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

Abstract

A growing body of linguistic studies is now deploying linguistic judgment experiments to probe both syntactic and phonological knowledge. A research question arises as to what kind of judgment format is useful to probe our linguistic knowledge. Against this 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 previous 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 wug-test reveals the influence of Lyman’s Law on rendaku more clearly than the naturalness judgment experiment. While the impact of the results may be limited and modest, this study provides a first step toward understanding how different tasks in phonological experimentation relate to one another. It is hoped that the current study will plant a seed for a research program which addresses what 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 what native speakers of a particular language feel about a set of sentences (in syntax) or phonological structures or 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), and has now received renewed interests from various perspectives (see Cowart 1997; Dąbrowska 2010; Gibson & Fedorenko 2010; Goldrick 2011; Ohala 1986; Kawahara 2011c; Marantz 2005; Myers 2009; Phillips 2009; Schütze 1996, 2011; Sprouse & Almeida 2012a,b; 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 as unreliable as one would argue, but at the same time experimentation complements this traditional approach in a number of useful ways. Therefore, we witness the rise of interests in eliciting linguistic data by way of experimentation using a large number of naive participants, both in syntax and phonology.

There are many conceivable types of experiments that one can do to obtain grammatical-acceptability judgments from naive speakers in experimental settings. One issue that arises in this theoretical context is how different tests are related to one another. This question is important partly because different studies deploy different types of experiments, and also because we want to know which test is more powerful in detecting grammatical differences under investigation.\(^1\) Sprouse & Almeida (2012b) present an extensive comparison between different types grammaticality judgment experiments in syntax, but there is nothing comparable in the domain of phonological judgment experimentation. This paper thus provides a more modest study in the domain of phonology. This work can also be understood as a part of a larger research enterprise addressing task effects in phonetic and phonological experimentation (Berent, 2008; Kawahara, 2013; Yu & Lee, 2014).

As a part of this research enterprise on linguistic experimentation, this paper reports a result of comparing a 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 elicit the plural form of nonce words, including the famous nonce word *wug*. This methodology has been deployed in many experimental work in phonology (see Kawahara 2011b for a review). Another test is a naturalness judgment task, which corresponds closely to the standard practice in syntax, where sentences are assigned a scale of grammaticality; e.g. * (totally ungrammatical) > *? (ungrammatical) > ?? (very questionable) > ? (questionable) > unmarked (grammatical). This sort of task was also deployed in phonological work as well (e.g. Daland et al. 2011; Gouskova & Roon 2013; Kawahara 2011a,c).

This paper takes up on comparing these two tasks, using rendaku as a case study.\(^2\) Rendaku is a morphological process whereby the initial consonant of the second member of compounds appear as voiced (e.g. /tako/ ‘octopus’ → /oo-dako/ ‘big octopus’).\(^3\) The application of rendaku is not automatic in the sense that many factors affect the applicability of rendaku (Vance, to appear; 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 (/toko/ ‘lizard’ → /oo-tokage/ ‘big lizard’)—and this blockage is known as Lyman’s Law (Lyman 1894

---

\(^1\)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.

\(^2\)There are other types of judgment experiments in linguistics, including a magnitude estimation task (Bard et al., 1996) and a binary yes/no task (Kawahara, 2013).

\(^3\)For the sake of simplicity, this paper uses phonemic transcription rather than IPA transcription (Vance, 2008).
et seq.). Two experiments in the past have explored rendaku and the effect of Lyman’s Law experimentally using nonce words, one with wug-test format (Kawahara & Sano, to appear) and one with naturalness judgment format (Kawahara, 2012), with an overlapping set of stimuli. These data allow us to compare the two types of tests, which the current study aims to do.

2 The two previous studies

One study used a naturalness judgment task to explore the effect of Lyman’s Law on rendaku (Kawahara, 2012). The study used a 5-point Likert scale from “very natural” to “very unnatural”, and asked native speakers of Japanese to rate the naturalness of forms that underwent rendaku (e.g. “How naturalness does /X+gotedon/ from /kotoni/ sound for you?”). The study found that Japanese speakers rate rendaku less naturally when it violates Lyman’s Law than when it does not (e.g. /X+gitage/ from /kitage/ is worse than /X+gimane/ from /kimane/). Another study explored a similar issue, using a forced-choice wug-test (Kawahara & Sano, to appear), a task also known as “head-to-head” (Daland et al., 2011). In that experiment, the participants were asked to choose the better inflected form, either with rendaku or without rendaku (“Which one of the following sounds better, /X+gotedon/ or /X+kotoni/, when /X/ and /kotoni/ is combined?”). 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 in both Kawahara (2012) and Kawahara and Sano (to appear).

<table>
<thead>
<tr>
<th>No LL violation</th>
<th>LL violation</th>
</tr>
</thead>
<tbody>
<tr>
<td>hinumi</td>
<td>haboke</td>
</tr>
<tr>
<td>honara</td>
<td>hekazu</td>
</tr>
<tr>
<td>katon</td>
<td>hemiga</td>
</tr>
<tr>
<td>kikake</td>
<td>hobasa</td>
</tr>
<tr>
<td>kikade</td>
<td>hogore</td>
</tr>
<tr>
<td>sekato</td>
<td>hokada</td>
</tr>
<tr>
<td>semaro</td>
<td>kabomo</td>
</tr>
<tr>
<td>sutane</td>
<td>kamagi</td>
</tr>
<tr>
<td>tamuma</td>
<td>kidake</td>
</tr>
<tr>
<td>taruna</td>
<td>kitage</td>
</tr>
<tr>
<td>tatuka</td>
<td>kobono</td>
</tr>
</tbody>
</table>

The two studies used a set of overlapping nonce words for stimuli, as shown in Table 1, which allows us to compare the two different tasks deployed in these experiments. There are 11 items

---

4For previous experimental studies on rendaku and Lyman’s Law, see Vance (1980) and Ihara et al. (2009)
that would not violate Lyman’s Law after rendaku, and there are 22 items that would violate Ly-
man’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.

For each item, average naturalness rating scores and average rendaku response proportions are
calculated across all the participants (forty-three for the naturalness experiment and thirty-eight for
the 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 is more likely to undergo rendaku in the wug-test. 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 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 violate Lyman’s Law and the other items which involve no Lyman’s Law violations, as shown in Figure 2. For those items whose rendaku involve no 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 \((p = 0.14)\). For those items whose rendaku result in 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](image)

**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.

Closely looking at the data then, after all, there may be a significant difference between the two tasks. 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 how many percentages 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 wug-test is a 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 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 grammatically 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).

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 a closer investigation, it seems that the forced-choice wug-test is more powerful in 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 two limitations. First, this task 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 with-subject comparisons would be desirable. Second, this sort of comparison should be made using phenomena other than rendaku, and moreover 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 task.

With these limitations explicitly noted, however, the current study offers intriguing similarities and differences between two types of phonological judgment experiment. It is hoped that the current study will plant a seed for a research program which addresses what kind of phonological
judgment experimentation is best to reveal our phonological knowledge.

References


