Comparing a forced-choice wug test and a naturalness rating test: An exploration using rendaku

Abstract

A growing body of linguistic studies is now deploying judgment experiments to probe both syntactic and phonological knowledge. Thus a research question arises as to what kind of judgment format is useful for probing our linguistic knowledge. Against this theoretical background, this study compares two types of phonological judgment experimentation: a scale-based naturalness judgment task and a forced-choice wug test. The current analysis uses the data from two previously published studies on rendaku, a famous voicing phenomenon found in Japanese compound formation, and Lyman's Law, which is known to inhibit rendaku. Although the two tasks at first sight show a close correlation with each other, a detailed examination of the data shows that the forced-choice wug-test reveals the influence of Lyman's Law on rendaku more clearly than the naturalness judgment experiment. To the extent that the effect of Lyman's Law is real, the current comparison shows that a forced-choice wug experiment is better than a naturalness judgment experiment. While the impact of the current results is limited and modest, this study provides a first step toward understanding how different tasks in phonological experimentation may compare to one another. It is hoped that the current study will plant a seed for a research program which addresses which kind of phonological judgment experimentation is best-suited to reveal our phonological knowledge.

1 Introduction

Generative linguistics has long been relying on the data based on native speakers' "intuitions" or "introspection". Theories have been built based on how native speakers of a particular language feel about a set of sentences (in syntax) or phonological structures and processes (in phonology), and it is not unusual that these native speakers are the authors of the papers themselves. This method has been criticized since the early years of the generative enterprise (Hill, 1961; Spencer, 1973). To briefly summarize the general concerns raised against this approach (see Schütze 1996

especially), first of all, such introspection data may be biased, because authors themselves produce the data. Second, they may also be oversimplified, sometimes under the rubric of "idealization", so that they would fit the theory that is being proposed. Third, introspection-based data may not be replicable with other speakers, because introspection is a matter of inner sensation which cannot be observed from outside. Finally, in the domain of phonology, concerns have been raised regarding whether some "phonological patterns," revealed through introspection, are truly productive (see Kawahara 2015).

In recent years, the problem of the heavy reliance on introspection in theory construction in generative linguistics has received renewed interests from various perspectives (see Coetzee et al. 2009; Cowart 1997; Dąbrowska 2010; Edelman & Christianson 2003; Ferreira 2005; Gibson & Fedorenko 2010; Goldrick 2011; Kawahara 2011c, 2015; Marantz 2005; Myers 2009; Ohala 1986; Phillips 2009; Riemer 2009; Schütze 1996, 2011; Sprouse & Almeida 2012a,b; Ueyama 2010; Wasow & Arnold 2005 among many others for various perspectives on this issue). The general consensus that is emerging from this debate is that the traditional introspection-based approach is not entirely unreliable, but at the same time experimentation complements this traditional approach in a number of useful ways. Therefore, we are witnessing the rise of interest in eliciting linguistic data by way of experimentation using a large number of naive native participants, both in syntax and phonology (see Myers 2009 for the former; see Coetzee et al. 2009 for the latter).

There are many conceivable types of experiments that one can perform to obtain acceptability judgments from native speakers in experimental settings. One question that arises in this theoretical context is how different tests may compare to one another. This question is important partly because different studies deploy different types of experiments, and also because we want to know which tests are more reliable in detecting grammatical differences under investigation.[1] Sprouse & Almeida (2012b) present an extensive comparison between different types of grammaticality judgment experiments in syntax, but there is nothing comparable in the domain of phonological judgment experimentation. This paper thus provides a case study in the domain of phonology, although the scale of the current study is admittedly much more limited and modest. This work can also be understood as a part of the larger research enterprise addressing task effects in phonetic and phonological experimentation (for other papers that address this general issue, see Berent 2008; Daland et al. 2011; Kawahara 2013; Perkins 2014; Yu & Lee 2014).

As part of this research enterprise in linguistic experimentation, this paper reports the results of comparing a forced-choice wug test and a naturalness judgment experiment. Wug-testing was made famous by the seminal work by Berko (1958), which asked English-speaking children to inflect nonce words, including the famous nonce word *wug*. This methodology has been deployed in

[1]This is not to say that any experimental format can be used for any kind of phonological pattern. For example, wug-tests are impossible to use to test patterns that do not involve any sort of word formation; e.g. phontactic judgment experiments.

many experimental works in phonology (see Kawahara 2011b for a recent review). In naturalness judgment tasks, which correspond closely to the standard practice in syntax, sentences are assigned a scale of acceptability; e.g. * (totally unacceptable) > *? (unacceptable) > ?? (very questionable) > ? (questionable) > unmarked (acceptable).[2] This sort of task has also been deployed in phonological experimentation as well (e.g. Daland et al. 2011; Gouskova & Roon 2013; Kawahara 2011a,c).

This paper compares these two tasks, using rendaku as a case study.[3] Rendaku is a morphophonological process whereby the initial consonants of the second members of compounds appear as voiced (e.g. /**t**ako/ 'octopus' → /oo-**d**ako/ 'big octopus').[4] In the rest of this paper, the second elements—potential targets of rendaku—are referred to as E2 (for Element 2), and the first elements of compounds are referred to as E1. The application of rendaku on E2 is not automatic in the sense that many factors affect the applicability of rendaku (Vance, 2015; Vance & Irwin, to appear), and many experiments have been run to test the psychological reality of such rendaku-affecting factors (see Kawahara to appear for a review). For example, rendaku is known to be blocked when the second member already contains a voiced obstruent (/**t**oka**g**e/ 'lizard' → /oo-**t**oka**g**e/ 'big lizard')—and this blockage is known as Lyman's Law (Lyman 1894 *et seq.*; Vance 2007, 2015). Two experiments in the past have explored rendaku and the effect of Lyman's Law experimentally using nonce words, one with the wug-test format (Kawahara & Sano, 2014) and one with the naturalness judgment format (Kawahara, 2012), with an overlapping set of stimuli.[5] These data allow us to compare the two types of tests, which the current study aims to do.

2 The two previous studies

Kawahara (2012) used a naturalness judgment task to explore the effect of Lyman's Law on rendaku. The study used a 5-point Likert scale: 5. "very natural", 4. "somewhat natural", 3. "neither natural nor unnatural", 2. "somewhat unnatural", and 1. "very unnatural". The experiment asked native speakers of Japanese to rate the naturalness of forms that underwent rendaku (e.g. "How natural does /X+**g**atoni/ from /**k**atoni/ sound for you?"). The stimuli consisted of two conditions: those items whose rendaku would not violate Lyman's Law (e.g. /katoni/) and those items whose rendaku would violate Lyman's Law (e.g. /kabomo/). The study found that Japanese speakers rate rendaku less naturally when it violates Lyman's Law than when it does not; e.g. /X+**g**a**b**omo/ from

[2]For example, Lasnik & Saito (1984) (pp. 206-209), cited and discussed by Pullum (2013) (pp. 504-505), distinguished 5-levels of grammaticality: "*", "?*", "??", "?" and no mark.

[3]There are other types of judgment experiments in linguistics, including a magnitude estimation task (Bard et al., 1996), a binary yes/no task (Kawahara, 2013), and a free-production/elicitation task (which is common in fieldwork studies).

[4]For the sake of simplicity, this paper uses phonemic transcription rather than IPA transcription.

[5]For other previous experimental studies on rendaku and Lyman's Law, see Vance (1980) and Ihara et al. (2009).

/kabomo/ is judged to be less natural than /X+**g**atoni/ from /katoni/.

Kawahara & Sano (2014) explored a similar issue, using a forced-choice wug-test. In that experiment, within each trial, the participants were first presented a nonce word as E2, and then given two inflected (i.e. compound) forms of that E2, one with rendaku and one without rendaku. Then, the participants were asked which one sounds better (e.g. "Which one of the following choices sounds better, /X+**g**otoni/ or /X+**k**otoni/, when /X/ and /kotoni/ are combined?"). This task is called a "wug-test," because it involves inflection of nonce words by the participants (i.e. how nonce word E2s are pronounced in compound formation)—this format is also forced-choice, because the set of possible choices is pre-determined.[6] This format is also known as "head-to-head" (Daland et al., 2011), or 2 alternative-forced choice (2AFC) (Macmillan & Creelman, 2005). This study too found that rendaku is less likely when it violates Lyman's Law than when it does not.

Table 1: The list of nonce word stimuli used as E2 in both Kawahara (2012) and Kawahara & Sano (2014). There were some stimulus items used only in one of the two experiments, but such items are not analyzed in the current study. Both experiments used *nise* 'fake' as E1.

No LL violation	LL violation	
hinumi	haboke	koriga
honara	hekazu	sebato
katoni	hemiga	segeha
kikake	hobasa	sekabo
kimane	hogore	sukaza
sekato	hokada	taguta
semaro	kabomo	tatuga
sutane	kamagi	tegura
tamuma	kidake	tenago
taruna	kitage	tomiba
tatuka	kobono	tozumi

The two studies used a set of overlapping nonce words stimuli for E2, as shown in Table 1. This set of stimuli allows us to compare the two different tasks deployed in these experiments. Both studies used *nise* 'fake' as E1 and hence this factor is controlled across the two experiments. For E2, there are 11 items that would not violate Lyman's Law after rendaku, and there are 22 items that would violate Lyman's Law after rendaku. There are twice as many items in the second condition, because the two previous studies tested the locality effect of Lyman's Law (i.e. whether the blocking consonant is in the second syllable or in the third syllable). Since the locality effect was not evident in either of the studies, they are collapsed together in the current analyses.

[6]This format is different from other wug-tests which involve free elicitation of inflected forms; i.e. it is not necessarily the case that all wug-tests need to be run in a forced-choice format. Neither is it the case that all forced-choice tests are wug-tests.

For each item, average naturalness rating scores and average rendaku response proportions are calculated across all the participants (forty-three participants for the naturalness experiment and thirty-eight participants for the forced-choice wug-test). These scores allow us to compare the relationship between the two tasks.

3 Result

Figure 1 plots, for each item, the average naturalness rating on the x-axis and the average rendaku response proportion from the wug-test on the y-axis. There is a positive correlation between the two dimensions ($\rho = 0.64, p < .001$) in such a way that an item whose rendaku is rated more naturally in Kawahara (2012) is more likely to undergo rendaku in the wug-test in Kawahara & Sano (2014). A regression analysis shows that as the naturalness score goes up by 1, the rendaku-undergoing probability goes up by 0.18.

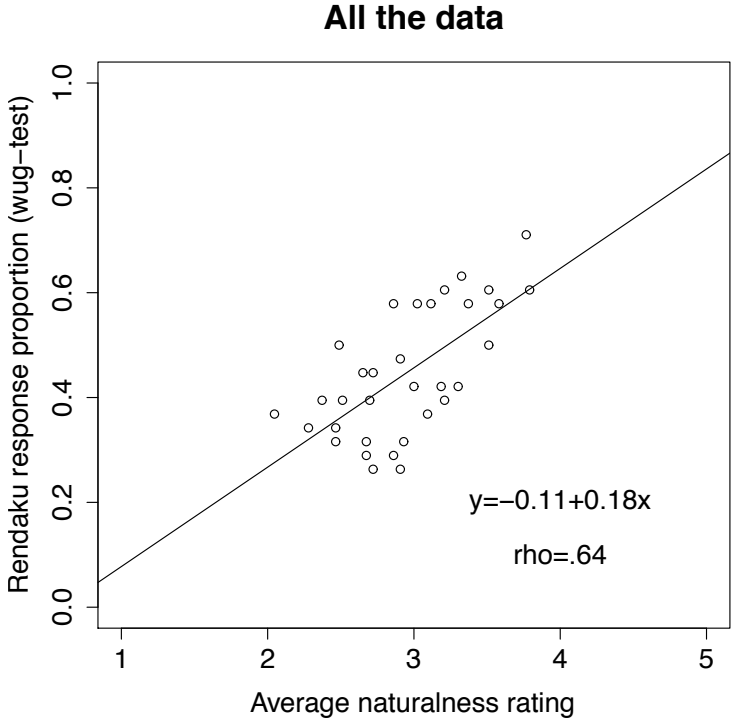

Figure 1: The correlation between the naturalness rating and rendaku response proportion, based on all the data. Each dot represents an item with its average naturalness rating on the x-axis and its average rendaku response proportion on the y-axis.

This analysis at first sight shows that the two tasks are correlated with another, and this result in and of itself may not be that surprising, given that the both tasks target the same phenomenon, rendaku.

However, something interesting emerges, when we separately analyze those items whose rendaku violates Lyman's Law and the other items which involve no Lyman's Law violations, as shown in Figure 2. For those items whose rendaku would not result in a Lyman's Law violation (shown with circles), there is still a tangible correlation between the two tasks ($\rho = 0.47$—the regression line shown with a solid line), although it does not reach statistical significance ($p = 0.14$). On the other hand, for those items whose rendaku results in a Lyman's Law violation (shown with squares), there is barely a correlation ($\rho = 0.11$—the regression line shown with a dashed line), which is not statistically significant ($p = 0.61$).

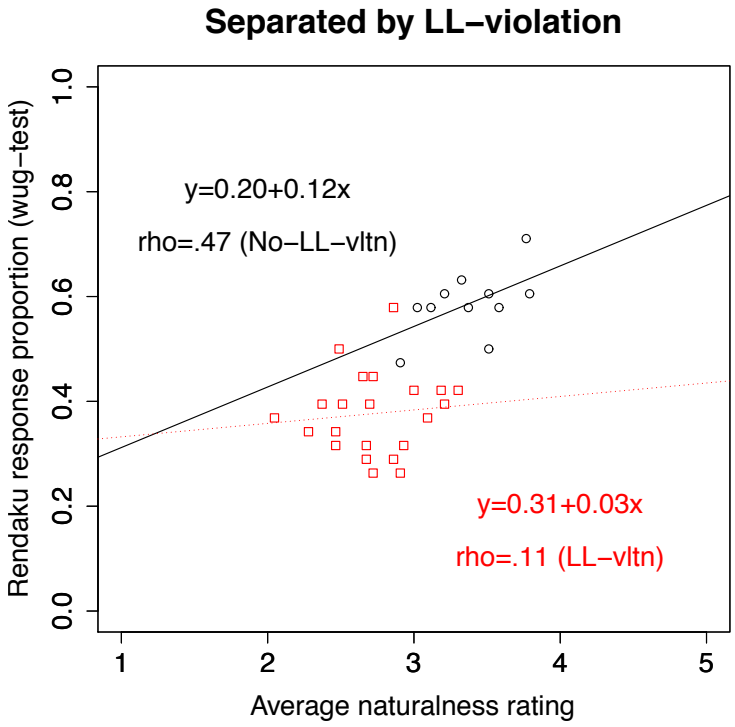

Figure 2: The correlation between the naturalness ratings and rendaku response proportions, separately by violation profiles of Lyman's Law. Circles with a solid regression line=no Lyman's Law violation; squares with a dotted regression line=Lyman's Law violations.

Looking closely at the data shows that there may be a significant difference between the two tasks, after all. The correlation that we observed in Figure 1 may thus be a spurious correlation in the sense that it arose from mixing up the two separate conditions.

Now looking at Figure 2, we observe that the y-axis—the wug-test—separates the two condi-

tions (circles and squares) better than the x-axis—the naturalness rating. In other words, there is not much overlap between the two conditions when the dots are projected on the y-axis, but there is some non-negligible amount of overlap when projected on the x-axis.

To further investigate this observation statistically, a linear discriminant analysis was run to see how each task—the naturalness rating and the wug test—separates the two conditions. This statistical technique finds an optimal boundary between the two groups, and calculates what percentage of the items are correctly categorized according to that optimal boundary. The linear discriminant analyses show that the wug-test can successfully distinguish 91% of the items in terms of Lyman's Law violation, whereas the accuracy rate for the naturalness rating is 82%, which is lower.

4 Discussion

Everything else being equal, then, the forced-choice wug-test is better at detecting a difference between Lyman's-Law-violating items and non-Lyman's-Law-violating items. This difference may come from the fact that in the naturalness judgment task, the participants evaluated only rendaku-undergoing items, whereas in the forced-choice wug-test, the participants compared both rendaku-undergoing forms and non-rendaku-undergoing forms. This finding is interesting because in the domain of syntactic experimentation, Sprouse & Almeida (2012b) found that grammaticality differences were most reliably detected when the participants were asked to compare the grammaticality of two sentences (referred to as "forced-choice" in their work). See also Daland et al. (2011) for a similar observation in phonological experimentation, although the paper does not go into details about this task effect in their experiments.

In summary, then, both the naturalness rating study and the forced-choice wug-test can reveal a difference between Lyman's Law violating items and non-Lyman's Law violating items. However, upon closer investigation, it seems that the forced-choice wug-test is more reliable when detecting the difference due to Lyman's Law.

This conclusion offers only a first step toward the general comparison of task effects in phonological experimentation, as the current study has three major limitations. First, the current analysis is based on cross-experimental comparisons: the two previous studies, although designed similarly, were run at separate times with different pools of participants. A follow-up study that would allow us to make within-subject comparisons would be desirable. Second, this sort of comparison should be made using phenomena other than rendaku. Third, the task effects in phonological judgment should be explored with other types of judgment experiments, such as magnitude estimation tasks (Bard et al., 1996), yes/no judgment tasks (Kawahara, 2013), and possibly free production tasks.

With these limitations explicitly noted, however, the current study has revealed intriguing similarities and differences between two types of phonological judgment experiments. It is hoped

that the current study will plant a seed for a research program which addresses which kind of phonological judgment experimentation is best suited to reveal our phonological knowledge.

Acknowledgements

This project is supported by JSPS Grant #26770147. I am grateful to Haruka Fukazawa, Hinako Masuda, Mayuki Matsui, Shin-ichiro Sano, Yoko Sugioka, Yuko Sugiyama, and two anonymous reviewers for careful comments on the previous version of the paper and/or suggestions on this general project. Thanks to Helen Stickney for her careful proofreading of the pre-final draft. All remaining errors are mine.

References

Bard, Ellen. G., Dan. Robertson, & Antonella Sorace (1996) Magnitude estimation of linguistic acceptability. *Language* **72**: 32–68.

Berent, Iris (2008) Are phonological representations of printed and spoken langauge isomorphic? Evidence from the restrictions on unattested onsets. *Journal of Experimental Psychology: Human Perception and Performance* **34**(5): 1288–1304.

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

Coetzee, Andries W., René Kager, & Joe Pater (2009) Introduction: phonological models and experimental data. *Phonology* **26**(1): 1–8.

Cowart, Wayne (1997) *Experimental syntax: Applying objective methods to sentence judgments*. London: Sage Publications.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis, & Ingrid Norrmann (2011) Explaining sonority projection effects. *Phonology* **28**(2): 197–234.

Dąbrowska, Ewa (2010) Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* **27**(1): 1–23.

Edelman, Simon & Morten H. Christianson (2003) How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* **7**(2): 60–61.

Ferreira, Fernanda (2005) Psycholinguistics, formal grammars and cognitive science. *The Linguistic Review* **22**: 365–380.

Gibson, Edward & Evelina Fedorenko (2010) Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* **14**(6): 233–234.

Goldrick, Matthew (2011) Utilizing psychological realism to advance phonological theory. In *The Handbook of Phonological Theory, 2nd Edition*, John A. Goldsmith, Jason Riggle, & Alan Yu, eds., Oxford: Blackwell-Wiley, 631–660.

Gouskova, Maria & Kevin Roon (2013) Gradient clash, faithfulness, and sonority sequencing effects in Russian compound stress. *Laboratory Phonology* **4**(2): 383–334.

Hill, Archibald (1961) Grammaticality. *Word* **17**: 1–10.

Ihara, Mutsuko, Katsuo Tamaoka, & Tadao Murata (2009) Lyman's Law effect in Japanese sequential voicing: Questionnaire-based nonword experiments. In *Current Issues in Unity and Diversity of Languages: Collection of the papers selected from the 18th International Congress*

of Linguists, The Linguistic Society of Korea, ed., Seoul: Dongam Publishing Co., Republic of Korea, 1007–1018.

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

Kawahara, Shigeto (2011b) Experimental approaches in theoretical phonology. In *The Blackwell companion to phonology*, Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, & Keren Rice, eds., Oxford: Blackwell-Wiley, 2283–2303.

Kawahara, Shigeto (2011c) Japanese loanword devoicing revisited: A rating study. *Natural Language and Linguistic Theory* **29**(3): 705–723.

Kawahara, Shigeto (2012) Lyman's Law is active in loanwords and nonce words: Evidence from naturalness judgment experiments. *Lingua* **122**(11): 1193–1206.

Kawahara, Shigeto (2013) Testing Japanese loanword devoicing: Addressing task effects. *Linguistics* **51**(6): 1271 – 1299.

Kawahara, Shigeto (2015) Can we use rendaku for phonological argumentation? *Linguistic Vanguard* .

Kawahara, Shigeto (to appear) Psycholinguistic studies of rendaku. In *Perspectives on rendaku: Sequential voicing in Japanese compounds*, Timothy Vance & Mark Irwin, eds., Berlin: Mouton.

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

Lasnik, Howard & Mamoru Saito (1984) On the nature of proper government. *Linguistic Inquiry* **15**: 235–189.

Lyman, Benjamin S. (1894) Change from surd to sonant in Japanese compounds. *Oriental Studies of the Oriental Club of Philadelphia* : 1–17.

Macmillan, Neil & Douglas Creelman (2005) *Detection Theory: A User's Guide. 2nd Edition*. Mahwah: Lawrence Erlbaum Associates Publishers.

Marantz, Alec (2005) Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* **22**: 429–445.

Myers, James (2009) Syntactic judgment experiments. *Language and Linguistic Compass* **3**(1): 406–423.

Ohala, John J. (1986) Consumer's guide to evidence in phonology. *Phonology* **3**: 3–26.

Perkins, Jeremy (2014) *Consonant-tone interaction in Thai*. Doctoral dissertation, Rutgers University.

Phillips, Colin (2009) Should we impeach armchair linguists? In *Japanese/Korean Linguistics 17*, Shoichi Iwasaki, ed., Stanford: CSLI, 49–64.

Pullum, Geoffrey K. (2013) The central question in comparative syntactic metatheory. *Mind and Language* **28**(4): 492–521.

Riemer, Nicholas (2009) Grammaticality as evidence and as prediction in a Galilean linguistics. *Language Sciences* **31**: 612–633.

Schütze, Carlson (1996) *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Schütze, Carlson (2011) Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* **2**(2): 206–211.

Spencer, N.J. (1973) Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* **2**(2): 83–93.

Sprouse, Jon & Diogo Almeida (2012a) Assessing the reliability of textbook data in syntax:

Adger's Core Syntax. *Journal of Linguistics* **48**: 609–652.
Sprouse, Jon & Diogo Almeida (2012b) Power in acceptability judgment experiments. Ms. University of California, Irvine.
Ueyama, Ayumi (2010) Model of judgment making and hypotheses in generative grammar. In *Japanese/Korean Linguistics 17*, Shoichi Iwasaki, Hajime Hoji, Patricia M. Clancy, & Sung-Ock Sohn, eds., Stanford: CSLI Publications, 27–47.
Vance, Timothy (1980) The psychological status of a constraint on Japanese consonant alternation. *Linguistics* **18**: 245–267.
Vance, Timothy (2007) Have we learned anything about *rendaku* that Lyman didn't already know? In *Current issues in the history and structure of Japanese*, Bjarke Frellesvig, Masayoshi Shibatani, & John Carles Smith, eds., Tokyo: Kurosio, 153–170.
Vance, Timothy (2015) Rendaku. In *The Handbook of Japanese Language and Linguistics: Phonetics and Phonology*, Haruo Kubozono, ed., Berlin: Mouton de Gruyter.
Vance, Timothy & Mark Irwin, eds. (to appear) *Perspectives on rendaku: Sequential voicing in Japanese compounds*. Mouton.
Wasow, Thomas & Jennifer Arnold (2005) Intuitions in linguistic argumentation. *Lingua* **115**: 1481–1496.
Yu, Alan & Hyunjung Lee (2014) The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *Journal of the Acoustical Society of America* **136**(1): 382–388.