

# Japanese speakers can infer specific sub-lexicons using phonotactic cues\*

Shigeto Kawahara and Gakuji Kumagai  
Keio University and Kansai University

## Abstract

Phonotactic restrictions do not usually hold uniformly across the entire lexicon of a particular language, and thus the lexicon of a natural language is often assumed to be organized into different sub-lexicons. A question that arises is how specific these sub-lexicons can be. A classic, conservative approach is to posit only broad distinctions, such as the distinction between native words vs. borrowed words. An alternative approach is to posit more specific miniature lexicons, such as a set of morphemes that show a particular morpheme-specific pattern or a set of loanwords from a particular language. With this general theoretical issue in mind, this paper first points out that there are phonotactic restrictions that can cue a very specific class of words in Japanese; e.g. a geminate /rr/ occurs almost exclusively in loanwords from Italian. Building upon these novel observations, the current experiments tested whether Japanese speakers can infer specific word classes such as “snack names”, “Italian restaurant names” and “German names” based on particular phonotactic cues. The results of the two experiments support the idea that the lexicon of a natural language can be organized into very specific sub-lexicons, at least going beyond the often posited native vs. loanword distinction, and that these specific sub-lexicons can be psychologically real.

---

\*We would like to thank Hiro Katsuda for his advice on using the BCCWJ to confirm the initial intuitions we had, and Tomoko Tanaka for her help with the analyses. We also received helpful feedback on the design of the experiments from the participants of a seminar that the first author gave at International Christian University, especially Honoka Endo, Rina Furusawa and Haseru Ida. All remaining errors are ours.

# 1 Introduction

## 1.1 General theoretical background

Phonotactic restrictions—restrictions on which sounds can appear in what position and how these sounds can be arranged—do not often hold uniformly across the entire lexicon of a given language. Typical cases come from differences between native words and loanwords, in which particular sounds and/or sound sequences are often only allowed in loanwords but not in native words. For instance, in English most if not all words with stress on their final syllables are loanwords from French (Gelbart 2005). Japanese also exhibits such patterns, in which the lexicon is considered to consist of several lexical strata: native words, Sino-Japanese, recent loanwords and mimetic words (Ito & Mester 1995, 1999, 2008). These sub-lexicons are subject to different sets of phonotactic restrictions; e.g. voiced obstruent geminates are allowed only in recent loanwords but not in other types of words; similarly, Sino-Japanese morphemes and mimetic roots are maximally bimoraic (Ito & Mester 1996), but no such size restrictions appear to hold in native words or recent loanwords. See Ito & Mester (1995) for a review of similar examples from other languages.

Experimental studies have shown that such phonotactic restrictions that hold specifically to a subset of the lexicon are, at least in some cases, psychologically real, in that they can influence the perception of ambiguous acoustic signals—listeners can use these cues to decipher the lexical class of the stimuli (Gelbart 2005; Gelbart & Kawahara 2007; Moreton & Amano 1999). For instance, Sino-Japanese does not allow either singleton /p/ or long /aa/, whereas loanwords allow both. When Japanese listeners hear a nonce word with /p/, they tend to perceive the /a/-/aa/ continuum more likely to be long than the control condition (Moreton & Amano 1999). On the other hand, recent loanwords generally do not exhibit a palatalized /ɾ/, which is very common in Sino-Japanese compounds. Upon hearing nonce words with a palatalized /ɾ/, Japanese listeners are biased toward perceiving the continuum to be short compared the control condition. This experiment by Moreton & Amano (1999) thus shows that listeners can use phonotactic cues—such as singleton /p/ and palatalized /ɾ/—to decipher which lexical class the incoming acoustic signal belongs to, which in turn affects their short /a/ vs. long /aa/ perception.

One question that arises in this context is how fine-grained the sub-division of the lexicon can be in natural languages. The “classic” (and also conservative) view is to postulate only very general divisions, perhaps with independent etymological motivations. Take the case of Japanese, for example—Ito & Mester (1995, 1999, 2008) posit four general strata (native words, Sino-Japanese, foreign words and mimetics). Similarly, in English, we could posit a very general distinction between native words and recent loanwords (Gelbart 2005), and perhaps within “native” words, a distinction between Germanic roots vs. Latinate roots (Chomsky & Halle 1968).

On the other hand, there is an alternative, less-conservative proposal. For example, the re-

36 search approach now widely known as “co-phonology” argues that there can co-exist many types  
37 of morphologically-conditioned phonological patterns in a single language, and posits that there  
38 can be as many number of phonological sub-systems as the number of such morpheme-specific  
39 patterns (Anttila 2002; Inkelas et al. 1996, 1997; Inkelas & Zoll 2007; Orgun 1996; Sande 2020  
40 among many others—see also Pater 2005). A similar idea is proposed by a line of work dubbed  
41 “sublexical phonology” (Becker & Gouskova 2016; Blake & Becker 2015; Gouskova & Ahn 2024;  
42 Gouskova et al. 2015), which posits that “learning lexically specific morphological and phonologi-  
43 cal rules involves separating the lexicon into sublexicons. Phonological generalizations about the  
44 application of such rules are encoded in part as phonotactic grammars learned over sublexicons.”  
45 (Gouskova & Ahn 2024: 6).<sup>1</sup>

46 With this general theoretical debate in mind, the current study explores how specific such sub-  
47 lexicons can be. Our case study is based on novel observations about the phonotactic restrictions  
48 that are associated with very specific parts of the Japanese lexicon. The experiments reported  
49 below examined whether Japanese speakers are only sensitive to the broad sub-lexicons that  
50 have been traditionally posited (i.e. native, Sino-Japanese, foreign and mimetics), or alternatively,  
51 whether they are sensitive to more fine-grained distinctions. This question is addressed through  
52 the investigation of the three specific phonological observations, explained in detail in the next  
53 subsection.

54 Before we proceed to the specifics, one general remark is in order. Japanese offers a particu-  
55 larly interesting testing ground to address this general issue, because not only are the traditional  
56 distinctions between native words, Sino-Japanese words and recent loanwords clearly motivated  
57 on phonological grounds (Ito & Mester 1995, 1999, 2008), Japanese uses different orthographic  
58 systems for different word classes; i.e. the *hiragana* system for native words, the *katakana* system  
59 and mimetics, and the *kanji characters* for native and Sino-Japanese words. In addition, learning  
60 these lexical classes is a part of elementary school education. On the one hand, the use of the  
61 *katakana* orthography, for example, may give rise to a sense of uniformity among recent loan-  
62 words. On the other hand, when there is a phonotactic cue that signals a specific set of loanwords  
63 from a particular language, that cue may indeed be psychologically associated with that specific  
64 sub-lexicon.

## 65 **1.2 The specific hypotheses tested**

66 The current study tested three phonotactic tendencies that can potentially cue a specific sub-  
67 lexicon in Japanese, listed in (1):

- 68 (1) Phonotactic tendencies that can potentially cue a specific sub-lexicon

---

<sup>1</sup>For proposals regarding how such sub-lexicons are learned, see Morita (2018), Pater (2005) and Shaw (2006).

- 69 a. Singleton /p/ often appears in snack names.
- 70 b. Geminate /rr/ appears almost exclusively in loanwords from Italian.
- 71 c. Geminate /hh/ appears almost exclusively in loanwords from German.

72 The first observation is reported and discussed in a recent popular science book (Kawahara 2023),  
73 which grew out of the dialogue-based linguistics lecture that the author gave to elementary school  
74 children. There, one student pointed out that there are many snack names that contain singleton  
75 /p/ in Japanese (e.g. /papiko/, /poiuru/, /pokii/, /porinkii/ and /pai-no-mi/), and asked why.  
76 Kawahara (2023) did not offer a quantitative backup of this observation, but assuming that /p/  
77 is indeed overrepresented in snack names in Japanese, he speculated that singleton /p/ is used  
78 to represent European—as opposed to traditional, Japanese—snack names, because singleton /p/  
79 appears only in loanwords in Japanese (Ito & Mester 1995, 1999, 2008); in a sense, singleton /p/  
80 cues “foreignness” in Japanese, as the experiment by Moreton & Amano (1999) demonstrates,  
81 which the product companies may be taking an advantage of.<sup>2</sup> He additionally entertains the  
82 possibility that /p/ is preferred in snack names because /p/ is known to convey a sound symbolic  
83 meaning of “cuteness” in Japanese (Kawahara 2019; Kumagai 2019). Regardless of whether these  
84 conjectures offered by Kawahara (2023) are on the right track, this dialogue raised an interesting  
85 question that should be empirically tested: whether Japanese speakers indeed associate singleton  
86 /p/ with snack names, i.e., whether the observation made by the elementary school student is  
87 psychologically real or not.

88 As for the second hypothesis, we are not aware of any previous systematic study on the  
89 observation, but it was instead initially based on our intuitions as native speakers of Japanese.  
90 The native phonology of Japanese does not allow geminate /rr/ (Kawahara & Pangilinan 2017;  
91 Kuroda 1965; Labrune 2014), perhaps because it is a flap which is intrinsically characterized by  
92 its short duration. However, geminate /rr/s appear in recent loanwords, and they appear mainly  
93 in names for Italian foods (e.g. /huzirri/ ‘fusilli’, /farufarre/ ‘Farfalle’ and /kitarra/ ‘chitarra’), in  
94 which case geminate /rr/ tends to be produced with lateral-like articulations (see also Morimoto  
95 2020 for the articulatory study of geminate /rr/ in Japanese). The use of /rr/ in Japanese was likely  
96 prompted by the fact that Italian has geminate /ll/ and /rr/ as their phonemes (De Benedetto &  
97 De Nardis 2021), which are clearly represented in their orthography as well.

98 To more objectively access our initial intuition in a quantitative manner, we have consulted  
99 the frequency list of the Balanced Corpus of Contemporary Written Japanese (BCCWJ),<sup>3</sup> which

---

<sup>2</sup>This idea can be formally captured by a family of EXPRESS(x) constraints proposed by Alderete & Kochetov (2017), which requires that certain sounds be used to express a particular semantic notion; e.g. use palatal consonants and/or high front vowels to express smallness. It is not hard to imagine that a stochastic version of a constraint like EXPRESS(P)FORSNACK is at work here.

<sup>3</sup><https://clrd.ninjal.ac.jp/bccwj/en/index.html>, consulted May 2024. Since the analysis is based on a published corpus, we are not able to make our analysis file publicly available, but the analysis files are

100 revealed 64 types of words containing /ʀʀ/. Among those, most, if not all, of them are borrowed  
101 from Italian, although there were 9 words whose origin was unclear after searching with Google  
102 and other tools. The only exception is /aruhurra/, which may come from ‘al-hurra’ in Arabic; in  
103 addition to this, albeit not being found in the corpus, we think that /arraa/ is a possible pronun-  
104 ciation of the word ‘Allah.’ However, these two seem to be the sole exceptions, perhaps because  
105 Japanese borrowed more words from Italian than from Arabic. At any rate, it seems safe to  
106 conclude that /ʀʀ/ appears mostly if not exclusively appear in loanwords from Italian. Having  
107 established this connection, the experiment reported below addressed whether Japanese speak-  
108 ers do indeed associate /ʀʀ/ with the sub-lexicon of “Italian names,” thereby also testing whether  
109 such a very specific sub-lexicon can be psychologically real.

110 The third hypothesis is concerned with a certain pattern found in loanwords from German—  
111 Japanese borrowed many medical, philosophical and other technical terms from German, and  
112 these borrowings can contain geminate /hh/. According to our intuition, such words with /hh/  
113 are largely limited to those loanwords from German, such as /bahha/ ‘Bach’, /mahha/ ‘mach’ and  
114 /riibihhi/ ‘Liebig’, which are also characterized by copy vowel epenthesis in word-final position  
115 (Kawahara 2007). These geminates, surrounded by the identical vowels, do not occur in native  
116 words, and rarely occur in other types of loanwords.

117 We have searched BCCWJ for those words containing a geminate /hh/, which revealed 73  
118 types of words, many of which are based on German words. There were 9 words, such as  
119 /boitʃahha/ and /uratahha/, whose origin was not very clear even after searching with Google.  
120 The non-German origin loanwords containing /hh/ that we found in the corpus were /sutahhu/  
121 ‘stuff’, /hurahhu/ ‘fluff’ and /sunahhu/ ‘snuff’, which are borrowed from English, as well as  
122 /buhha/ ‘Tunisian alcohol’ and /ʃorusutahhu/ ‘Falstaff (English/Italian) opera name’, but these  
123 tokens do not have identical vowels surrounding the geminate. The truly exceptional example  
124 was /gohho/ ‘Gogh’ (and another token of a compound including this word), which is a Dutch  
125 name.<sup>4</sup> Thus, at least many of the relevant tokens were of the German origin. There thus seems  
126 to be some connection between geminate /hh/ and German names, which is useful for addressing  
127 the question of whether a sub-lexicon that is as specific as German names can be psychologically  
128 real in the minds of Japanese speakers.

---

available upon request.

<sup>4</sup>It is interesting and telling that both authors of the paper were considering /gohho/ to be a German artist, until pointed out by Tomoko Tanaka who helped us analyzed the data

## 129 2 Experiment 1

### 130 2.1 Method

#### 131 2.1.1 Task

132 The experiment was designed to address whether Japanese speakers associate /p/, /rr/ and /hh/  
133 with snack names, Italian names, and German names, respectively. In the current experiment,  
134 the participants were presented with one stimulus item, and were asked to judge whether that  
135 item is better suited as a name of one genre (=the intended category) or the other (=the control  
136 category).

137 For the first hypothesis (the connection between /p/ and snack names), the participants were  
138 asked whether each name is an European snack name or a cosmetic brand name, the latter of  
139 which is used for comparison because most cosmetic names in Japan are loanwords, just like  
140 European snack names, and are usually written with the *katakana* orthography. For the second  
141 hypothesis (the connection between /rr/ and Italian names), the participants were asked wether  
142 each stimulus is a name for an Italian restaurant or a name for a French restaurant. These two  
143 types of restaurants are both common in the current Japanese community, and since French does  
144 not have geminate liquids, no borrowed words from French in Japanese contain /rr/. For the  
145 third hypothesis (the connection between /hh/ and German names), the participants were asked  
146 wether a given name is a German celebrity name or an English celebrity name.

#### 147 2.1.2 Stimuli

148 The list of the stimuli is shown in Table 1. Since we used the Buy Response function of Survey-  
149 Monkey (see below), we were limited to have 50 questions, including one question to present the  
150 consent form and one question to check qualification for participation. Therefore we had 8 pairs  
151 of items for the first two hypotheses ( $8 \times 2 \times 2$ ) and 7 pairs of items for the third hypothesis ( $7$   
152  $\times 2$ ).

Table 1: The list of nonce words used in the experiment.

	Target	Control
/p/=snack	/pariko/ /pasomi/ /penaro/ /posone/ /pamore/ /parase/ /pesemo/ /poniru/	/tarikō/ /tasomi/ /tenaro/ /tosone/ /kamore/ /karase/ /kesemo/ /koniru /
/rr/=Italian	/metorra/ /makorro/ /nesorra/ /nokirre/ /temerro/ /tamirra/ /kanorri/ /tonorre/	/metobba/ /makobbo/ /nesodda/ /nokidde/ /temeddo/ /tamigga/ /kanoggi/ /tonogge/
/hh/=German	/bohho/ /kuhhu/ /gehhe/ /gahha/ /rehhe/ /bihhi/ /nahha/	/boppo/ /kuppu/ /geppe/ /gatta/ /rette/ /bikki/ /nakka/

153 For the first hypothesis, the target items began with /p/, whereas the control items began with  
 154 either /t/ or /k/. The rest of the words was identical between the target items and the control items.  
 155 We avoided using high vowels after /t/, because they may cause affrication of the preceding stop  
 156 (Vance 2008). For the second hypothesis, the target items included /rr/ whereas the control items  
 157 contained voiced obstruent geminates, which signaled that control items were also loanwords (Ito  
 158 & Mester 1995, 1999, 2008). For the third hypothesis, the target items contained /hh/, whereas  
 159 the control items contained voiceless stop geminates. We included geminates in the control items  
 160 for the second and third hypotheses to make sure that it is not a mere presence of any kind of  
 161 geminates that cue a particular sub-lexicon.

### 162 2.1.3 Participants

163 The experiment was conducted online using SurveyMonkey. The participants were collected  
 164 using the Buy Response function offered by SurveyMonkey. A total of 162 native speakers of

165 Japanese, who confirmed that they have not studied either sound symbolism (related to /p/ or  
166 not) or Japanese phonetics/phonology, completed the experiment.

#### 167 **2.1.4 Procedure**

168 In the instructions, the participants were told that in each trial, they are given one name and two  
169 categories and were asked to choose which category better fits that name. Example questions  
170 were thus, “given /pariko/, is the name better for a European snack or a cosmetic product?”  
171 and “given /metorra/, is the name better for an Italian restaurant name or a French restaurant  
172 name?” Each stimulus was presented in isolation, not in a pair, i.e. the experiment was a forced-  
173 choice task, not a two-alternative forced choice (2AFC) task. The stimuli were all written in  
174 the *katakata* orthography, which is used primarily for loanwords in the Japanese orthographic  
175 system. Although the stimuli were presented in the written form, the participants were asked to  
176 produce each form before they register their response. The order of the stimuli was randomized  
177 by SurveyMonkey, with each participant assigned a uniquely randomized order.

#### 178 **2.1.5 Statistical analyses**

179 For statistical analyses, we fit a Bayesian mixed effects logistic regression model, using the *brms*  
180 package (Bürkner 2017) implemented in R (R Development Core Team 1993–). For accessible  
181 introduction to Bayesian modeling, we would like to refer to the readers to Franke & Roettger  
182 (2019), Kruschke (2014), Kruschke & Liddell (2018), McElreath (2020), and Vasishth et al. (2018).  
183 Simply put, Bayesian analyses take a prior distribution and the obtained data to yield a posterior  
184 distribution of a parameter that we would like to estimate.

185 One common way to interpret the results of Bayesian regression models is to examine the  
186 middle 95% of the posterior distribution of an estimate parameter,  $\hat{\beta}$ , which is known as a 95%  
187 credible interval (also known as high density interval). If its credible interval does not include  
188 zero, then that effect can be considered to be meaningful. However, in Bayesian analyses, we  
189 do not need to be bound by the “credible” vs. “non-credible” dichotomy, unlike in a frequentist  
190 statistical testing with a strict “significant” vs “non-significant” dichotomy. That is, another way  
191 to interpret the results of Bayesian regression models is to calculate how many samples of the  
192 coefficient of interest are in the expected direction in the posterior distribution. In the current  
193 paper, we present both measures to interpret the results.

194 The details of the current model specifications were as follows. The dependent variable was  
195 whether each item was chosen as the target category name (“snack names”, “Italian names”, “Ger-  
196 man names”) or not. The main independent variable is the fixed factor encoding whether each  
197 item was the target (containing the phonotactic cue) or the control (not containing the phono-  
198 tactic cue). We included a random intercept for items and a random intercept and slope for



199 participants associated with that fixed factor.

200 The prior distributions were a Normal(0, 1) weakly informative prior for the intercept (Lemoine  
201 2019) and a Cauchy prior with scale of 2.5 for the slope (Gelman et al. 2018). Four chains with 2,000  
202 iterations were run, and the first 1,000 iterations from each chain were disregarded as warmups.  
203 All the  $\hat{R}$ -values were 1.00 and no divergent transitions were detected. The raw data, the R Mark-  
204 down file with the R syntax as well as the posterior samples are available in an OSF repository.<sup>5</sup>

## 205 2.2 Results

### 206 2.2.1 Snak names

207 Figure 1 shows the results concerning the first hypothesis, which is the violin plot representing  
208 the distributions of “snack responses ratios” for control conditions (those that begin with either  
209 /t/ or /k/) and for the target items (=those that begin with /p/). Transparent circles, slightly jittered  
210 to avoid overlap, represent averaged responses for each condition from each participant. Solid red  
211 circles are the grand averages in each condition and the red bars around the circles represent the  
212 bootstrap 95% confidence intervals around these averages, calculated by the `ggplot` package  
213 (Wickham 2016).

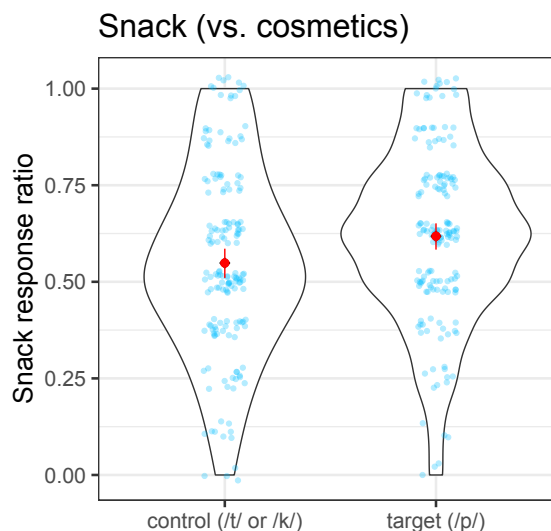


Figure 1: The comparison between the control and the target items in terms of the probabilities of being chosen as snack names, as opposed to cosmetic names.

214 The grand averages for the control and the target conditions were 0.55 vs. 0.62, respectively,  
215 which shows that the items beginning with /p/ were more likely to be judged as snack names

<sup>5</sup><https://osf.io/97zc5/?viewonly=4efac77f1b1a4e7c80798f7efbe86160>

216 than the control items.

217 The result of the Bayesian analysis shows that the central coefficient estimate ( $\hat{\beta}$ ) of the dif-  
218 ference between the control and the target is 0.33, suggesting that target items with singleton /p/  
219 induced more the snack responses than control items with /t/ or /k/. The 95% credible interval of  
220 this coefficient is [-0.09, 0.75]. Although this interval includes 0, it is heavily skewed toward the  
221 positive values, and the posterior probability of this coefficient being positive ( $\hat{\beta} > 0$ ) is 0.95.

222 With this result, it seems safe to conclude with a reasonable amount of confidence that  
223 Japanese speakers indeed associate names with /p/ with snack names—the observation discussed  
224 in Kawahara (2023) is corroborated by the current behavioral experiment.

### 225 2.2.2 Italian names

226 To assess the second hypothesis, Figure 2 shows the violin plot of the “Italian responses ratios” for  
227 the control condition and target condition, the latter of which contained /rr/. The grand average  
228 was 0.64 for the control condition vs. 0.66 for the target condition.

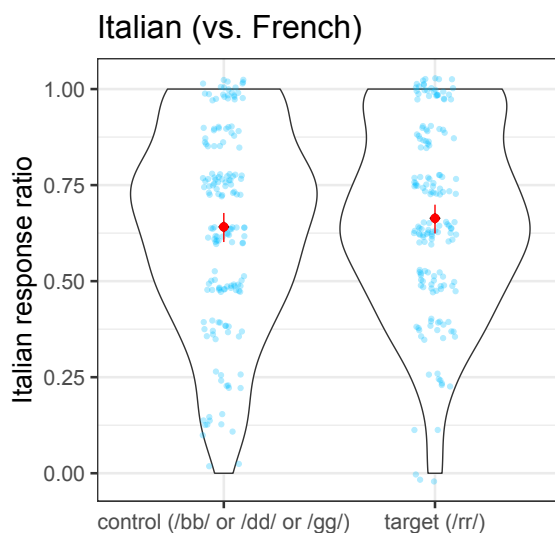


Figure 2: The comparison between the control and the target items in terms of the probabilities of being chosen as Italian restaurant names as opposed to French restaurant names.

229 The difference is in the expected direction—those that contain /rr/ were slightly more likely  
230 to be judged to be Italian names than those that contain voiced obstruent geminates. However,  
231 the magnitude of this difference is very small. The central coefficient estimate of this difference is  
232 0.11, with its 95% credible interval being [-0.13, 0.35]. The posterior probability of this coefficient  
233 being positive is 0.83. The evidence for this difference thus appears to be weak or at best modest.

234 **2.2.3 German names**

235 Figure 3 shows the probability distribution of the German response ratios for the stimuli that  
236 were intended to address the third hypothesis. The grand average for the control items was  
237 0.60 and the average for the target items is 0.71, suggesting that those nonce items containing  
238 /hh/, surrounded by the same vowels, were more likely to be judged as German celebrity names.  
239 The central estimate of this coefficient  $\hat{\beta}$  for this difference is 0.67, with its 95% credible interval  
240 being [-0.11, 1.40]. The probability of this coefficient being positive in the posterior distribution,  
241  $p(\hat{\beta} > 0)$  is 0.95. These results suggest that Japanese speakers do indeed tend to associate /hh/  
242 with German names, as opposed to English names.

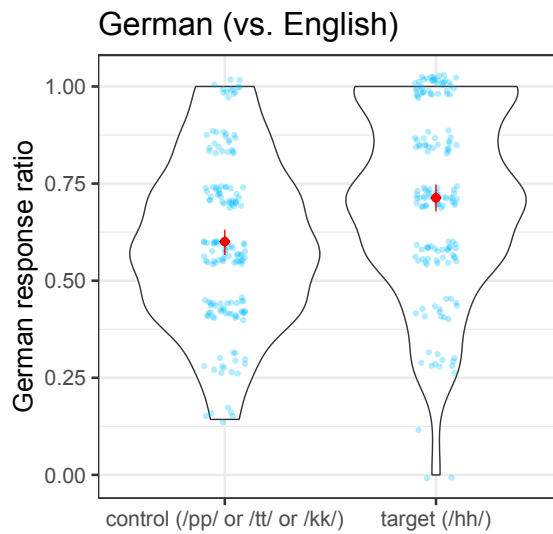


Figure 3: The comparison between the control and the target items in terms of the probabilities of being chosen as German celebrity names, as opposed to English names.

243 **2.3 Discussion**

244 The results of the first experiment show that we have reasonable amount of evidence to believe  
245 that Japanese speakers associate /p/ with snack names and /hh/ with German celebrity names,  
246 and we have weak or perhaps modest evidence showing that Japanese speakers associate /rr/ with  
247 Italian names. More generally speaking, the results show that Japanese speakers can associate  
248 nonce words with a particular phonotactic cue to a very specific lexical class, at least going beyond  
249 the traditional distinction between native words vs. loanwords.

250 The connection between /rr/ and Italian restaurant names, as revealed in this experiment,  
251 was not very robust, however. One possibility is that those who are not familiar with linguistics—  
252 which was exactly the pool of participants in this experiment—may not clearly distinguish French

253 names and Italian names. Another possibility is that the participants may have avoided associat-  
254 ing names with voiced obstruents with French names—if we look back the results in Figure 2, the  
255 control items (containing voiced obstruent geminates) were associated with Italian names with  
256 the probability that is much higher than chance (=0.64). And actually, with hindsight, there seems  
257 to be a good reason for them to have done so. Our post-experimental search shows, for example,  
258 that the website<sup>6</sup> which lists 131 popular French loanwords in Japanese, contains no instance of  
259 a French loanword containing a voiced obstruent geminate.

260 For the third hypothesis, an alternative interpretation was pointed out by our anonymous  
261 colleague—it may be the case that the participants were not so much familiar with German names,  
262 but instead they were only familiar with English names. As long as they have a general idea of  
263 how English loanwords look like, they may realize that /hh/ is absent in such words, and may  
264 have associated those words with /hh/ with “the non-English option.” This alternative, together  
265 with the possibility discussed in the preceding paragraph, point to a general methodological issue  
266 in Experiment 1: the participants may have been making judgments about the control categories  
267 (“no voiced obstruents in French names” and “no /hh/ in English names”) rather than the target  
268 categories.

269 To address this possibility, in the next experiment, we avoided comparing target categories  
270 and control categories, because after all, we are not directly interested in how control categories  
271 behave. Instead, we asked the participants to judge how suitable each name is for the three  
272 respective target categories.

273 In addition, after we ran Experiment 1, we realized that one of the control items for the third  
274 hypothesis, /geppe/, could have sounded too close to /gebberusu/ ‘Goebbles’—an (in)famous Ger-  
275 man historical figure—whose /bb/ can be devoiced, because of its OCP(voice) violation due to its  
276 co-occurrence with /g/ (Kawahara 2006). As a post-hoc analysis, we compared the German re-  
277 sponse ratio for /geppe/ and that of the other items, which showed the /geppe/ was judged to be  
278 a German name 81.4% of the time, while the average response for the other control items is 56.8%.  
279 The use of this particular item was thus not ideal, an issue that we also fixed in Experiment 2.<sup>7</sup>

---

<sup>6</sup><https://origamijapan.net/origami/2019/07/09/france-gairaigo/>, last access May 2024

<sup>7</sup>We refrain from running a new statistical analysis with /geppe/ excluded, to avoid HARKing (Hypothesizing After the Results are Known) (Kerr 1998).

## 3 Experiment 2

### 3.1 Methods

The crucial methodological differences between Experiment 1 and Experiment 2 were as follows. In Experiment 2, we presented each stimulus and asked how suitable each name is for the three respective target categories, using a 4-point Likert scale, with the following labels: 1=“not at all suitable”, 2=“not so suitable”, 3=“suitable” and 4=“very suitable.” Example questions are thus, “How suitable is /pariko/ as a snack name?” and “How suitable is /bohho/ as a German celebrity name?” We used the same set of stimuli as Experiment 1 (see Table 1), except that for the reason discussed above, we replaced /gehhe/ and /geppe/ with /gihhi/ and /gippi/, respectively. As with Experiment 1, the participants were asked to read each stimulus before registering their responses.

Like Experiment 1, we used the Buy Response function of SurveyMonkey, and the data from 162 native speakers of Japanese were collected. Since the responses were obtained using a Likert scale, we used an *ordinal* logistic regression. Other details of the statistical analysis were identical to that of Experiment 1. The files used for the statistical analyses for this experiment are also made available at the above-mentioned OSF repository.

### 3.2 Results

#### 3.2.1 Snack names

Figure 4 is a violin plot showing how suitable the control items (with /t/ or /k/) and the target items (with /p/) were judged as snack names. We observe that the latter items were judged to be more suitable (the grand averages were 2.53 vs. 2.86, respectively). The central estimate of this coefficient ( $\hat{\beta}$ ) for this difference is 0.86 with its 95% credible interval being [-0.44, 2.19]. The posterior probability of this coefficient being positive ( $\hat{\beta} > 0$ ) is 0.91.

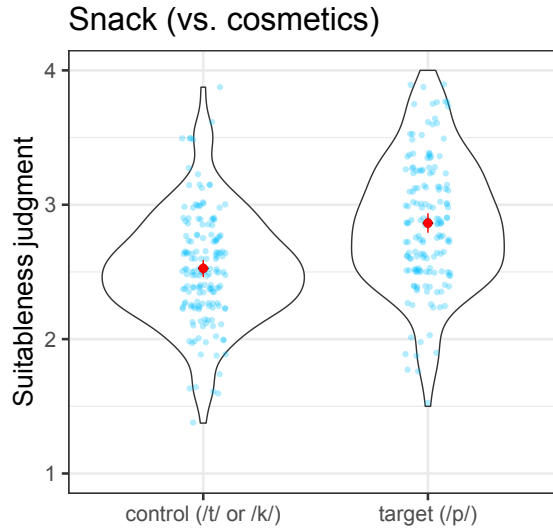


Figure 4: The comparison between the control and the target items in terms of how suitable they were judged as snack names.

### 303 3.2.2 Italian names

304 Figure 5 shows how suitable the control items (with a voiced obstruent geminate) and the target  
 305 items (with a geminate /ʀʀ/) were judged as the names of Italian restaurants. Compared to the  
 306 results of Experiment I, we observe a rather substantial difference between the two conditions  
 307 (2.22 vs. 2.94). The central coefficient estimate for this difference was 2.12, with its 95% credible  
 308 interval being [1.56, 2.69]. All the posterior samples were positive ( $\hat{\beta} > 0 = 1$ ), suggesting that  
 309 the connection between /ʀʀ/ and Italian names is extremely robust in this experiment. This result  
 310 supports the idea that in Experiment 1, the participants avoided associating names with a voiced  
 311 obstruent geminate with French restaurant names, hence shrinking the difference between the  
 312 control items and the target items.

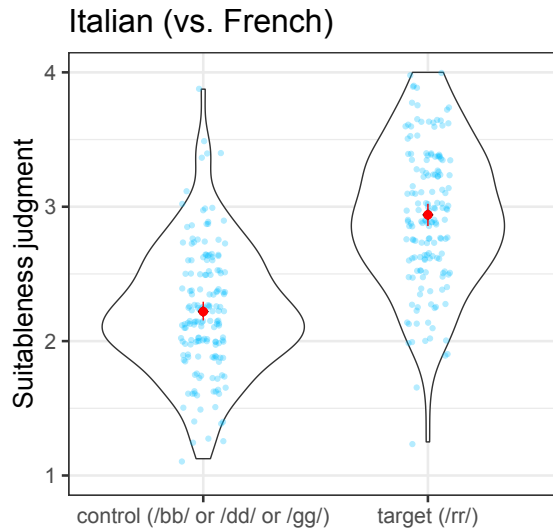


Figure 5: The comparison between the control and the target items in terms of how suitable they were judged as Italian restaurant names.

### 313 3.2.3 German names

314 Figure 6 shows the difference between the control items and the target items in terms of how  
 315 suitable they were judged as German names. Those target items with /hh/ were judged to be  
 316 more suitable than the control items with voiceless stop geminates (2.93 vs. 2.35); the central  
 317 coefficient estimate is 1.62, with its 95% credible interval being [0.62, 2.57]. More than 99% of the  
 318 posterior samples were positive, showing that this association is very robust.

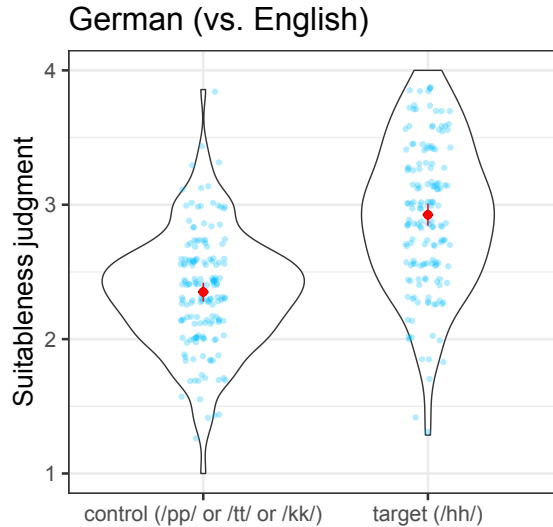


Figure 6: The comparison between the control and the target items in terms of how suitable they were judged as German names.

### 3.3 Discussion

The results were all in the expected direction and were generally clearer in Experiment 2 than in Experiment 1. Overall, the results suggest that Japanese speakers can infer, based a particular phonotactic cue, a very specific subpart of the lexicon, that is something as specific as “snack names”, “Italian names” and “German names”. In this sense, the traditional lexical stratification into native, Sino Japanese, loanwords and mimetics does not suffice to explain the current experimental results.

## 4 General discussion

We started with a general question concerning how the lexicon can be organized into smaller sub-lexicons in a natural language. Phonotactic restrictions do not hold uniformly across the lexicon of a whole language and the lexicon thus seems to be organized into sub-lexicons—then, how fine-grained can this sub-division be? The classic, and conservative, view is that such division should be very general, perhaps with independent etymological/orthographic motivations (Chomsky & Halle 1968; Ito & Mester 1995, 1999, 2008). The alternative view is that the sub-lexicons can be very specific, as specific as “a set of roots that a particular affix is attached to” (Becker & Gouskova 2016; Gouskova et al. 2015; Inkelas et al. 1996; Inkelas & Zoll 2007; Pater 2005).

To address this general question, the paper first pointed out three cases in which particu-



336 lar phonotactic tendencies can cue a very specific word-class in Japanese, two of which were  
337 quantitatively examined by a corpus-study. We then moved on to two experiments, which have  
338 demonstrated that Japanese speakers can infer specific lexical classes such as “snack names” and  
339 “German names” based on phonotactic cues, which is compatible with the second view.

340 The current study opens up several opportunities for future studies. The current study focused  
341 on three specific cases of phonotactic cues that are associated with a specific sub-lexicon, but it  
342 is not hard to imagine that there can be other similar cases in Japanese and other languages.  
343 For example, it has been pointed out that back vowels are overrepresented in ice cream names  
344 in English (Jurafsky 2014), and therefore, it would be interesting to explore whether English  
345 speakers can use this cue to identify an ice cream name. It is also not hard to imagine that among  
346 the loanwords in English, forms with only light syllables may be overrepresented in those that  
347 are borrowed from Japanese, such as *Toshiba*, *Pikachu* and *Yamaha*. It would be interesting to test  
348 whether English speakers would tend to associate new nonce words with only light syllables as  
349 those borrowed specifically from Japanese.

## 350 **Funding**

351 This project is supported by JSPS grants #22K00559 to the first author and #19K13164 to the second  
352 author.

## 353 **Conflicts of interest**

354 We declare no conflicts of interest.

## 355 **Availability of data and material**

356 The data are available at  
357 <https://osf.io/ym79p/?viewonly=ce17de5a39834ae397c44a19e74db082>

## 358 **Code availability (software application or custom code)**

359 The code is also available at  
360 <https://osf.io/ym79p/?viewonly=ce17de5a39834ae397c44a19e74db082>

## 361 **Authors' contributions**

362 Both authors contributed to the conception and execution of the experiments. The first author  
363 wrote the manuscript and the second author revised it. The statistical analysis was primarily  
364 conducted by the first author. The second author checked the details.

## 365 **Ethics approval**

366 The current experiments were conducted with an approval from the authors' institute.

## 367 **Consent to participate**

368 The participants read the written consent form before participating in the experiments.

## 369 **Consent for publication**

370 Both authors approve that the current manuscript be evaluated for publication in the journal.

## 371 **References**

- 372 Alderete, John & Alexei Kochetov. 2017. Integrating sound symbolism with core grammar: The  
373 case of expressive palatalization. *Language* 93. 731–766.
- 374 Anttila, Arto. 2002. Morphologically conditioned phonological alternations. *Natural Language  
375 and Linguistic Theory* 20(1). 1–42.
- 376 Becker, Michael & Maria Gouskova. 2016. Source-oriented generalizations as grammar inference  
377 in Russian vowel deletion. *Linguistic Inquiry* 47(3). 391–425.
- 378 Blake, Allen & Michael Becker. 2015. Learning alternations from surface forms with sublexical  
379 phonology. Ms. [lingbuzz/002503](https://lingbuzz/002503).
- 380 Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models using Stan.  
381 *Journal of Statistical Software* 80(1). 1–28.
- 382 Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.
- 383 De Benedetto, Maria Gabriella & Luca De Nardis. 2021. Consonant gemination in Italian: The  
384 nasal and liquid case. *Speech Communication* 133. 62–80.
- 385 Franke, Michael & Timo B. Roettger. 2019. Bayesian regression modeling (for factorial designs):  
386 A tutorial. Ms. <https://doi.org/10.31234/osf.io/cdxv3>.
- 387 Gelbart, Ben. 2005. *Perception of foreignness*: University of Massachusetts, Amherst Doctoral  
388 dissertation.
- 389 Gelbart, Ben & Shigeto Kawahara. 2007. Lexical cues to foreignness in Japanese. In *Proceedings  
390 of formal approaches to Japanese linguistics 4 [MIT working papers in linguistics 55]*, 49–60.  
391 Cambridge: MITWPL.

- 392 Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau & Yu-Sung Su. 2018. A weakly informative  
393 default prior distribution for logistic and other regression models. *Annual Applied Statistics*  
394 2(4). 1360–1383.
- 395 Gouskova, Maria & Suzy Ahn. 2024. Sublexical phonotactics and English comparatives. *University*  
396 *of Massachusetts Occasional Papers in Linguistics 50th Anniversary Volume* .
- 397 Gouskova, Maria, Sofya Kasyanenko & Luiza Newlin-Ł ukowicz. 2015. Selectional restrictions as  
398 phonotactics over sublexicons. *Lingua* 41–81.
- 399 Inkelas, Sharon, Orhan Orgun & Cheryl Zoll. 1996. Subregularities as cogrammars: the theoretical  
400 status of nonproductive patterns and exceptions in grammars. In *Proceedings of the nels 26*, Ms.,  
401 University of California, Berkeley.
- 402 Inkelas, Sharon, Orhan Orgun & Cheryl Zoll. 1997. The implications of lexical exceptions for  
403 the nature of grammar. In Iggy Roca (ed.), *Derivations and constraints in phonology*, 393–418.  
404 Oxford: Oxford University Press.
- 405 Inkelas, Sharon & Cheryl Zoll. 2007. Is grammar dependence real? A comparison between  
406 cophonological and indexed constraint approaches to morphologically conditioned phonology.  
407 *Linguistics* 45. 133–172.
- 408 Ito, Junko & Armin Mester. 1995. Japanese phonology. In John Goldsmith (ed.), *The handbook of*  
409 *phonological theory*, 817–838. Oxford: Blackwell.
- 410 Ito, Junko & Armin Mester. 1996. Stem and word in Sino-japanese. In T. Otake & A. Cutler (eds.),  
411 *Phonological structure and language processing: Cross-linguistic studies*, 13–44. Berlin: Mouton  
412 de Gruyter.
- 413 Ito, Junko & Armin Mester. 1999. The phonological lexicon. In Natsuko Tsujimura (ed.), *The*  
414 *handbook of Japanese linguistics*, 62–100. Oxford: Blackwell.
- 415 Ito, Junko & Armin Mester. 2008. Lexical classes in phonology. In Shigeru Miyagawa & Mamoru  
416 Saito (eds.), *The Oxford handbook of Japanese linguistics*, 84–106. Oxford: Oxford University  
417 Press.
- 418 Jurafsky, Dan. 2014. *The language of food: A linguist reads the menu*. New York: W. W. Norton &  
419 Company.
- 420 Kawahara, Shigeto. 2006. A faithfulness ranking projected from a perceptibility scale: The case  
421 of [+voice] in Japanese. *Language* 82(3). 536–574.
- 422 Kawahara, Shigeto. 2007. Copying and spreading in phonologica theory: Evidence from echo  
423 epenthesis. In Leah Bateman, Michael O’Keefe, Ehren Reilly & Adam Werle (eds.), *University*  
424 *of Massachusetts occasional papers in linguistics 32: Papers in Optimality Theory iii*, 111–143.  
425 Amherst: GLSA.
- 426 Kawahara, Shigeto. 2019. What’s in a precure name? *ICU Working Papers in Linguistics 7:*  
427 *Festschrift for Professor Tomoyuki Yoshida on his 60th birthday* 15–22.
- 428 Kawahara, Shigeto. 2023. *Why snack names often contain [p] in japanese: A linguist interacts with*  
429 *elementary school students [in Japanese]*. Discover 21.
- 430 Kawahara, Shigeto & Melanie Pangilinan. 2017. Spectral continuity, amplitude changes, and  
431 perception of length contrasts. In Haruo Kubozono (ed.), *Aspects of geminate consonants*, 13–  
432 33. Oxford: Oxford University Press.
- 433 Kerr, N.L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Psychology*  
434 *Review* 2(3). 196–217.
- 435 Kruschke, John K. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*.  
436 Waltham: Academic Press.

- 437 Kruschke, John K. & Torrin M. Liddell. 2018. The Bayesian new statistics: Hypothesis testing, esti-  
438 mation, meta-analysis, and power analysis from a Bayesian perspective. *Psychological Bulletin*  
439 *and Review* 25. 178–206.
- 440 Kumagai, Gakuji. 2019. A sound-symbolic alternation to express cuteness and the orthographic  
441 Lyman’s Law in Japanese. *Journal of Japanese Linguistics* 35(1). 39–74.
- 442 Kuroda, S.-Y. 1965. *Generative grammatical studies in the Japanese language*: MIT Doctoral dis-  
443 sertation.
- 444 Labrune, Laurence. 2014. The phonology of Japanese /r/: A panchronic account. *Journal of East*  
445 *Asian Linguistics* 23(1). 1–25.
- 446 Lemoine, N.P. 2019. Moving beyond noninformative priors: Why and how to choose weakly  
447 informative priors in bayesian analyses. *Oikos* 128. 912–928.
- 448 McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan,*  
449 *2nd edition*. London: Taylor & Francis Ltd.
- 450 Moreton, Elliot & Shigeaki Amano. 1999. Phonotactics in the perception of Japanese vowel length:  
451 Evidence for long distance dependencies. *Proceedings of the 6th European Conference on Speech*  
452 *Communication and Technology* .
- 453 Morimoto, Maho. 2020. *Geminated liquids in Japanese: A production study*: University of Califor-  
454 nia, Santa Cruz Doctoral dissertation.
- 455 Morita, Takashi. 2018. *Unsupervised learning of lexical subclasses from phonotactics*: MIT Doctoral  
456 dissertation.
- 457 Orgun, C. Orhan. 1996. *Sign-based morphology and phonology: With special attention to Optimality*  
458 *Theory*: University of California, Berkeley Doctoral dissertation.
- 459 Pater, Joe. 2005. Learning a stratified grammar. In Alejna Brugos, Manuella R. Clark-Cotton & Se-  
460 ungwan Ha (eds.), *Proceedings of the 29th boston university conference on language development,*  
461 482–492. Somerville: Cascadilla.
- 462 R Development Core Team. 1993–. *R: A language and environment for statistical computing*. R  
463 Foundation for Statistical Computing Vienna, Austria.
- 464 Sande, Hannah. 2020. Cophonologies by Ph(r)ase. *Natural Language and Linguistic Theory* 38.  
465 1211–1261.
- 466 Shaw, Jason. 2006. Learning a stratified lexicon. *Proceedings of NELS* 36. 519–530.
- 467 Vance, Timothy. 2008. *The sounds of Japanese*. Cambridge: Cambridge University Press.
- 468 Vasishth, Shravan, Bruno Nicenboim, Mary Beckman, Fangfang Li & Eun Jong Kong. 2018.  
469 Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonet-*  
470 *ics* 71. 147–161.
- 471 Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.