular):

385

386 1) Like HG, for each candidate, harmonic score (H-score) is calculated as the sum of C_i*w_i,
387 where the candidate's violation of each constraint (C_i) is multiplied by its weight (w_i);
388 2) Each candidate's "bare" probability is calculated as $e^{-(\text{H-score})}$;
389 3) $e^{-(\text{H-score})}$ is summed over all candidates;
390 4) P(x), the predicted probability of candidate x, is its $e^{-(\text{H-score})}$ divided by the sum of the $e^{-(\text{H-}}$
391 $^{\text{score})}$ of all of the candidates.

392

393 We used the MaxEnt Grammar Tool (Hayes et al. 2009) to implement the analysis, which cal-
394 culates optimal weights for each constraint, given the frequency distributions of actual out-
395 comes. To implement the MaxEnt analysis, we use the following four constraints. First,
396 *SON(C_2) > SON(C_1) disfavors forms in which the second word begins with a less sonorous

consonant than the first word (e.g., /m/ > /h/ in ***m***ona#***h***ana; /r/ > /m/ in ***r***ina#***m***ona). Second, OCP(nasal) is a constraint that is violated by two consecutive nasal consonants across a word boundary (e.g., ha***n***a#***m***ona; ri***n***a#***m***ona).[9] Since the experimental results did not generally show a substantial difference between sequences of two nasals and sequences of three nasals, their violation profiles were not distinguished. Third, OCP(C) is violated if the two nasals across the word boundary are identical (e.g., ku***m***i#***m***ona; ru***m***i#***m***ona). Fourth, OCP(CV) is violated if there is a pair of adjacent identical CV-moras (e.g., ha***na***#***na***mi; ri***na***#***na***mi). The violation profiles of these constraints as well as the candidate sets fed to the MaxEnt Grammar Tool are shown in (1) and (2).

Table 6 shows the results of constraint weight that we obtained by the MaxEnt Grammar Tool. Each MaxEnt analysis is given in (1) and (2), respectively.[10] (3) and (4) compare the observed probabilities obtained in the experiment with the predicted probabilities by the MaxEnt Tool. We observe that the two probabilities are highly correlated, indicating the success of the MaxEnt analysis.

[9] Since two consecutive nasal consonants within a word (e.g. ***m***o***n***a) are shared by compared candidates, they can be ignored in our tableaux thanks to Cancellation Lemma (Prince and Smolensky 1993/2004).

[10] The harmonic scores of candidates can be used to model acceptability judgments as well (e.g., Coetzee & Pater 2008). The idea is that, provided that the optimal candidate of each candidate set has the same violation profile, the lower a candidate's harmonic-score is *across* candidate sets, the more unlikely it is to be considered acceptable. To take the analysis in (1) for example, we can predict that ha***n***a#***m***ona (= -0.082) is the most harmonic, ha***na***#***na***mi (= -0.924) is the least harmonic, and ku***m***i#***m***ona (= -0.345) is in-between; as a result, ha***na***#***na***mi is judged to be less acceptable than ku***m***i#***m***ona and also that ku***m***i#***m***ona is less acceptable than ha***n***a#***m***ona.

413 Table 6: Constraints used, and their weights obtained by the MaxEnt Grammar Tool

Constraints	Weight
*SON(C_2) > SON(C_1)	0.11
OCP (nasal)	0.082
OCP (C)	0.263
OCP (CV)	0.579

414

415 (1) MaxEnt analysis (the obstruent condition)

	*S(C_2) > S(C_1)	OCP (nasal)	OCP (C)	OCP (CV)				
weights	0.11	0.082	0.263	0.579		H-score	$e^{-(\text{H-score})}$	Predicted Prob.
mona + (hana/kana)								
mona # (hana/kana)	-1					-0.11	0.8958	0.493
(hana/kana) # mona		-1				-0.082	0.9213	0.507
mona + (kumi/fumi)								
mona # (kumi/fumi)	-1					-0.11	0.8958	0.5585
(kumi/fumi) # mona		-1	-1			-0.345	0.7082	0.4415
nami + (hana/kana)								
nami # (hana/kana)	-1					-0.11	0.8958	0.693
(hana/kana) # nami		-1	-1	-1		-0.924	0.3969	0.307

416

417 (2) MaxEnt analysis (the sonorant condition)

	*S(C_2) > S(C_1)	OCP (nasal)	OCP (C)	OCP (CV)				
weights	0.11	0.082	0.263	0.579		H-score	$e^{-(\text{H-score})}$	Predicted Prob.
mona + (rina/rena)								
mona # (rina/rena)						0	1	0.5479
(rina/rena) # mona	-1	-1				-0.192	0.8253	0.4521
mona + (rumi/remi)								
mona # (rumi/remi)						0	1	0.6118
(rumi/remi) # mona	-1	-1	-1			-0.455	0.6344	0.3882
nami + (rina/rena)								
nami # (rina/rena)						0	1	0.7377
(rina/rena) # nami	-1	-1	-1	-1		-1.034	0.3556	0.2623

418

(3) Set 1 & 2

Forms	Observed P	Predicted P
mona # (hana/kana)	0.47	0.49
(hana/kana) # mona	0.53	0.51
mona # (kumi/fumi)	0.57	0.56
(kumi/fumi) # mona	0.43	0.44
nami # (hana/kana)	0.70	0.69
(hana/kana) # nami	0.30	0.31

(4) Set 3 & 4

Forms	Observed P	Predicted P
mona # (rina/rena)	0.57	0.55
(rina/rena) # mona	0.43	0.45
mona # (rumi/remi)	0.60	0.61
(rumi/remi) # mona	0.40	0.39
nami # (rina/rena)	0.73	0.74
(rina/rena) # nami	0.27	0.26

2.6 Summary

In this section, we examined the group-name formation pattern in Japanese, in which two names are combined together to form a group name. We observed that similarity avoidance plays a visible role in this word formation in such a way that similarity at the word boundary is avoided, and the higher the similarity, the more strongly it is disfavored. In particular, sequences of two nasals and sequences of CV-moras with two identical nasals were particularly disfavored. Importantly, however, no phonological constraints were deterministic, i.e. inviolable. They simply reduced the probability of nasal clash. In this sense, identity avoidance constraints stochastically affect the word formation pattern. We modeled these gradient patterns using a MaxEnt grammar as well as different types of OCP constraints. We also found that Japanese speakers may prefer to have less sonorous consonant word-initially. Although this preference toward lower sonority has been known cross-linguistically (Smith 2002), we believe that it is a new finding for Japanese.

3. Rendaku as evidence for stochastic phonological knowledge

3.1. Identity Avoidance in rendaku

We next turn to an analysis of another word formation pattern, rendaku, which shows another case of the stochastic and systematic influences of identity avoidance constraints on a word formation pattern. This section analyzes the experimental data presented by Kawahara & Sano (2016), in order to show the generality of the constraints and analysis developed in section 2.

Before delving into the analysis, we first briefly review the experiment design and results.

The purpose of Kawahara & Sano (2016) was to examine whether identity avoidance blocks rendaku application. The set of stimuli in Table 7 was used to test the effects of identity avoidance at the consonantal level (i.e., OCP(C)), and the set in Table 8 was used to test the effect of identity avoidance at the CV-moraic level (i.e., OCP(CV)). In each set, their stimuli contained four first elements (E1s) and three different second elements (E2s) for each consonant /k, t, s, h/ that potentially undergoes rendaku, which yielded 12 E2s for each E1. There were thus 48 combinations in total.

Table 7: The list of the stimuli used in Set 1. All combinations of E1 and E2 (4 * 12 =48) were tested. E2 were nonce words.

E1		E2		
/iga/	*	/keniro/	/komoke/	/korimo/
/aza/		/seniro/	/somoke/	/sorimo/
/kuda/		/teniro/	/tomoke/	/torimo/
/kaba/		/heniro/	/homoke/	/horimo/

Table 8: The list of the stimuli used in Set 2.

E1		E2		
/iga/	*	/kaniro/	/kamoke/	/karimo/
/aza/		/saniro/	/samoke/	/sarimo/
/kuda/		/taniro/	/tamoke/	/tarimo/
/kaba/		/haniro/	/hamoke/	/harimo/

The participants were 43 native speakers of Japanese, who were undergraduate students of a Japanese university. None of them participated in the experiment presented in Section 2. The experiment was conducted online using SurveyMonkey. In the test, they were presented with two elements (E1 and E2) and two forms (rendaku and non-rendaku forms), and were then

asked which was more natural; i.e. it was a forced-choice wug test (Berko 1958). The stimuli were presented in the hiragana Japanese orthography, the standard orthography to write native words (rendaku generally applies only to native words). The order of the stimuli was randomized. See Kawahara & Sano (2016) for further details.

Figure 3 shows the results of the applicability of rendaku for each condition. The results showed that there was a significant difference between cases that violate moraic identity avoidance and those without any violation (0.27 vs. 0.44; $z = 5.32$, $p < .001$). The results also show that there was a significant difference between consonantal identity avoidance and the control group (0.39 vs. 0.45; $z = 2.23$, $p < .05$). They also found a significant difference between moraic identity avoidance and consonantal identity avoidance ($z = 4.55$; $p < .001$), which suggests that the effect of identity avoidance is stronger at the CV-moraic level (the first bar) than at the consonantal level (the third bar).

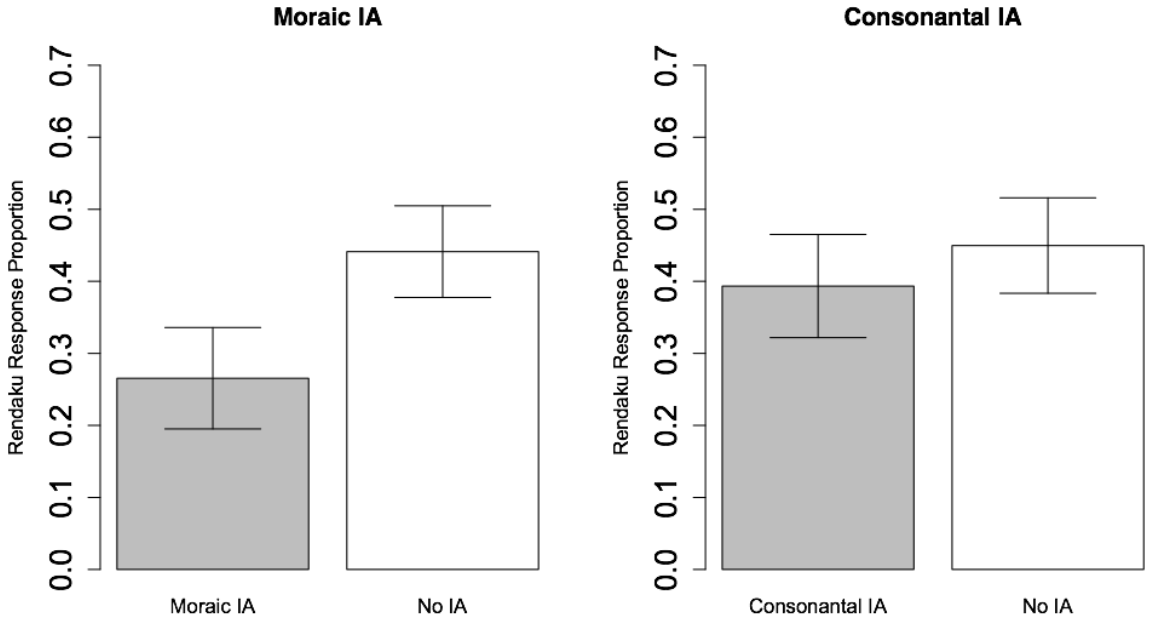

Figure 3: Proportion of rendaku application in each condition (adapted from Kawahara & Sano 2016)

To sum up, Kawahara & Sano (2016) showed that rendaku is less likely to occur when it results in identical consonants in adjacent moras. Furthermore, the applicability of rendaku was even more reduced when rendaku resulted in adjacent identical CV moras. These results exemplified a case in which the more similar the strings of segments that rendaku creates, the more likely it is avoided, again a case of gradient phonological effects on word-formation patterns. Like the case we analyzed in section 2, the effects of phonological constraints were stochastic; they did not deterministically block rendaku—they merely reduced the probability of rendaku application.

3.2. A MaxEnt analysis

For the current MaxEnt analysis of rendaku, we used four constraints. Following the most comprehensive OT analysis of Japanese rendaku presented by Ito & Mester (2003), we use REALIZE MORPHEME (RM) and IDENT (voice); the former constraint encourages rendaku, assuming that rendaku is a realization of a compound juncture morpheme. IDENT(voice) disfavors rendaku, because rendaku changes the underlying specification of a [voice] feature. We also used OCP(C) and OCP(CV) defined in Section 2.

Like the MaxEnt analysis presented in section 2, two candidates (rendaku and non-rendaku forms) were evaluated for each input form, with the violation profiles shown in (5). The results appear in Table 9 and (6). The MaxEnt Tool learned the experimental results with success, with the multiple OCP constraints we posited in section 2; the predicted probabilities are almost identical to the observed probabilities, as shown in (6).

(5) A MaxEnt analysis of rendaku

	RENDAKU	IDENT (voice)	OCP (C)	OCP (CV)		H-Score	$e^{-(\text{H-score})}$	Predicted Prob.
weights	4.89	5.1	0.24	0.60				
/…pa+ta…/								
…pata…	-1					-4.89	$7.52_{*10\text{^}-4}$	0.55
… pada…		-1				-5.1	$6.10_{*10\text{^}-4}$	0.45
/…ga+ko…/								
…gako…	-1					-4.89	$7.52_{*10\text{^}-4}$	0.61
…gago…		-1	-1			-5.34	$4.80_{*10\text{^}-4}$	0.39
/…ga+ka…/								
… gaka	-1					-4.89	$7.52_{*10\text{^}-4}$	0.74
… gaga…		-1	-1	-1		-5.69	$3.38_{*10\text{^}-4}$	0.26

Table 9: The posited constraints, and the obtained weights.

Constraints	Weight
RM	4.89
IDENT (voice)	5.1
OCP (C)	0.24
OCP (CV)	0.6

(6) Observed and predicted probabilities.

Forms	Observed	Predicted
…pa#ta…	0.55	0.55
…pa#da…	0.45	0.45
…ga#ko…	0.61	0.61
…ga#go…	0.39	0.39
…ga#ka…	0.74	0.74
…ga#ga…	0.26	0.26

4. Conclusions

The current paper explored a stochastic yet systematic aspect of Japanese word formation in group name formation and rendaku. In both types of word formation, sequences of two moras with the same nasal consonants are avoided, and sequences of two identical moras are avoided even more. However, it is not that case that a violation of one of these constraints entirely dictates the word formation pattern; the effects of phonological constraints are probabilistic, suggesting that phonological constraints can impose stochastic influences on word formation.

We also showed that a MaxEnt grammar is a general, useful tool to model such stochastic patterns. Overall, this research contributes to the growing body of literature showing that phonological knowledge can be stochastic and systematic.

In addition to these contributions to the issue of gradiency, we would like to highlight the fact that we are the first ones to systematically analyze the formation of group names in Japanese from the perspective of formal phonological theory. The results in section 2 show that this method is useful in revealing some aspects of phonological knowledge that Japanese speakers possess. In particular, we discovered that Japanese speakers may favor to place less sonorous consonant word-initially. We hope that this methodology will be used to explore the nature of other phonological and morphological patterns in other languages. In particular, since identity avoidance is observed across many languages, it is of interest to test the generality of how identity avoidance may affect the formation of new coordinate compounds, like those tested in section 2 of this paper (see also Shih 2013; Shih and Zuraw 2017).

The current study examined OCP effects in Japanese group name formation with experimentation; an interesting question that arises is whether the patterns we observed hold in existing words as well. Unfortunately, to the best of our knowledge, there is no large-scale corpus of Japanese group names consisting of two personal names, like *mana-kana*. However, there is an alternative way to address the OCP effects in Japanese: Many Japanese names consist of two disyllabic Sino-Japanese morphemes, such as *kazu-taka,* where *kazu* and *taka* are Sino-Japanese morphemes. If OCP(CV) is an active constraint in Japanese phonology, the prediction is that the order like *kazu-taka* is more frequent than the order like *taka-kazu*, as the latter violates OCP(CV). A future research can explore whether this prediction is borned out, using corpuses of Japanese names, in order to further test the activity of OCP(CV) in Japanese.

Acknowledgments

References

Alderete, John and Stefan Frisch. 2007. Dissimilation in grammar and the lexicon. *The Cambridge Handbook of Phonology*, ed. by Paul de Lacy, 379–398. Cambridge: Cambridge University Press.

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bat-El, Outi. 1996. Selecting the best of the worst: the grammar of Hebrew blends. *Phonology* 13: 283–328.

Benor, Sarah Bunin and Roger Levy. 2006. The chicken or the egg? A probalistic analysis of English binomials. *Language* 82: 233–277.

Berent, Iris, Donca Steriade, Tracy Lennertz and Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104: 591–630.

Berent, Iris and Joseph Shimron. 1997. Co-occurrence restrictions on identical consonants in the Hebrew lexicon: Are they due to similarity? *Journal of Linguistics* 39: 31–55.

Berkley, Deborah. 1994. The OCP and gradient data. *Studies in the Linguistic Sciences* 24: 59–72.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14: 150–177.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10): 341–345.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45–86.

Bresnan, Joan and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua* 118(2): 245–259.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1965. Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of speech*. Cambridge, MA: MIT Press.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology 1: Between the grammar and physics of speech*, ed. by John Kingston and Mary Beckman, 283–333. New York: Cambridge University Press.

Coetzee, Andries W. and S. Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 31(1): 47–89.

Coetzee, Andries W. and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory* 26: 289–337.

595 Cohn, Abigail. 2006. Is there gradient phonology? *Gradience in grammar: Generative per-*
596 *spectives*, ed. by Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesew-
597 sky, 25–44. Oxford: Oxford University Press.
598 Fnaselow, Gisbert, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky. 2006. eds. *Gradi-*
599 *ence in grammar: Generative perspectives*. Oxford: Oxford University Press.
600 Colavin, Rebecca, Roger Levy and Sharon Rose. 2014. Modeling OCP-place in Amharic with
601 the maximum entropy phonotactic learner. *Proceedings from the Annual Meeting of the*
602 *Chicago Linguistic Society* 46: 27–41.
603 Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis and Ingrid Norr-
604 mann. 2011. Explaining sonority projection effects. *Phonology* 28: 197–234.
605 Frisch, Stefan, Janet Pierrehumbert and Michael Broe. 2004. Similarity avoidance and the
606 OCP. *Natural Language & Linguistic Theory* 22: 179–228.
607 Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maxi-
608 mum entropy model. *Proceedings of the Stockholm Workshop on variation within Opti-*
609 *mality Theory*, ed. by Jennifer Spenader, Anders Erikson and Osten Dahl, 111–120.
610 Stockholm: Stockholm University.
611 Greenberg, Joseph H. and James J. Jenkins. 1964. Studies in the psychological correlates of
612 the sound system of American English. *Word* 20: 157–177.
613 Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the
614 sounds of their language. *Linguistic theory and psychological reality*, ed. by Morris Halle,
615 Joan Bresnan and George Miller, 294–303. Cambridge, MA: MIT Press.
616 Hayes, Bruce. 2017. Varieties of Noisy Harmonic Grammar. Proceedings of the 2016 Annual
617 Meeting of Phonology. Online Publication.
618 Hayes, Bruce and Colin Wilson 2008. A maximum entropy model of phonotactics and phono-
619 tactic learning. *Linguistic Inquiry* 39: 379–440.
620 Hayes, Bruce, and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: the case
621 of Hungarian vowel harmony. *Phonology* 23: 59–104.
622 Hayes, Bruce, Kie Zuraw, Peter Siptar and Zsuzsa Londe. 2009. Natural and unnatural con-
623 straints in Hungarian vowel harmony. *Language* 85: 822–863.
624 Hayes, Bruce, Colin Wilson and Anne Shisko. 2012. Maxent Grammars for the metrics of
625 Shakespeare and Milton. *Language* 88: 691–731.
626 Hayes, Bruce, Colin Wilson and Benjamin George. 2009. Manual for Maxent Grammar Tool.
627 Downloadable at http://www.linguistics.ucla.edu/people/hayes/Maxent-GrammarTool/
628 Irwin, Mark. 2016. A rendaku bibliography. *Sequential voicing in Japanese compounds: Pa-*
629 *pers from the NINJAL Rendaku Project*, ed. by Timothy Vance and Mark Irwin, 235–250.
630 Berlin: John Benjamins.
631 Ito, Junko. 1989. A prosodic theory of epenthesis. *Natural Language & Linguistic Theory* 7:
632 217–260.
633 Ito, Junko and Armin Mester. 2003. *Japanese morphophonemics: Markedness and word*
634 *structure*. Cambridge, MA: MIT Press.

635 Jäger, Gerhard and Anette Rosenbach. 2006. The winner takes it all - almost: Cumulativity in
636 grammatical variation. *Linguistics* 44: 937–971.
637 Jaeger, Jeri and John J. Ohala. 1984.On the structure of phonetic categories. *Tenth Annual*
638 *Meeting of the Berkeley Linguistics Society* (*BLS* 10), 15–26.
639 Katamba, Francis. 1993. *Morphology*. London: Macmillan Press, LTD.
640 Kawahara, Shigeto. 2013. Testing Japanese loanword devoicing: Addressing task effects. *Lin-*
641 *guistics* 51(6): 1271-1299.
642 Kawahara, Shigeto. 2015. Japanese /r/ is not feature-less: A rejoinder to Labrune (2014).
643 *Open Linguistics* 1: 432–443.
644 Kawahara, Shigeto, Hajime Ono and Kiyoshi Sudo. 2006. Consonant co-occurrence re-
645 strictions in Yamato Japanese. *Japanese/Korean Linguistics* 14, 27–38. Stanford: CSLI
646 Publications.
647 Kawahara, Shigeto and Shin-ichiro Sano. 2016. Rendaku and identity avoidance: Consonantal
648 identity and moraic identity. *Sequential voicing in Japanese compounds: Papers from the*
649 *NINJAL Rendaku Project*, ed. by Timothy J. Vance and Mark Irwin, 47–55. Amsterdam:
650 John Benjamins.
651 Keller, Frank. 2006. Linear Optimality Theory as a model of gradience in grammar. *Gradi-*
652 *ence in grammar: Generative perspectives*, ed. by Gisbert Fanselow, Caroline Féry, Ralph
653 Vogel, and Matthias Schlesewsky, 270–287. Oxford: Oxford University Press.
654 Kenstowiz, Michael. 1994. *Phonology in generative grammar*. Oxford: Blackwell.
655 Kilbourn-Ceron, Oriana and Morgan Sonderegger. to appear. Boundary phenomena and vari-
656 ability in Japanese high vowel devoicing. *Natural Language and Linguistic Theory*.
657 Kumagai, Gakuji. 2017. Cumulative faithfulness effect in Māori loanword adaptation: The
658 case of repair for consonant clusters. *Phonological Studies* 20: 77–84.
659 Kubozono, Haruo. 1989. The mora and syllable structure in Japanese: evidence from speech
660 errors. *Language and Speech* 32: 249–278.
661 Kubozono, Haruo. 2015. Diphthongs and vowel coalescence. *The handbook of Japanese lan-*
662 *guage and linguistics: phonetics and phonology*, ed. by Haruo Kubozono, 215–252. Ber-
663 lin: Mouton de Gruyter.
664 Labrune, Lawrence. 2006. Phonemic preferences in Japanese non-headed binary compounds:
665 what waa-puro, mecha-kucha and are-kore have in common. *Gengo Kenkyû (Journal of*
666 *the Linguistic Society of Japan)* 129: 3–41.
667 Labrune, Laurence. 2012. Questioning the universality of the syllable: Evidence from Japa-
668 nese. *Phonology* 29: 113–152.
669 Labrune, Lawrence. 2014. The phonology of Japanese /r/: A panchronic account. *Journal of*
670 *East Asian Linguistics* 23: 1–25.
671 Lasnik, Howard. 2004. Pronouns and non-coreference. *University of Maryland Working Pa-*
672 *pers in Linguistics* 13: 214–227.
673 Lasnik, Howard and Mamoru Saito. 1984. On the nature of proper government. *Linguistic In-*
674 *quiry* 15: 235–289.

675 Lass, Roger. 1976. *English phonology and phonological theory*. Cambridge: Cambridge Uni-
676 versity Press.
677 Leben, William R. 1973. Suprasegmental phonology. Doctoral dissertation. Massachusetts In-
678 stitute Technology.
679 Legendre, Géraldine, Yoshiro Miyata and Paul Smolensky. 1990. Harmonic Grammar—a for-
680 mal multi- level connectionist theory of linguistic wellformedness: An application. *Pro-*
681 *ceedings of the 20th annual conference of the Cognitive Science Society*, 884–891. Cam-
682 bridge: Lawrence Erlbaum.
683 Legendre, Géraldine, Antonella Sorace and Paul Smolensky. 2006. The optimality theory-har-
684 monic grammar connection. *The harmonic mind: From neural computation to optimality*
685 *theoretic grammar, vol. 2: Linguistic and philosophical implications*, ed. by Paul Smo-
686 lensky and Géraldine Legendre, 339–402. Cambridge, MA: MIT Press.
687 Lohmann, Arne. 2014. *English coordinate constructions: A processing perspective on constit-*
688 *uent order*. Cambridge, Cambridge University Press.
689 Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact
690 within the grammar. *Language* 87: 751–770.
691 McCarthy, John J. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17:
692 207–263.
693 McCawley, James D. 1968. *The phonological component of a grammar of Japanese*. The
694 Hague: Mouton.
695 McPherson, Laura and Bruce Hayes. 2016. Relating application frequency to morphological
696 structure: The case of Tommo So vowel harmony. *Phonology* 33: 125–167.
697 Mollin, Sandra. 2012. Revisiting binomial order in English: ordering constraints and reversi-
698 bility. *English Language and Linguistics* 16: 81–103.
699 Moore-Cantwell, Claire and Joe Pater. 2016. Gradient exceptionality in Maximum Entropy
700 Grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15: 53–66.
701 Neeleman, Ad. 2013. Comments on Pullum. *Mind & Language* 28(4): 522–531.
702 Otake, Takashi, Giyoo Hatano, Anne Cutler and Jacques Mehler. 1993. Mora or syllable?
703 Speech Segmentation in Japanese. *Journal of Memory and Language* 32: 258–278.
704 Padgett, Jaye. 1992. OCP subsidiary features. *Proceedings of the North East Linguistic Soci-*
705 *ety* 22, 335–346. University of Massachusetts, Amherst. GLSA.
706 Parker, Steve. 2002. Quantifying the sonority hierarchy. Doctoral dissertation. University of
707 Massachusetts, Amherst.
708 Parker, Steve. 2011. Sonority. *The Blackwell companion to phonology*, ed. by Marc van
709 Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice, 1160–1184. West Sussex,
710 UK: Wiley-Blackwell.
711 Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33: 999–
712 1035.
713 Pater, Joe. 2016. Universal grammar with weighted constraints. *Harmonic grammar and har-*
714 *monic serialism*, ed. by John McCarthy and Joe Pater, 1–46. London: Equinox Press.
715 Pierrehumbert, Janet B. 1997. Stochastic phonology. *Glot International* 5(6): 195–207.

716 Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt and Michael Becker. 2010. Har-
717 monic grammar with linear programming: From linear systems to linguistic typology.
718 *Phonology* 27: 77–117.
719 Prince, Alan. 1983. Relating to the grid. *Linguistic Inquiry* 14: 19–100.
720 Prince, Alan and Paul Smolensky. 1993/2004. *Optimality theory: Constraint interaction in*
721 *generative grammar*. Malden, MA & Oxford, UK: Blackwell.
722 Pullum, Geoferey K. 2013a. The central question in comparative syntactic metatheory. *Mind*
723 *& Language* 28(4): 492–521.
724 Pullum, Geoferey K. 2013b. Consigning phenomena to performance: A response to
725 Neeleman. *Mind & Language* 28(4): 532–537.
726 Sagey, Elizabeth. 1986. The representation of feature and relations in nonlinear phonology.
727 Doctoral dissertation, Massachusetts Institute of Technology.
728 Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and*
729 *linguistic methodology*. Chicago, IL. University of Chicago Press.
730 Schütze, Carson T. 2016. *The empirical base of linguistics: Grammaticality judgments and*
731 *linguistic methodology*. Berlin: Language Science Press.
732 Shih, Stephanie S. 2014. *Towards Optimal Rhyme*. Doctoral dissertation, Stanford University.
733 Shih, Stephanie S. 2016. Super additive similarity in Dioula tone harmony. *Proceedings of the*
734 *33rd West Coast Conference on Formal Linguistics*, ed. by Kyeong-min Kim et al., 361–
735 370. Somerville, MA: Cascadilla Proceedings Project.
736 Shih, Stephanie S. and Sharon Inkelas. 2016. Morphologically-conditioned tonotactics in
737 multilevel Maxent Entropy grammar. *Proceeding of Phonology 2015*. Online Publication.
738 The Linguistic Society of America. Washington, DC.
739 Shih, Stephanie S. and Kie Zuraw. to appear. Phonological conditions on variable adjective-
740 noun word order in Tagalog. *Phonological Analysis*.
741 Smith, Jennifer. 2002. Phonological Augmentation in Prominent Position. Doctoral disserta-
742 tion, University of Massachusetts, Amherst.
743 Sorace, Antonella and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11):
744 1497–1524.
745 Sprouse, Jon. 2007. A program a program for experimental syntax: Finding the relationship
746 between acceptability and grammatical knowledge. Doctoral dissertation. University of
747 Maryland, College Park.
748 Sprouse, Jon. 2015. Three open questions in experimental syntax. *Linguistics Vanguard*.
749 Online Publication.
750 Tanaka, Yu. 2017. *The Sound Patterns of Japanese Surnames*. Doctoral dissertation, UCLA.
751 Vance, Timothy J. 1980. The psychological status of a constraint on Japanese consonant alter-
752 nation. *Linguistics* 18: 245–267.
753 Vance, Timothy J. 1987. *An introduction to Japanese phonology*. New York: SUNY Press.
754 Vance, Timothy J. 2015. Rendaku. *The handbook of Japanese language and linguistics: pho-*
755 *netics and phonology*, ed. by Haruo Kubozono, 397–441. Berlin: Mouton de Gruyter.

Vance, Timothy J. and Mark Irwin. 2016. ed. *Sequential voicing in Japanese compounds: Papers from the NINJAL Rendaku Project*. Amsterdam: John Benjamins.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30: 945–982.

Yip, Moira. 1998. Identity avoidance in phonology and morphology. *Morphology and its Relation to Phonology and Syntax*, ed. by Steven G. Lapointe, Diane K. Brentari and Patrick M. Farrell, 216–246. Stanford: CSLI Publications.

Zhang, Jie, Yuwen Lai, and Craig Sailor. 2011. Modeling Taiwanese speakers' knowledge of tone sandhi in reduplication. *Lingua* 121: 186–206.

Zuraw, Kie. 2000. Patterned exceptions in phonology. Doctoral dissertation. University of California, Los Angeles.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language & Linguistic Theory* 28(2): 417–472.

Zuraw, Kie and Bruce Hayes. to appear. Intersecting constraint families: An argument for Harmonic Grammar. *Language*.

Zuraw, Kie and Yu-An Lu. 2009. Diverse repairs for multiple labial consonants. *Natural Language & Linguistic Theory* 17: 197–224.