

Modes of phonological judgment*

Shigeto Kawahara
Rutgers University

Abstract

Generative linguistics has primarily used introspection-based data for its theory construction. However, we now witness the rise of experimental approaches to linguistic judgments, in which linguistic judgment patterns are investigated through experimentation. Using patterns of obstruent devoicing in Japanese loanwords as a test case, the current project attempts to contribute to this growing body of work by investigating how different experimental variables affect phonological judgments. The three variables tested in the current experiments are (i) scalar rating judgments vs. binary yes/no judgments, (ii) real words vs. nonce words, and (iii) orthography stimuli vs. audio stimuli. The results show that (i) scalar rating tests and binary yes/no tests show very similar patterns, (ii) nonce words show less variability in acceptability across different grammatical conditions than real words, and (iii) orthography stimuli and audio stimuli yield comparable results, but (iv) audio-based experiments exaggerate the effect of particular phonetic implementation patterns as compared to orthography-based tests. Building on these results, this paper provides some suggestions for future experimentation on phonological judgments.

1 Introduction

1.1 The general aim

Generative linguistics has primarily used intuition-based data for theory construction. Oftentimes authors ask themselves whether certain linguistic structures or processes are grammatical or not.

*I could not have even started a project of this scale without the help of many people. For gathering participants, I am grateful for Yuki Hirose (the University of Tokyo), Yukino Kobayashi (Sophia University), Mutsuto Kawahara (Chuo University), Toshio Matsuura (Hokusei University), Noriko Nakanishi (Aizu University), and Mariko Sugahara (Doshisha university) for arranging their students to take these tests. I am also grateful to Osamu Fujimura, Kazu Kurisu, Julien Musolino, Jeremy Perkins, Jason Shaw, Mariko Sugahara, Kristen Syrett, Kyoko Yamaguchi, [AND PUT YOUR NAMES HERE!!], and the audience at the colloquium talks at the University of Pennsylvania and Johns Hopkins University, especially Geraldine Legendre, Mike McCloskey, Brenda Rapp, Paul Smolensky, and Colin Wilson, for their insightful comments. Please feel free to let me know any remaining errors.

Sometimes these introspection-based data are checked against the intuition of a few colleagues or friends. However, some concerns have been raised against this introspection-based approach (e.g. Schütze 1996 and see below for more references), and we now witness the rise of experimental approaches to linguistic judgments, in which linguistic judgments are elicited from a large number of theoretically naive speakers, using a protocol that is familiar from psychological research. The current project attempts to contribute to this general research enterprise by investigating how various experimental variables affect phonological judgment patterns. In particular, the current experiments compare different experimental paradigms that potentially affect phonological judgments. The three variables tested in the current experiments are (i) scalar rating judgments vs. binary yes/no judgments, (ii) real words vs. nonce words, and (iii) orthography-based testing vs. audio-based testing.

1.2 Empirical and theoretical background

1.2.1 The phenomena

As a case study, this study uses patterns of obstruent devoicing in Japanese loanword phonology in order to investigate the nature of phonological judgments. This section lays out some empirical and theoretical background for the discussion that follows. Starting with the empirical background, in Japanese, voiced geminates are not allowed in native phonology (Itô and Mester, 1995, 1999, 2008). However, in recent loanwords, voiced geminates do appear (Itô and Mester, 1995, 1999, 2008), as word-final consonants preceded by a lax vowel are often borrowed as geminates with a following epenthetic vowel (Katayama, 1998; Kaneko and Iverson, 2009; Kubozono et al., 2009; Shirai, 2002).

Nishimura (2003) pointed out that voiced geminates optionally devoice when they co-occur with another voiced obstruent, as exemplified in (1). This devoicing of geminates is caused by a restriction against two voiced obstruents within the same stem, the OCP(voice) (Itô and Mester, 1986, 1998, 2003a,b; Suzuki, 1998) (see Leben 1973; McCarthy 1986; Odden 1986; Suzuki 1998, among others, for general OCP). Nishimura (2003) and Kawahara (2006) contrast the OCP-violating geminates in (1) with voiced consonants in two other contexts: non-OCP-violating voiced geminates and OCP-violating singletons, whose devoicing, according to their introspective judgments, is ungrammatical, as in (2)-(3).

- (1) Optional grammatical devoicing of OCP-violating geminates
 - a. **baddo** → **batto** ‘bad’
 - b. **baggu** → **bakku** ‘bag’
 - c. **doggu** → **dokku** ‘dog’

- (2) Ungrammatical devoicing of non-OCP-violating geminates
 - a. sun**bbu** → *sun**ppu** ‘snob’
 - b. reddo → *retto ‘red’
 - c. eggu → *ekku ‘egg’
- (3) Ungrammatical devoicing of OCP-violating singletons
 - a. g**ibu** → *g**ipu** ‘give’
 - b. bagu → *baku ‘bug’
 - c. dagu → *daku ‘Doug’

1.2.2 The theoretical concern

Since Nishimura (2003), many theoretical claims have been made based on the patterns in (1)-(3) (e.g. Itô and Mester 2008; Kawahara 2006; Pater 2009; Tateishi 2002; see Kawahara 2011b for a summary). However, Kawahara (2011b) raised one concern: several theoretical claims have been made based on the behavior of [+voice] in Japanese, but the data are primarily based on the intuitions of Nishimura (2003) and Kawahara (2006). Kawahara (2011b) summarizes five concerns about a purely-intuition based approach, listed in (4) (see Alderete and Kochetov 2009; Coetzee 2005; Dabrowska 2010; Gibson and Fedorenko 2010; Griner 2001; Labov 1975, 1996; Myers 2009; Ohala 1974, 1986; Schütze 1996; Sprouse and Almeida 2010; Vance 1980; Wasow and Arnold 2005, among others, for further discussion).

- (4) Concerns about a purely-intuition based approach
 - a. **PRODUCTIVITY:** Some patterns that were used for theory construction have been shown to be non-productive under experimental settings.
 - b. **GENERALIZABILITY:** It is not clear whether the data are about Nishimura and Kawahara or the population of Japanese speakers.
 - c. **REPLICABILITY:** The intuitions are what Nishimura and Kawahara felt inside their minds, which cannot be observed from outside; i.e. cannot be replicated.
 - d. **OVERSIMPLIFICATION:** The introspection-based data may be oversimplified.
 - e. **BIAS:** The theoretical orientations of Nishimura and Kawahara could have skewed the data.

To address these concerns, Kawahara (2011b) conducted a rating study using naive native speakers of Japanese. The results basically supported the introspection-based data provided by Nishimura (2003) and Kawahara (2006) in that Japanese speakers found devoicing of OCP-violating geminates most natural. However, the results also revealed further aspects of devoicing patterns

in Japanese loanwords. First, Japanese speakers found devoicing of non-OCP-violating geminates (as in (2)) more natural than devoicing of OCP-violating singletons (as in (3)), both of which were judged to be ungrammatical by Nishimura (2003) and Kawahara (2006). This result shows that there was no clear line that divides the continuum into two dichotomous categories, “grammatical devoicing” and “ungrammatical devoicing”. Second, the devoicing pattern within OCP-violating geminates itself was not monolithic; other phonological and lexical factors affect naturalness ratings of devoicing of OCP-violating geminates (Kawahara, 2011a).

Kawahara (2011b) thus concludes that intuition-based data provide a useful first step in theory construction—Nishimura (2003) and Kawahara (2006) were not wrong when they provided their introspection-based judgments, and their data provided bases for further theoretical developments. However, the introspection-based data missed two aspects of naive native speakers’ actual behaviors: (i) the actual judgment patterns are more gradient than the binary dichotomy assumed by Nishimura (2003) and Kawahara (2006); (ii) the phonological patterns may not be as simple as Nishimura (2003) and Kawahara (2006) once contended in that various phonological and lexical factors affect the naturalness judgments of devoicing.

Generally speaking, then, a systematic experimental investigation can complement the traditional introspection-based approach by providing further insights into our phonological knowledge. Kawahara (2011b) is not an isolated case: there is a growing body of literature on how experimental studies can be used in tandem with the traditional introspection-based approach (see Alderete and Kochetov 2009; Coetzee 2005, to appear; Dabrowska 2010; Gibson and Fedorenko 2010; Griner 2001; Labov 1975, 1996; Myers 2009; Ohala 1974, 1986; Schütze 1996; Sprouse and Almeida 2010; Vance 1980; Wasow and Arnold 2005 among others for discussion). This project is intended to contribute to this growing body of literature. The current experimental studies start with the assumption that experimentation is useful in phonological research,¹ and investigate how different ways of running phonological experiments affect phonological judgment patterns.

1.3 The current study

The main goal of this study is, therefore, to test how different modes of experimental paradigms affect actual phonological judgment patterns. This paper takes the devoicing phenomenon in Japanese as a case study, and reports a set of studies that systematically vary different experimental variables. The experimental variables that are tested in this paper are listed in (5).

- (5) Three experimental variables
 - a. Scalar rating judgments vs. binary yes/no judgments.

¹I am not denying the usefulness of an introspection-based approach: Nishimura (2003) and Kawahara (2006) did provide bases for further theoretical discussion.

- b. Real word stimuli vs. nonce word stimuli.
- c. Orthography stimuli vs. audio stimuli.

The first experimental variable tested is a difference between judgments based on a scale and those based on a binary yes/no format. For example, given a word [doggu] ‘dog’, we can ask native speakers how natural they would find it to pronounce it as [dokku] on a scale, or we can ask them if it is possible to pronounce the word [doggu] as [dokku] with a binary choice format. Testing this difference is in part motivated by the debate concerning the gradient nature of phonological judgments. It has been known that grammatical judgments show distinctions beyond a “grammatical” vs. “ungrammatical” dichotomy, especially in experimental settings (e.g. Albright 2009; Berent et al. 2007; Coetzee 2008, 2009, to appear; Goldrick to appear; Greenberg and Jenkins 1964; Hayes 2000, 2009; Kawahara and Kao to appear; Pertz and Bever 1975; Pierrehumbert 2001; Shademan 2007; Zuraw 2000; see also Adli 2010; Chomsky 1965; Myers 2009; Schütze 1996; Sorace and Keller 2005 for a similar observation in syntactic judgments). However, one may contend that we obtain gradient results in experimental settings because the scales used in experimental settings are often scalar. Therefore, it is important to test whether phonological judgment patterns show gradient results even in a binary yes/no task. See Bader and Mäussler (2010) who test a similar issue in syntactic judgments.

The second variable tested is a difference between real words and nonce words. The standard assumption in generative phonology is that real words and nonce words are treated alike by grammar. Halle’s (1978) classic example—that *brick* and *blick* are assigned the same status (“grammatical”)—illustrates this assumption. Also, a popular test on phonological productivity is a wug-test (Berko, 1958), in which the participants are asked to inflect nonce words. In some cases, wug-tests fail to replicate phonological patterns that apply to real words, in which case it is often concluded that alleged phonological patterns are not productive i.e. lexicalized (Griner, 2001; Ohala, 1974; Sanders, 2003). Vitevitch and Luce (1998, 1999) moreover showed that real words and nonce words are affected differently by phonotactic probabilities and lexical neighborhood densities in speech processing (see also Shademan (2007) for some related discussion). A question that this paper addresses is how a difference between real words and nonce words affects phonological judgments. A more practical question is whether we should use real words or nonce words in testing phonological judgments.

The final variable tested is a difference between orthography-based test and audio-based test. When testing phonological judgments, the null hypothesis may be that, since phonology is about sounds, not about letters, we should use audio-based tests when possible. However, using orthography stimuli has virtues as well. Orthography stimuli are easier to use, especially in web-based experimentation, which has been receiving a growing body of interests in linguistics and elsewhere (Collins et al., 2009; Hayes et al., 2009; Kawahara, 2011a,b; Kawahara and Kao, to appear;

Sprouse, 2011b; Zuraw, 2006; Reips, 2002). Moreover, orthography-based tests avoid a problem of listeners' potentially mishearing the stimuli, which could affect the results (though see also Berent 2008). The current project thus compared phonological judgment patterns between these two modes of judgment.

One final note is in order. All the experiments reported in this paper are judgment experiments on a phonological process, i.e., devoicing. The task is for speakers to judge the naturalness or possibility of a phonological pattern, or in other words, a pairing between an underlying form and a surface form. This task therefore differs from phonotactic wellformedness judgment tasks in which speakers judge the wellformedness of surface forms only (Bailey and Hahn 2001; Coetzee 2008, 2009; Hay et al. 2003; Greenberg and Jenkins 1964; Kawahara and Kao to appear; Shademan 2007 and many others). For production studies comparing different mappings between an underlying forms and surface forms, see Davidson (2006, 2010).

This paper reports five experiments to address the three questions in (5), as summarized in (6). The first three experiments use orthography stimuli, which will be contrasted with the final two experiments, which use audio stimuli. Experiment I and IV use a scale, while the other three experiments use a yes/no format. The difference between real words and nonce words is tested as a within-subject variable. Experiment III addresses an order effect on the difference between real words and nonce words.

- (6) The five experiments
 - a. Experiment I: Orthography-based rating experiment
 - b. Experiment II: Orthography-based yes/no experiment
 - c. Experiment III: Orthography-based yes/no experiment II
 - d. Experiment IV: Audio-based rating experiment
 - e. Experiment V: Audio-based yes/no experiment

2 Experiment I: Web-based rating experiment

The first experiment is a web-based (i.e. orthography-based) rating experiment.

2.1 Method

2.1.1 Stimuli

All the experiments reported in this paper used the same set of stimuli, which consisted of four grammatical conditions: OCP-violating geminates, non-OCP-violating geminates, OCP-violating

singletons, and non-OCP-violating singletons, as summarized in (7). In this design, two factors—OCP and GEM—were fully crossed. This paper uses CAPITAL LETTERS to represent variable names.

- (7) The four grammatical conditions
- a. OCP-violating geminates (e.g. [b**aggu**])
 - b. non-OCP-violating geminates (e.g. [e**ggu**])
 - c. OCP-violating singletons (e.g. [d**agu**])
 - d. non-OCP-violating singletons (e.g. [m**agu**]).

The experiment had 9 items per each condition. The stimuli were all disyllabic. Among 9 items, 6 items contained [d] followed by epenthetic [o], 3 items contained [g] followed by epenthetic [u]. No stimuli with [b] were used, because [bb] is very rare in Japanese loanwords (Katayama, 1998; Shirai, 2002). The real word stimuli are listed in Table 1. Short vowels were used before geminates and [g]. Long vowels and diphthongs were used before singleton [d], because disyllabic loanwords with an initial short vowel almost always have a geminate [dd] ([bado] is a truncated form of [badominton]).

Table 1: The list of the stimuli: real words.

OCP GEM		GEM		OCP SING		SING	
baddo	‘bad’	heddo	‘head’	bado	‘badminton’	muudo	‘mood’
beddo	‘bed’	reddo	‘red’	gaido	‘guide’	waido	‘wide’
daddo	‘dad’	uddo	‘wood’	zoido	common name	haido	‘hide’
deddo	‘dead’	kiddo	‘kid’	boodo	‘board’	roodo	‘road’
guddo	‘good’	maddo	‘mad’	gaado	‘guard’	riido	‘lead’
goddo	‘god’	roddo	‘rod’	baado	‘bird’	huudo	‘food’
baggu	‘bag’	eggu	‘egg’	dagu	‘Doug’	hagu	‘hug’
biggu	‘big’	reggu	‘leg’	bagu	‘bug’	magu	‘mug’
doggu	‘dog’	taggu	‘tag’	jogu	‘jog’	ragu	‘rag’

The nonce word stimuli are listed in Table 2. The nonce word stimuli had the same phonological structures as the real-word stimuli, except that all the nonce word stimuli had short initial vowels.

2.1.2 Task

In this experiment Japanese speakers rated the naturalness of devoicing in the four grammatical conditions. The instructions explained that the questionnaire was about the naturalness of devoicing in Japanese loanwords. For each question, the participants were presented with one stimulus

Table 2: The list of the stimuli: nonce words.

OCP GEM	GEM	OCP SING	SING
buddo	keddo	budo	hudo
boddo	koddo	dado	rado
doddo	ruddo	dodo	rudo
geddo	yuddo	dedo	rido
gaddo	taddo	gado	yudo
giddo	kuddo	gudo	wado
boggu	uggu	degu	hegu
gaggu	oggu	dogu	negu
goggu	naggu	gegu	mugu

and asked to judge the naturalness of the form that undergoes devoicing of word-internal consonants (e.g. given [baddo], how natural would you find it to pronounce it as [batto]?). The instructions and the stimuli were presented in Japanese orthography. The *katakana* orthography was used for the stimuli, both for real words and nonce words, because it is used for loanwords and nonce words in the standard Japanese orthography convention. Although the test was based on orthography, the participants were asked to read each stimulus in their head, and make judgments based on their auditory impression rather than on orthography.²

In this experiment, the speakers judged the naturalness of devoicing on a 5-point scale: A. “very natural”, B. “somewhat natural”, C. “neither natural nor unnatural”, D. “somewhat unnatural”, and E. “very unnatural”.³ The software that ran the experiment (see below) could not present the scale numerically, so the responses were converted to a numerical scale later.

The main session was blocked into two parts. The first block presented all the real word stimuli, followed by a break sign. The second block presented all the nonce word stimuli. The entire experiment was blocked this way because it was assumed that making judgments about real words is easier for the participants. See Experiment III which addresses a possible order effect in this design. The participants went through both real words and nonce words, and hence the difference between real words and nonce word is a within-subject variable. (The other two experimental variables tested in this paper—scalar rating vs. yes/no and orthography stimuli vs. audio stimuli—are between-subject variables.)

²Although the instructions encouraged the participants to use their auditory impressions, the results of the orthography experiments differed (slightly) from the results of the purely auditory experiments. See Experiments IV and V for experiments using audio stimuli.

³The magnitude estimation task (Bard et al., 1996) could have been an alternative to the current rating study with a Lickert scale. See Sprouse (2009, 2011a) for a critical assessment of the use of magnitude estimation tasks for linguistic experiments.

2.1.3 Procedure

Sakai (<https://sakai.rutgers.edu/portal>) was used to run the online experiment (see Reips 2002 and Sprouse 2011b for general discussion on online experimentation in psychology and linguistics). The first page of the experimental website presented a consent form, followed by the instructions of the experiment. After the instructions, each page presented one trial. Sakai randomized the order of the stimuli. At the end of the experiment, the participants were asked if they were familiar with the devoicing phenomenon. To avoid bias effects due to their theoretical orientation, data from those who answered positively to this question were excluded.

2.1.4 Participants

Thirty-two native speakers of Japanese participated in this experiment. None of them reported that they are familiar with the devoicing phenomenon.

2.2 Statistics

The responses were first converted to numerical values as follows: “very natural”=5; “somewhat natural”=4; “neither natural nor unnatural”=3; “somewhat unnatural”=2; “very unnatural”=1. For statistical analyses, first, a general linear mixed model was run (Baayen et al., 2008; Baayen, 2008; Bates, 2005; Jaeger, 2008) using R (R Development Core Team, 1993–2011) with the `lme4` package (Bates et al., 2011). The p-values were calculated by the Markov chain Monte Carlo method using the `pval.func()` function of the `languageR` package (Baayen, 2009).

2.3 Results

Figure 1 illustrates average rating scores in the web-based rating experiment. In real words, the naturalness rating showed the following order: OCP-violating geminates (4.23) > non-OCP-violating geminates (3.29) > OCP-violating singletons (2.69) > non-OCP-violating singletons (2.21), replicating the previous studies (Kawahara, 2011a,b). Statistically, for real words, all factors were significant: OCP ($t = 5.29, p < .001$), GEM ($t = 11.81, p < .001$), and the interaction ($t = 2.68, p < .01$). These results show that OCP and GEM each affect naturalness ratings, and that the effect of OCP is bigger on the geminate pair (4.23-3.29=0.94) than on the singleton pair (2.69-2.21=0.48).

For nonce words, the order of the naturalness ratings is the same as the real word condition: OCP-violating geminates (3.64) > non-OCP-violating geminates (3.41) > OCP-violating singletons (3.06) > non-OCP-violating singletons (2.81). The statistical analysis shows that OCP ($t = 2.56, p < .05$) and GEM ($t = 6.44, p < .001$) are significant, but their interaction is not

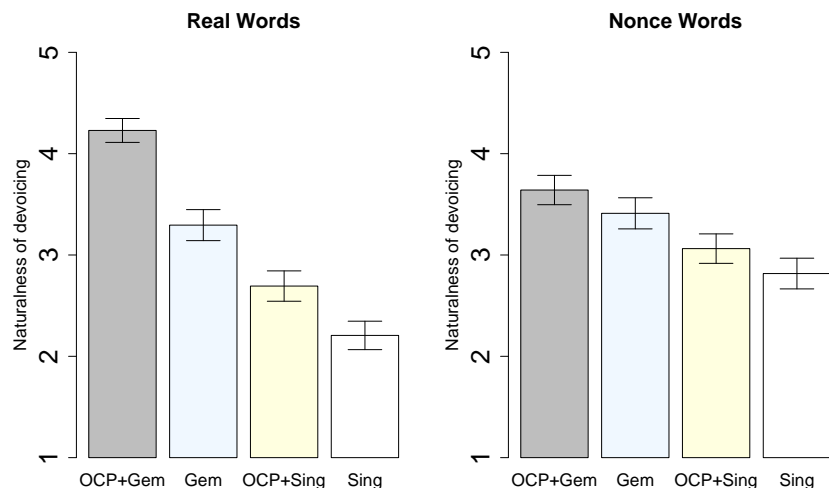


Figure 1: The average naturalness ratings in the web-based rating experiment (Experiment I). The error bars represent 95% confidence intervals.

($t = 0.06, n.s.$). For nonce words, the effect of OCP on naturalness ratings is comparable between the singleton condition (3.64-3.41=0.25) and the geminate condition (3.06-2.81=0.25).

2.4 Discussion

2.4.1 Gradiency

The results generally replicate the previous rating studies of the same phenomena in finding gradient grammatical distinctions (Kawahara, 2011a,b). There does not seem to be an objective line between “grammatical devoicing” and “ungrammatical devoicing”. In other words, in Figure 1, there does not seem to be an objective ground on which we could say that OCP-violating geminates are different from the other three conditions.

One question that arises is whether this four-way distinction is due to a non-homogeneous speech community. That is, one could argue that response from each speaker is always binary which follows a “grammatical” vs. “ungrammatical” dichotomy, but averaging over the responses from different speakers resulted in gradient patterns. This hypothesis predicts bimodal distributions of responses at two extremes, because people should consistently rate each devoicing pattern either as completely grammatical (=5 in rating) or completely ungrammatical (=1 in rating). In this view, the differences between the four grammatical conditions arise from the difference in the number of speakers who assign grammatical status to each condition. To examine this prediction, Figures 2 and 3 provide histograms that show the distributions of average scores for each speaker in each grammatical condition. We observe that, contra the hypothesis, there are many speakers who show

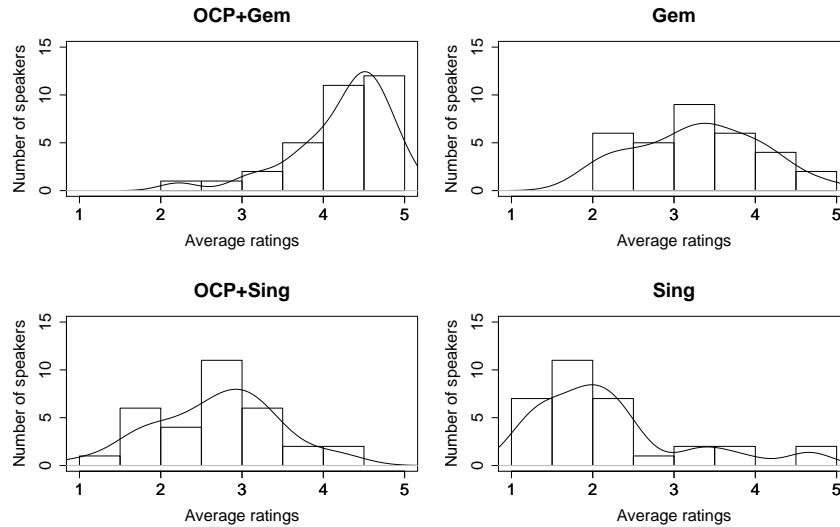


Figure 2: A histogram with a density plot: Web-based rating experiment, real words by speaker.

intermediate average scores in each grammatical condition.

An alternative to the hypothesis we examine in Figures 2 and 3 is to say that items within each grammatical condition showed a binary grammatical vs. ungrammatical pattern, but averaging over non-homogeneous set of items resulted in a gradient pattern. To check this possibility, Figures 4 and 5 illustrate the distributions of average naturalness ratings for each item. The hypothesis predicts that average scores for each item distribute bimodally at the two extreme ends, around grammatical (=5 in rating) and ungrammatical (=1 in rating). This prediction, however, is not supported by the actual data in Figures 4 and 5.

In summary, gradiency does not come from averaging over a non-homogeneous speech community or a non-homogeneous set of test items. It seems safe to conclude that the acceptability patterns show a gradient distinction, which goes beyond the “grammatical” vs. “ungrammatical” dichotomy (Albright, 2009; Berent et al., 2007; Coetzee, 2008, 2009, to appear; Goldrick, to appear; Greenberg and Jenkins, 1964; Hayes, 2000, 2009; Kawahara and Kao, to appear; Pertz and Bever, 1975; Pierrehumbert, 2001; Shademan, 2007; Zuraw, 2000).

2.4.2 The difference between real words and nonce words

Second, concerning the difference between real words and nonce words, we observe less variability in acceptability across the four grammatical conditions in nonce words than in real words. As observed in Figure 1, the most natural devoicing (OCP-violating geminates) is judged to be less natural in nonce words than in real words, and the least natural devoicing (non-OCP-violating geminates) is judged to be more natural in nonce words than in real words. In other words, the

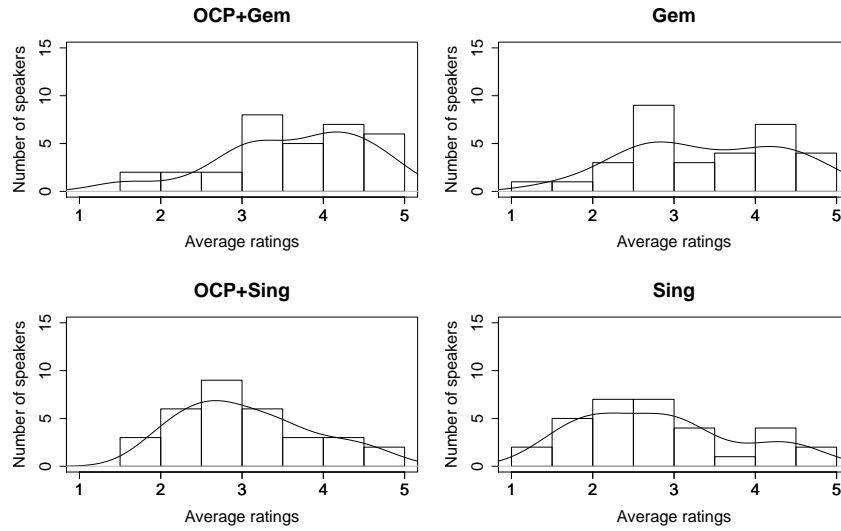


Figure 3: A histogram with a density plot: Web-based rating experiment, nonce words by speaker.

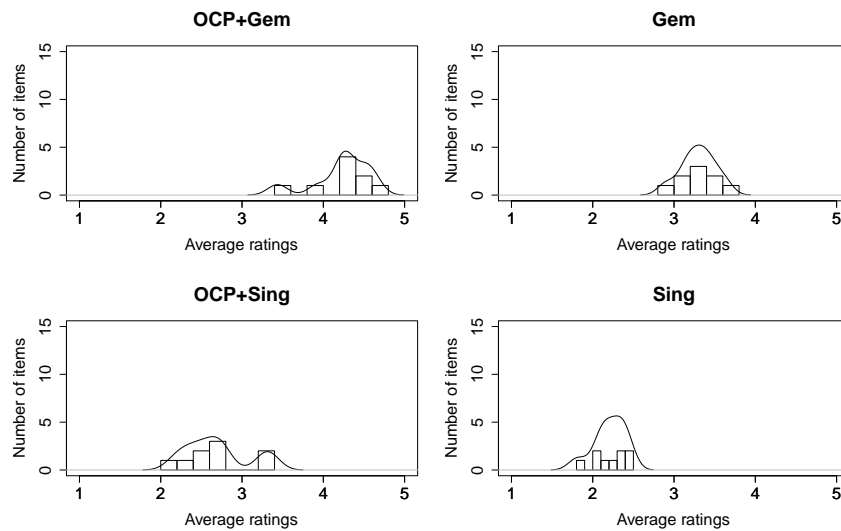


Figure 4: A histogram with a density plot: Web-based rating experiment, real words by item.

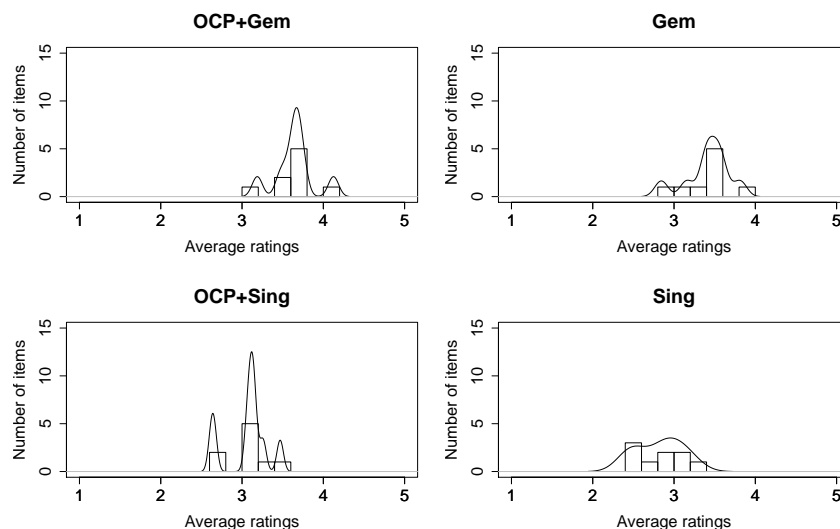


Figure 5: A histogram with a density plot: Web-based rating experiment, nonce words by item.

grammatical space—the range within which acceptability ratings can vary—is generally reduced in nonce words. This reduction of the grammatical space in nonce words may be responsible for the lack of a significant interaction between OCP and GEM in nonce words; there may not be a space left for OCP-violating geminates to have an acceptability rating that is high enough to yield a significant interaction between OCP and GEM.

To statistically assess this reduction of variability in rating in the nonce word condition, for each speaker, the standard deviations across all tokens were calculated separately for real words and nonce words. These standard deviations were compared between the two conditions using a within-subject Wilcoxon test. This analysis shows that the average standard deviations are 1.30 for the real words and 1.03 for the nonce words, and that the difference is significant ($p < .001$). Therefore, we can conclude that acceptability ratings vary less for nonce words than for real words.

Where does the difference between real words and nonce words come from? Presumably the participants have encountered real instances of devoicing in real words, which would make them “more confident” about what would happen to each target word in experimental settings. On the other hand, the participants have not seen nonce words before, and therefore they may feel less committed about making extreme grammatical judgments in general.

3 Experiment II: Web-based yes/no experiment

Experiment II is a web-based experiment, which used a yes/no, rather than scalar rating, format. The aim of this experiment is to compare judgment patterns made using a scale and those made

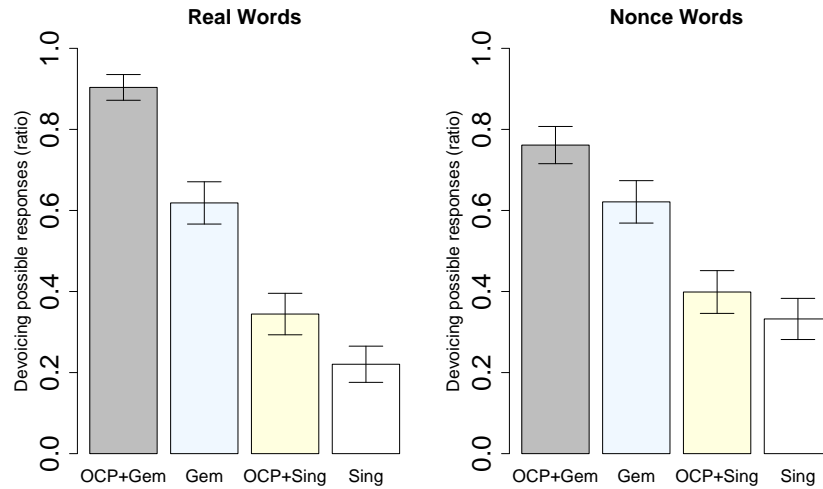


Figure 6: Average DEVOICING POSSIBLE response ratios in a web-based yes/no test (Experiment II). The error bars represent 95% confidence intervals.

using a binary yes/no format.

3.1 Method

Experiment II is similar to Experiment I, but it instead asked native speakers whether devoicing in each of the four grammatical conditions is possible or not in a binary yes/no format. Thirty-seven native speakers of Japanese participated in this experiment. No participants reported that they were familiar with the devoicing phenomenon. Since the responses were binary, a logistic linear mixed model (Jaeger, 2008; Quené and van den Berg, 2008) was used to analyze the results, again using R (R Development Core Team, 1993–2011).

3.2 Results

Figure 6 illustrates the average ratios of DEVOICING POSSIBLE responses—the average numbers of items participants chose DEVOICING POSSIBLE divided by the total number of trials—of each condition, both for real words and nonce words. The ratio followed the same hierarchy as the rating experiment for both real words and nonce words: OCP-violating geminates (0.90) > non-OCP-violating geminates (0.62) > OCP-violating singletons (0.34) > non-OCP-violating singletons (0.22) for real words, and OCP-violating geminates (0.76) > non-OCP-violating geminates (0.62) > OCP-violating singletons (0.40) > non-OCP-violating singletons (0.33) for nonce words.

A logistic linear mixed model on real words shows that OCP ($z = 4.17, p < .001$), GEM ($z = 11.09, p < .001$), and their interaction ($z = 3.67, p < .01$) are all significant. OCP and

GEM each increase the possibility of devoicing. The significant interaction shows that the effect of OCP is bigger on the geminate pair (0.28 increase in ratio (0.90-0.62)) than on the singleton pair (0.12 increase in ratio (0.34-0.22)). For nonce words, OCP ($z = 2.17, p < .05$) and GEM ($z = 8.56, p < .001$) are significant, but their interaction is not ($z = 1.65, n.s.$). There is some difference in the effect of OCP between the geminate pair (0.76-0.62=0.14) and the singleton pair (0.40-0.33=0.07), but the difference did not reach statistical significance.

3.3 Discussion

3.3.1 Rating experiments vs. yes/no experiments

First of all, the rating experiment (Experiment I) and the binary yes/no experiment (Experiment II) showed the same ordering between the four grammatical conditions. The results of the statistical tests on these two experiments are identical: for real words, both experiments showed significant main effects of OCP and GEM as well as a significant interaction effect between OCP and GEM; for nonce words, only the main effects of OCP and GEM were significant. These parallels show that a rating experiment and a yes/no experiment show very similar patterns. See Bader and Mäussler (2010) for a similar observation in syntactic experimentation.

3.3.2 Gradiency

Second, even when the speakers made binary yes/no judgments, we observe a four-way grammatical distinction. This result shows that the gradient pattern obtained in Experiment I was not due to the fact that the participants used a scale for their judgments; i.e. it was not a task effect. Acceptability patterns show a gradient distinction that goes beyond a “grammatical” vs. “ungrammatical” dichotomy, regardless of whether we use a scalar task or a binary yes/no task as an experimental format.⁴

3.3.3 Decrease in variability in nonce words

Third, we again observe reduction of the grammatical space in nonce words. As observed in Figure 6, OCP-violating geminates show fewer DEVOICING POSSIBLE responses in nonce words than in real words, and non-OCP violating singletons show more DEVOICING POSSIBLE responses in nonce words than in real words. To assess this decrease in variability in nonce words with respect to real words, standard deviations across the four grammatical conditions in the number of

⁴One may argue that this four-way grammatical distinction arose from averaging over a non-homogeneous speech community or a non-homogenous set of items. To address this possibility, analyses similar to those reported in Figures 2-5 were run for Experiment II, which showed that the four-way grammatical distinction does not arise from averaging over a non-homogeneous speech community or a non-homogeneous set of items.

DEVOICING POSSIBLE responses for each condition were calculated. The average standard deviations in the numbers of DEVOICING POSSIBLE responses were 3.04 for the real word condition and 2.36 for the nonce word condition, and the difference is significant according to a within-subject Wilcoxon test ($p < .001$). Responses to nonce words were indeed less variable than those to real words in Experiment II, just like in Experiment I.

3.4 Interim summary

Three observations have emerged from the results of the previous two experiments: (i) the acceptability hierarchy in devoicing shows a four-way distinction; (ii) a rating format and a binary yes/no format show a very similar pattern; (iii) variability across the four grammatical conditions is smaller for nonce words than for real words. The next experiment addresses one question regarding the third observation.

4 Experiment III: Web-based yes/no experiment 2

The previous two experiments have shown that we observe less variability across the four grammatical conditions in nonce words than in real words. However, in the previous two experiments, real words were presented in a block before the block for nonce words. The experiments were structured this way because making judgments about real words was expected to be easier than making judgments about nonce words. However, a question arises as to whether the difference between real words and nonce words can be due to an order effect. That is, the grammatical space may shrink as the participants proceed with an experiment. In other words, it is not the property of nonce words, but the fact that the nonce words were placed later in Experiments I and II, that is responsible for the reduction of variability in acceptability in nonce words.

4.1 Method

To address this question, a follow-up experiment was run, which was exactly the same as the previous yes/no experiment (Experiment II), except that nonce words are presented first before real words.⁵ Fifty-six speakers of Japanese participated in this experiment. Eight of them reported that they knew the devoicing phenomenon (some of them could have taken either of the previous two experiments). Hence the data from the remaining forty-eight speakers entered into the subsequent analysis.

⁵A yes/no format rather than a rating format was used in this experiment, because the interaction between OCP and GEM was closer to significance for the yes/no format. The reasoning was that if the reduction of variability was due to an order effect and is responsible for the lack of significant interaction between OCP and GEM, then it was expected that changing the order may make the interaction term significant in this experiment.

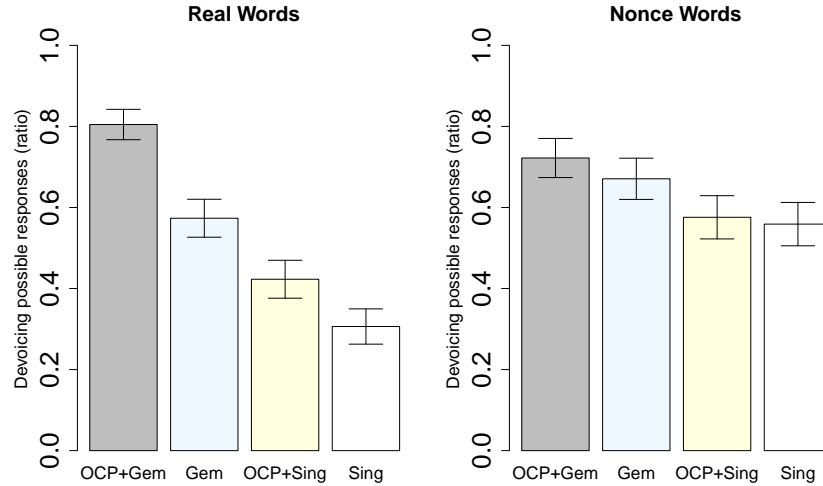


Figure 7: Average DEVOICING POSSIBLE response ratios in a web-based yes/no test in Experiment III. In this experiment, nonce words were presented before real words.

4.2 Results

Figure 7 shows the results of Experiment III. Experiment III yet again revealed the same ordering between the four-grammatical conditions: OCP-violating geminates (0.80) > non-OCP-violating geminates (0.57) > OCP-violating singletons (0.42) > non-OCP-violating singletons (0.31) for real words, and OCP-violating geminates (0.72) > non-OCP-violating geminates (0.67) > OCP-violating singletons (0.58) > non-OCP-violating singletons (0.56) for nonce words.

Statistically, for real words, OCP ($z = 4.81, p < .001$), GEM ($z = 9.71, p < .001$), and their interaction ($z = 2.66, p < .01$) were all significant; i.e. the same pattern as Experiments I and II. For nonce words, only GEM ($z = 4.55, p < .001$) was significant, and OCP ($z = 1.18, n.s.$) and the interaction ($z = 1.11, n.s.$) were non-significant. A simple analysis using only geminate data shows that OCP had a significant impact on the devoicability of geminates ($z = 2.70, p < .01$). The main effect of OCP in the general analysis was thus non-significant because its effect on the singleton pair was too small.

4.3 Discussion

4.3.1 Reduction of grammatical space in nonce words

We observe in Figure 7 that the grammatical space is again reduced in nonce words with respect to real words: DEVOICING POSSIBLE ratios differ less between the four grammatical conditions in nonce words than in real words. Average standard deviations in the numbers of DEVOICING POS-

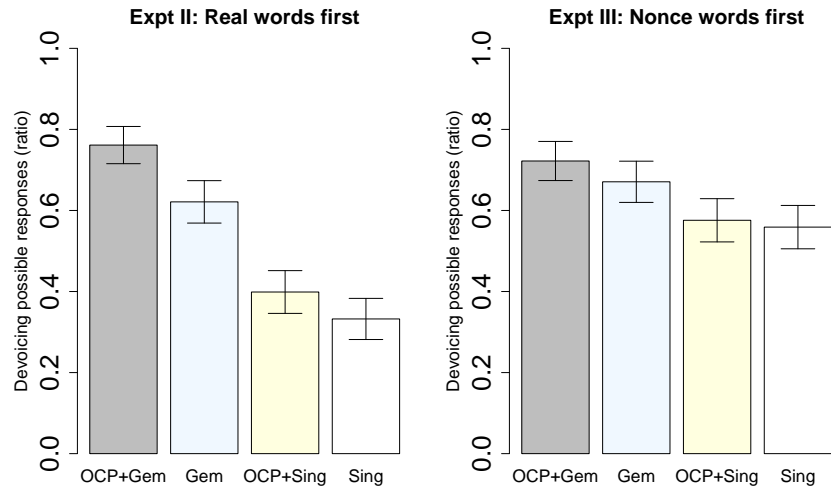


Figure 8: Comparison of the results of nonce words in Experiment II and Experiment III. The left figure=Experiment II; the right figure=Experiment III.

SIBLE responses were 2.46 for the real word condition and 1.48 for the nonce word condition, and they are different to a statistically significant degree ($p < .001$). The reduction of the grammatical space in nonce words is obtained even when nonce words were presented before real words. The reduction of variability in nonce words was not due to an order effect.

4.3.2 An order effect

However, ordering between real words and nonce words did have an effect on yes/no judgments in nonce words. To illustrate, Figure 8 compares DEVOICING POSSIBLE ratios in nonce words between Experiment II and Experiment III.

Figure 8 shows that there is actually an order effect after all, but not the kind that we expected. We observe even less variability in responses between the four grammatical conditions in Experiment III than in Experiment II, and the difference is significant according to a between-subject Wilcoxon test (average standard deviations: 2.36 vs. 1.48, $p < .01$). Judging real words first, as in Experiment II, enhances grammatical differences in nonce words later. Put differently, judging nonce words first would reduce differences between grammatical distinctions even further.⁶ Recall that when participants judged nonce words first in Experiment III, the grammatical space was reduced to the degree that the effect of OCP became non-significant.

To summarize the observations, in both Experiment II and Experiment III, the grammatical space is reduced in the nonce word condition, compared to the real word condition. In general

⁶A remaining question is what would happen if real words and nonce words are presented together within a single block. See Shademan (2007) for an example of such a wellformedness judgment study.

speakers make less extreme commitments about acceptability to nonce words than to real words. The grammatical space is even smaller when nonce words are presented before real words. In other words, judging real words first enforces grammatical differences between different grammatical conditions in judging nonce words later. Presumably, as discussed above, speakers make stronger commitments about grammaticality judgments for real words, because they have encountered (devoicing of) real words before in their lives. Having made judgments based on real words first may enforce the differences in acceptability across the different grammatical conditions, and this experience may help make judgments about nonce words (see Gulbertson and Gross 2009 for an effect of learning about making linguistic judgments.)

4.4 Interim conclusion

To summarize the results so far, OCP and GEM each affect naturalness ratings (in a rating study) and likelihood of devoicing (in binary yes/no studies). The effects of these two grammatical factors yield four-way distinctions in all the three experiments. In this sense, acceptability patterns go beyond a dichotomous, “grammatical” vs. “ungrammatical”, distinction. This gradient pattern is not due to averaging over data from different speakers or different items.

The rating study and the yes/no studies show very similar patterns. Not only are the orders between the four grammatical conditions identical, the patterns of statistical significance of each grammatical factor are almost identical between the two formats.

Finally, regarding the difference between real words and nonce words, the interaction between OCP and GEM is significant only in real words, in all three experiments. Acceptability differences across the four different conditions are reduced in nonce words. Judging nonce words before real words reduces grammatical differences in nonce words even more.

5 Experiment IV: Audio-based rating experiment

5.1 Introduction

The final two experiments tested another experimental variable: audio stimuli vs. orthography stimuli. When running experiments on phonological judgments, the null hypothesis may be that, audio-based experiments are better than orthography-based tests, since phonology is about sounds. However, logistically speaking, orthography-based tests are easier to prepare, especially for online experimentation. The last two experiments therefore investigated the comparability of audio-based experiments and orthography-based experiments.

In addition to this general aim, there was a secondary aim. Kawahara (2006) argued that geminates are more devoicable than singletons in Japanese loanword phonology, because a voicing

contrast is less perceptible in geminates than singletons, given how Japanese speakers phonetically implement voiced geminates. A judgment experiment using audio stimuli would help address this hypothesis.

5.2 Method

5.2.1 Stimuli

Experiments IV and V used the same set of stimuli as the web-based experiments. To obtain the auditory stimuli, a female native speaker of Japanese, who was naive to the purpose of this paper, pronounced all the stimuli seven times at a sound-attenuated booth. Her speech was recorded through an AT4040 Cardioid Capacitor microphone with a pop filter and amplified through an ART TubeMP microphone pre-amplifier (JVC RX 554V), digitized at a 44K sampling rate. From the seven repetitions, tokens that do not have phonetic deviance—such as heavy creakiness or unusual F0 contours—were chosen. To equalize the amplitudes of the stimuli, peak amplitude of all the stimuli was modified to 0.8 by Praat (Boersma and Weenink, 1999–2011). Then the files were converted to mp3 files and embedded in sakai tests. In her pronunciation, as expected, voiced geminates were semi-devoiced phonetically (Kawahara, 2006). As illustrated in the right panel of Figure 9, voicing during closure ceases at an early phase of the constriction interval. (However, see Kawahara 2006 for evidence that this phonetic semi-devoicing does not itself result in neutralization of a phonological voicing contrast.)

5.2.2 Participants and procedure

Experiment IV was a judgment experiment using a Lickert scale, as in Experiment I. Twenty-eight speakers participated in this experiment, but one speaker knew about the literature on devoicing in Japanese loanwords. The experiments were run in a quiet environment, using headphones. Other aspects of the experiment are identical to the previous three experiments, except that the experimenter sat with the participants as the experiment was run in Japan. As with Experiment I, within each trial, the participants were presented with an original form (e.g. [doggu] ‘dog’) and the form that undergoes devoicing (e.g. [dokku]). They were asked to rate the naturalness of the second form as a pronunciation of the original form. No orthographic representations of the stimuli were given—the participants only saw play buttons. Participants were allowed to listen to the stimuli as many times as they like.

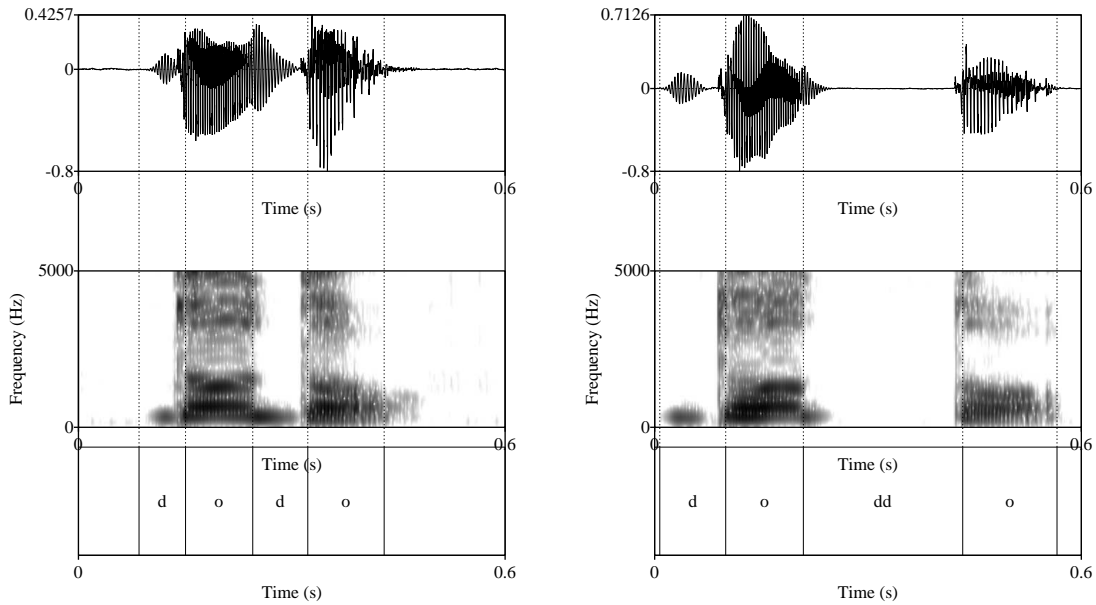


Figure 9: A comparison of a singleton [d] and a geminate [dd] in the current stimuli.

5.3 Results

Figure 10 illustrates the average naturalness ratings in Experiment IV. The results of the real words show the same hierarchy as the web-based experiments: OCP-violating geminates (3.89) > non-OCP-violating geminates (3.60) > OCP-violating singletons (1.92) > non-OCP-violating singletons (1.83). The statistical test shows that, for real words, GEM ($t = 17.75, p < .001$) was significant, but OCP ($t = 1.31, n.s.$) and the interaction ($t = 1.16, n.s.$) were not. Within geminates, OCP is significant ($t = 3.12, p < .01$). The main effect of OCP was therefore not significant in the general analysis because its effect on the singleton pair was too small.

Nonce words showed one reversal in that devoicing was rated higher for non-OCP-violating geminates (3.75) than for OCP-violating geminates (3.56). The rest of the orderings was identical to the previous experiments: OCP-violating singletons (2.57) > non-OCP violating singletons (2.46). Statically, GEM ($t = 12.00, p < .001$) is significant, but not OCP ($t = 0.78, n.s.$) or the interaction ($t = -1.85, n.s.$). The reversal in the geminate pair is significant ($t = -2.04, p < .05$).

5.4 Discussion

5.4.1 Orthography stimuli vs. audio stimuli

The ordering of the four grammatical conditions in real word condition is identical to the ordering we observed in the previous three experiments. In the nonce-word condition, we observe one

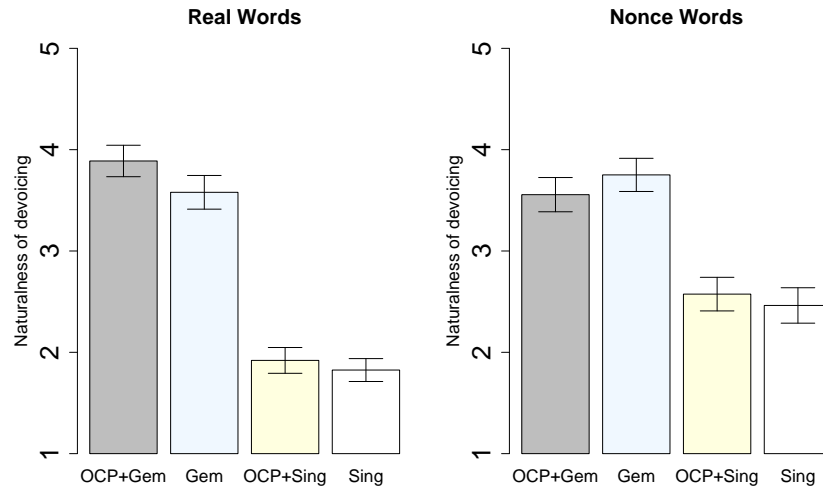


Figure 10: The average naturalness ratings in the audio rating experiment (Experiment IV).

reversal in the geminate pair. It therefore seems that orthography-based testing and audio-based testing show comparable results, especially in real words.

5.4.2 The magnified effects of GEM

Nevertheless, there is a difference between audio stimuli and orthography stimuli: the effect of GEM is magnified. In other words, the overall difference between the geminate conditions and the singleton conditions is magnified in this experiment, compared to Experiments I-III. To assess this difference statistically, for each speaker, the difference between the average ratings in the geminate conditions and the average ratings in the singleton conditions was calculated for Experiment I and Experiment IV. These values were compared using a between-subject Wilcoxon test, and it revealed a significant difference (0.94 in Experiment I vs. 1.49 in Experiment IV, $p < .05$). (See also Berent 2008 for a further discussion of differences between orthography stimuli and nonce stimuli.)

The reason for this magnified effect of GEM in Experiment IV perhaps lies in the phonetic semi-devoicing in Japanese voiced geminates. As we observe in Figure 9, the audio stimuli used in this experiment involved semi-devoiced voiced geminates. Therefore, the participants of this study heard renditions of voiced geminates that were already close to voiceless counterparts. On the other hand, voiced singleton stops were fully voiced, which sound very different from their voiceless counterparts. This difference in perceptibility of the [voice] contrasts was demonstrated in the perception experiment reported in Kawahara (2006). The current result thus supports Kawahara's (2006) hypothesis that the higher neutralizability of geminates may have its roots in the phonetic semi-devoicing of voiced geminates in Japanese.

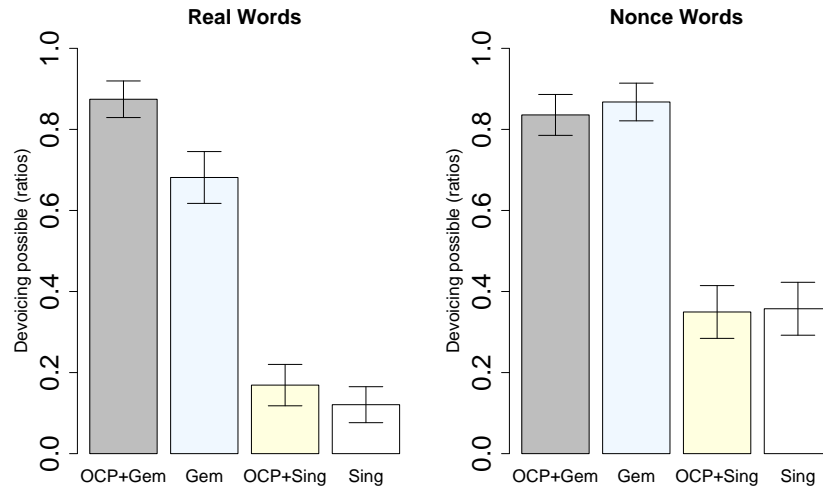


Figure 11: Average DEVOICING POSSIBLE response ratios in a audi-based yes/no test in Experiment V.

5.4.3 Reduction in variability in nonce words

Concerning the difference in variability between real words and nonce words, acceptability differences across all the four grammatical conditions are again reduced in nonce words. The difference in average standard deviations is 1.35 for the real words and 1.27 for the nonce words, and the difference between the two conditions is significant ($p < .01$).

6 Experiment V: Audio-based yes/no experiment

The final experiment is an audio-based experiment which used a yes/no format.

6.1 Method

Every aspect of the experiment was the same as that of Experiment IV, except that the experiment used a binary yes/no format; the participants were presented with an original form and a form that undergoes the devoicing in audio formats, and were asked if the second form was a possible pronunciation of the original form. Twenty-five speakers participated in this experiment.

6.2 Results

Figure 11 illustrates the results of Experiment V. The real words show the by-now familiar hierarchy: OCP-violating geminates (0.87) > non-OCP-violating geminates (0.68) > OCP-violating

singletons (0.17) > non-OCP-violating singletons (0.12). For real words, GEM ($z = 11.12, p < .001$) is significant, and OCP is not ($z = 1.42, n.s.$). However, the interaction is significant ($z = 2.18, p < .05$), reflecting the fact that OCP has a more tangible effect on the geminate pair than on the singleton pair. Within the geminate pair, OCP is significant ($z = 4.94, p < .001$).

The nonce words show non-significant reversals within the geminate and the singleton pairs: non-OCP-violating geminates (0.87) > OCP-violating geminates (0.84) > non-OCP-violating singletons (0.36) > OCP-violating singletons (0.35). The statistical test shows that only GEM ($z = 10.78, p < .001$) is significant, but not OCP ($z = -0.12, n.s.$) or the interaction ($z = -0.76, n.s.$). The reversal is not significant in the geminate pair ($z = -1.15, n.s.$) or in the singleton pair ($z = -0.13, n.s.$).

6.3 Discussion

6.3.1 Orthography-based testing and audio-based testing

The ordering between the four grammatical conditions in real words in the current experiment is identical to that observed in Experiments I-III. In nonce-words, the difference due to the OCP disappeared in both the singleton pair and the geminate pair. At least in the real word condition, we can conclude that orthography-based tests and audio-based tests yield comparable results.

6.3.2 The magnified effects of GEM

The effect of GEM is larger in the current audio-based experiment than in the orthography-based experiment (Experiment II) as well. The average difference between the geminate conditions and the singleton conditions in the number of DEVOICING POSSIBLE responses is 14.43 in Experiment II and 20.17 in Experiment V, and this difference is significant according to a between-subject Wilcoxon test ($p < .001$).

6.3.3 Reduction of variability in nonce words

Again, similar to all the previous experiments, acceptability differences across the four different conditions are reduced in nonce words. Average standard deviations in the numbers of DEVOICING POSSIBLE responses are 3.54 for the real words and 2.77 for the nonce words ($p < .001$).

7 General discussion

Before closing this paper, this section offers some general discussion.

7.1 Introspection-based data and experimental data

Concerning the status of OCP-violating geminates, which were treated as special by Nishimura (2003) and Kawahara (2006), all the experiments but the nonce word condition in Experiments IV and V showed that they are judged to be most likely to undergo devoicing. In this regard, the experiments show that the intuition by Nishimura (2003) and Kawahara (2006) is generally confirmed by the experimental findings, which indicates that an introspection-based approach provides a useful first-step in theory construction.

In the nonce word condition in the audio-based experiments, we observed reversals between OCP-violating geminates and non-OCP violating geminates, which was significant in Experiment IV and non-significant in Experiment V. Maybe these reversals occurred because in nonce words, acceptability differences are reduced in general, and in the audio-based experiments, devoicing of geminates was rated as highly acceptable. It may be that these two factors reduced the difference between OCP-violating geminates and non-OCP violating geminates (and somehow caused a reversal in the geminate pair in Experiment IV).

While the experimental results generally agree with the introspection-based data by Nishimura (2003) and Kawahara (2006), the experiments have also demonstrated that both the acceptability hierarchy (Experiment I and IV) and devoicability hierarchy (Experiment II, III and V) show a distinction that goes beyond a dichotomous “grammatical” vs “ungrammatical” distinction. This gradient pattern is observed even when the participants use a binary yes/no method. Even given such results, I acknowledge that one could still argue that grammar is dichotomous, and that it is performance that is gradient (e.g. Sprouse 2007). However, recall that generally OCP and GEM both contribute to the naturalness/possibility of devoicing, and these forces are most likely grammatical. This sort of view would then have to treat the effects of OCP and GEM as arising from performance factors, which is unlikely. Furthermore, recall that the gradient patterns were observed in the yes/no experiments as well, suggesting that the gradient results did not arise due to the fact that participants were forced to use a numerical scale.

Overall, the current studies show that experimentation provides further insights into phonological knowledge, which can be used in tandem with a traditional introspection-based approach (Alderete and Kochetov 2009; Dabrowska 2010; Gibson and Fedorenko 2010; Griner 2001; Labov 1975, 1996; Myers 2009; Ohala 1974, 1986; Schütze 1996; Sprouse and Almeida 2010; Vance 1980; Wasow and Arnold 2005, among others).

7.2 Summary of the effects of the experimental variables

The list in (8) summarizes the results of the current experiments, regarding how experimental variables affect phonological judgment patterns.

- (8) Summary of the effects of experimental variables
 - a. RATING VS. YES/NO: They show very similar patterns.
 - b. REAL VS. NONCE WORDS: Acceptability varies less across different grammatical conditions for nonce words than real words.
 - c. ORDER EFFECT: Judging nonce words before real words shrinks the grammatical space even more.
 - d. ORTHOGRAPHY STIMULI VS. AUDIO STIMULI: They yield comparable results especially in real words, but the effect of a particular phonetic implementation pattern is exaggerated in audio-based experiments.

The comparison between Experiments I and IV on the one hand and Experiments II, III and V on the other shows that experiments using a scalar rating and those using a binary yes/no format show very similar results.

Throughout all the experiments, nonce words show less variability across the four grammatical conditions in acceptability than real words. Moreover, the comparison between Experiment II and Experiment III shows that nonce words show even less variability when the participants were presented with nonce words before real words.

The comparison between Experiment I-III and Experiments IV-V show that audio stimuli and orthography stimuli yield comparable results, especially in real words. However, the effect of a particular phonetic implementation—semi-devoicing in Japanese voiced geminates—is exaggerated in audio-based experiments.

7.3 Lessons for future studies

The most important aim of this project has been methodological: how different tasks affect phonological judgment patterns. In (9) I summarize how we may utilize the current findings in future experimentation. I hasten to add however that these suggestions are purely based on the results of the current experiments, and we should be cautious about generalizing the current results to other cases.

- (9) Suggestions for future studies
 - a. The difference between scalar rating and a yes/no format should not matter.
 - b. Real words and nonce words show comparable, but slightly different, phonological judgment patterns.
 - c. The order of presentation between real words and nonce words matters.
 - d. Orthography stimuli and audio stimuli experiments show largely similar patterns.

- e. However, in audio-based experiments, a particular phonetic implementation may have stronger impact.

First of all, the current set of experiments did not yield substantial differences between experiments based on scalar rating and those based on a yes/no format. I acknowledge that it is dangerous to generalize this observation to other cases without actually testing this (lack of) difference with a wider range of data. However, until such differences are shown, this format difference does not seem crucial in running phonological judgment experiments. See Bader and Mäussler (2010) for similar results in syntactic judgments.

Second, we should bear in mind that real words and nonce words can show differences in acceptability patterns in phonological judgments. This paper does not address the issue of which one of these conditions—real words or nonce words—reflect phonological knowledge more directly. One could argue that phonological knowledge is acquired based on real words, and that real words therefore reflect phonological knowledge better. However, one could also argue that nonce words reflect phonological knowledge more directly, because nonce words are free(r) from the effects of lexical factors (see also Goldrick to appear; Vitevitch and Luce 1998, 1999; Shademan 2007 for related discussion). Until this debate is resolved, experiments eliciting linguistic judgments should include both real words and nonce words. At least, generalizing the results of real words to the results of nonce words, without testing the latter, runs the risk of overgeneralization.

Moreover, if we present both nonce words and real words in phonological judgment experiments, then the order of presentation matters. Recall that in Experiment III, in which nonce words were presented first, the differences between the four grammatical conditions were highly reduced in size. Therefore, presenting nonce words only may run the risk of missing grammatical differences, which could have been revealed by real word stimuli.

Finally, although orthography-based experiments and audio-based experiments show comparable results, in audio-based experiments, a particular phonetic implementation pattern may affect acceptability patterns more strongly. Therefore, when testing a phonological hypothesis in which a particular phonetic effect matters, then it would be safer to use auditory stimuli. However, we should also note that the effect of OCP was weakened in the two audio-based experiments, because the effect of GEM is also magnified. To the extent that the effects of the OCP are indeed robust in Japanese phonology, as evidenced in Experiments I-III, visual aids in phonological judgments may also be useful. I thus conclude that both modes of experiments should be used in future phonological judgment studies.

7.4 Remaining issues

Admittedly, this paper is just a beginning and it only scratched the surface of the intricacy of phonological judgment patterns. Many questions still remain—this paper probably raises more questions than it answers. For example, what if we mix real word stimuli and nonce word stimuli within the same block (Shademan, 2007)? What if we present audio stimuli together with orthography stimuli? We have observed from Experiments II and III that participants “learn the patterns” from the real word stimuli and can apply those patterns to nonce word stimuli later. Would we observe this learning effect within a block of real word stimuli if we have more real word items? How are the observed patterns reflected in actual production patterns? What kinds of models of grammar best account for the observed judgment patterns? How do the current observations generalize to other cases? These are all important questions, but beyond the scope of the current paper. In general more case studies with many other phonological patterns are necessary to address these questions. I hope that this paper stimulates more research on phonological judgments.

References

- Adli, Aria. 2010. Constraint cumulativity and gradience: Wh-scrambling in Persian. *Lingua* 120 (9): 2259–2294.
- Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26 (1): 9–41.
- Alderete, John, and Alexei Kochetov. 2009. Japanese mimetic palatalization revisited: Implications for conflicting directionality. *Phonology* 26 (3): 369–388.
- Baayen, Harald R. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baayen, Harald R. 2009. LanguageR. R package.
- Baayen, Harald R., Doug. J. Davidson, and Douglas. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412.
- Bader, Markus, and Jana Mäussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46: 273–330.
- Bailey, Todd, and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44: 568–591.
- Bard, Ellen. G., Dan. Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72: 32–68.
- Bates, Douglas. 2005. Fitting linear mixed models in R. *R News* 5: 27–30.
- Bates, Douglas, Martin Maechler, and Ben Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. R package.

- Berent, Iris. 2008. Are phonological representations of printed and spoken language isomorphic? evidence from the restrictions on unattested onsets. *Journal of Experimental Psychology: Human Perception and Performance* 34 (5): 1288–1304.
- 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.
- Berko, Jean. 1958. The child's learning of English morphology. *Word* 14: 150–177.
- Boersma, Paul, and David Weenink. 1999–2011. Praat: Doing phonetics by computer. Software.
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.
- Coetsee, Andries W. 2005. Using psycholinguistic data in phonology. ms. University of Michigan.
- Coetsee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84 (2): 218–257.
- Coetsee, Andries W. 2009. Grammar is both categorical and gradient. In *Phonological Argumentation: Essays on Evidence and Motivation*, ed. Steve Parker, 9–42. London: Equinox.
- Coetsee, Andries W. to appear. Gradient well-formedness in Harmonic Grammar: Phonological performance as a window on phonological competence. *Journal of the Phonetic Society of Japan*.
- Collins, Chris, Stephanie Guitard, and Wood Jim. 2009. Imposters: An online survey of grammaticality judgements. *NYU Working papers* 2.
- Dabrowska, Ewa. 2010. Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27 (1): 1–23.
- Davidson, Lisa. 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics* 34: 104–137.
- Davidson, Lisa. 2010. Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics* 38 (2): 272–288.
- Gibson, Edward, and Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14 (6): 233–234.
- Goldrick, Matthew. to appear. Utilizing psychological realism to advance phonological theory. In *The handbook of phonological theory II*, eds. John A. Goldsmith, Jason Riggle, and Alan Yu. Oxford: Blackwell-Wiley.
- Greenberg, Joseph, and James Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20: 157–177.
- Griner, Barry. 2001. Productivity of Japanese verb tense inflection: A case study. MA thesis, University of California Los Angeles.
- Gulbertson, Jennifer, and Steven Gross. 2009. Are linguists better subjects? *British Journal of Philosophical Science* 1: 1–16.
- Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the sounds of

- their language. In *Linguistic theory and psychological reality*, eds. Morris Halle, Joan Bresnan, and George A. Miller, 294–303. Cambridge: MIT Press.
- Hay, Jennifer, Janet Pierrehumbert, and Mary Beckman. 2003. Speech perception, well-formedness, and the statistics of the lexicon. In *Papers in laboratory phonology VI: Phonetic interpretation*, eds. John Local, Richard Ogden, and Rosalind Temple, 58–74. Cambridge: Cambridge University Press.
- Hayes, Bruce. 2000. Gradient well-formedness in Optimality Theory. In *Optimality Theory: Phonology, syntax, and acquisition*, eds. Joost Dekkers, Frank Van der Leeuw, and Jeroen Van de Weijer, 88–120. Oxford: Oxford University Press.
- Hayes, Bruce. 2009. Embedding grammar in a quantitative framework: Case studies from phonology and metrics. A handout for a minicourse at Brown University.
- Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85 (4): 822–863.
- Itô, Junko, and Armin Mester. 1986. The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry* 17: 49–73.
- Itô, Junko, and Armin Mester. 1995. Japanese phonology. In *The handbook of phonological theory*, ed. John Goldsmith, 817–838. Oxford: Blackwell.
- Itô, Junko, and Armin Mester. 1998. Markedness and word structure: Ocp effects in Japanese. ms. University of California, Santa Cruz.
- Itô, Junko, and Armin Mester. 1999. The phonological lexicon. In *The handbook of Japanese linguistics*, ed. Natsuko Tsujimura, 62–100. Oxford: Blackwell.
- Itô, Junko, and Armin Mester. 2003a. *Japanese morphophonemics*. Cambridge: MIT Press.
- Itô, Junko, and Armin Mester. 2003b. Lexical and postlexical phonology in Optimality Theory: Evidence from Japanese. *Linguistische Berichte* 11: 183–207.
- Itô, Junko, and Armin Mester. 2008. Lexical classes in phonology. In *The Oxford handbook of Japanese linguistics*, eds. Shigeru Miyagawa and Mamoru Saito, 84–106. Oxford: Oxford University Press.
- Jaeger, Florian T. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59: 434–446.
- Kaneko, Emiko, and Gregory Iverson. 2009. Phonetic and other factors in Japanese on-line adaptation of English final consonants. In *Studies in language sciences 8: Papers from the eighth annual conference of the Japanese Society for Language Science*, eds. Shunji Inagaki and Makiko Hirakawa. Tokyo: Kuroshio Publications.
- Katayama, Motoko. 1998. Optimality Theory and Japanese loanword phonology. Doctoral dissertation, University of California, Santa Cruz.
- Kawahara, Shigeto. 2006. A faithfulness ranking projected from a perceptibility scale: The case of

- voicing in Japanese. *Language* 82 (3): 536–574.
- Kawahara, Shigeto. 2011a. Aspects of Japanese loanword devoicing. *Journal of East Asian Linguistics* 20.
- Kawahara, Shigeto. 2011b. Japanese loanword devoicing revisited: A rating study. *Natural language and Linguistic Theory*.
- Kawahara, Shigeto, and Sophia Kao. to appear. The productivity of a root-initial accenting suffix, [-zu]: Judgment studies. *Natural Language and Linguistic Theory*.
- Kubozono, Haruo, Junko Itô, and Armin Mester. 2009. Consonant gemination in Japanese loanword phonology. In *Current issues in unity and diversity of languages. collection of papers selected from the 18th international congress of linguists*, ed. The Linguistic Society of Korea, 953–973. Republic of Korea: Dongam Publishing Co..
- Labov, William. 1975. What is a linguistic fact? In *Empirical foundations of linguistic theory: The scope of American linguistics*, ed. Robert Austerlitz, 77–113. Lisse: The Peter de Ridder Press.
- Labov, William. 1996. When intuitions fail. In *Papers from the the 32nd regional meeting of Chicago Linguistic Society: Papers from the parasession on theory and data in linguistics*, eds. Lisa McNair, Kora Singer, Lise Dolbrin, and Michelle Aucon, 77–106. Chicago: Chicago Linguistics Society.
- Leben, Will. 1973. Suprasegmental phonology. Doctoral dissertation, MIT.
- McCarthy, John J. 1986. Ocp effects: Gemination and antigemination. *Linguistic Inquiry* 17: 207–263.
- Myers, James. 2009. Syntactic judgment experiments. *Language and Linguistic Compass* 3 (1): 406–423.
- Nishimura, Kohei. 2003. Lyman’s Law in loanwords. MA thesis, Nagoya University.
- Odden, David. 1986. On the obligatory contour principle. *Language* 62: 353–383.
- Ohala, John J. 1974. Experimental historical phonology. In *Historical linguistics II: Theory and description in phonology. Proceedings of the first international linguistic conference on historical linguistics*, eds. J. M. Naderson and Charles Jones, 353–389. New York: Elsevier.
- Ohala, John J. 1986. Consumer’s guide to evidence in phonology. *Phonology* 3: 3–26.
- Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33: 999–1035.
- Pertz, D. L., and T. G. Bever. 1975. Sensitivity to phonological universals in children and adolescents. *Language* 51: 149–162.
- Pierrehumbert, Janet B. 2001. Stochastic phonology. *GLOT* 5: 1–13.
- Quené, Hugo, and Huub van den Berg. 2008. Examples of mixed effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59: 413–425.
- R Development Core Team. 1993–2011. *R: A language and environment for statistical computing*. Vienna, Austria. R Foundation for Statistical Computing. Software, available at <http://www.R->

project.org.

- Reips, Ulf-Dietrich. 2002. Standards for internet-based experimenting. *Experimental psychology* 49 (4): 243–256.
- Sanders, Nathan. 2003. Opacity and sound change in the Polish lexicon. Doctoral dissertation, University of California, Santa Cruz.
- Schütze, Carlson. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
- Shademan, Shabhame. 2007. Grammar and analogy in phonotactic well-formedness judgments. Doctoral dissertation, University of California, Los Angeles..
- Shirai, Setsuko. 2002. Gemination in loans from English to Japanese. MA thesis, University of Washington.
- Sorace, Antonella, and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115 (11): 1497–1524.
- Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1: 118–129.
- Sprouse, Jon. 2009. Magnitude estimation and the non-linearity of acceptability judgments. In *Proceedings of West Coast Conference on Formal Linguistics 27*, eds. Natasha Abner and Jason Bishop. Somerville: Cascadia Press.
- Sprouse, Jon. 2011a. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*.
- Sprouse, Jon. 2011b. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior and Research methods* 43.
- Sprouse, Jon, and Diogo Almeida. 2010. A quantitative defense of linguistic methodology. Ms. University of California, Irvine.
- Suzuki, Keiichiro. 1998. A typological investigation of dissimilation. Doctoral dissertation, University of Arizona.
- Tateishi, Koichi. 2002. Bunpoo no ichibutoshite no goisoo no zehi [Lexical strata as a part of grammar]. *Journal of the Phonetic Society of Japan* 6 (1): 34–43.
- Vance, Timothy J. 1980. The psychological status of a constraint on Japanese consonant alternation. *Linguistics* 18: 245–267.
- Vitevitch, Michael, and Paul Luce. 1998. When words compete: Levels of processing in spoken word recognition. *Psychonomic Science* 9: 325–329.
- Vitevitch, Michael, and Paul Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40: 374–408.
- Wasow, Thomas, and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115: 1481–1496.

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

Zuraw, Kie. 2006. Using the web as a phonological corpus: A case study from Tagalog. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics/Proceedings of the 2nd International Workshop on Web As Corpus*.