Learnability of length-referencing alternations

Hailang Jiang*
Supervised by James White

Abstract: The present study takes a starting point from the non-internalized pattern of Voiced Velar Nasalization (VVN) in compounds in Yamanote Japanese which is partly conditioned by the mora length of the compound in the corpus and then investigates the learnability of phonological generalizations conditioned by length. Through an artificial language learning experiment targeting a structure akin to Japanese VVN, this study reveals that length-referencing phonological alternations were always underlearned, no matter what exact length was involved as the condition in the context of a rule. The results of this experiment favor the idea that phonology lacks access to length during the learning process and there is a bias against it, which may be one reason for the typological rarity of such patterns.

Keywords: learnability, artificial language, experiment, alternation, nasalization

1 Introduction

1.1 Japanese voiced velar nasalization

In Yamanote Japanese (a traditional dialect of Standard Japanese spoken in an area in Tokyo), the phoneme /ɡ/ has two allophones depending on the context: [ɡ] and [ŋ]. The voiced velar oral [ɡ] and the velar nasal [ŋ] surface word-initially and word-medially, respectively. The process of /ɡ/ becoming [ŋ] is conventionally known as Velar Voiced Nasalization (VVN) in the literature (Ito & Mester 1996), which is formalized as the rule in (1).

(1) Voiced Velar Nasalization in Yamanote Japanese
/ɡ/→ [ɡ] / [$_{\text{word}}$ __
/ɡ/→ [ŋ] / [$_{\text{word}}$ X_0__

When we concatenate two members[1] together to form a new compound word, if the second member begins with /ɡ/, according to the rule in (1), it should surface as a /ŋ/. However, the data shows that there are times when /ɡ/ can only surface as [ɡ], and other times when both [ɡ] and [ŋ] are acceptable (Ito & Mester 1996). In other words, in cases where /ɡ/ serves as the initial phoneme in the second member of a compound, it either optionally[2] undergoes VVN surfacing as either [g] or [ŋ], or remains unchanged as [ɡ]. This is exemplified by the examples in (2), which illustrate different words that either exhibit optional VVN or never undergo VVN.

(2) Japanese two-member compound words (from Breiss et al. (2022))
i. that optionally undergo VVN
/doku + ga/ ➞ [doku-ga]~[doku-ŋa] "poison moth"

---

* Thanks to James White and Department of Linguistics at UCL for jointly funding this study. Thanks to James White for pronouncing the stimuli used in the experiment of this study.

[1] Each member is either a morpheme (e.g., bou-gai ぼう-がい/妨-害) or a few morphemes (himitsu-gaikou ひみつ-がいこう/秘密-外交).

[2] Note that in the context of Japanese VVN, the term 'optional' implies that for certain specific words such as the example in (2.i), the voiced velar /ɡ/ has two options of free variants to choose from, namely [ɡ] and [ŋ].

ii. that never undergo VVN
/noo + ɡeka/ ➞ [noo-ɡeka] *[noo-ŋeka] "brain surgery"

Whether the rule optionally or never applies might seem specific to each single word. Breiss et al. (2022)'s corpus study quantitively summarizes this morphophonological observation and proposes several factors that can condition the optional application of VVN in the data.

The first of them, which is the central concern of this essay, is the mora[3] length of the entire compound word. According to their findings, the mora length of a word has a significant effect on the word's probability of undergoing VVN. The following Figure 1[4], adapted from Figure 3 in Breiss et al. (2022), depicts the proportion of words undergoing VVN across two-member compound words of different mora lengths.

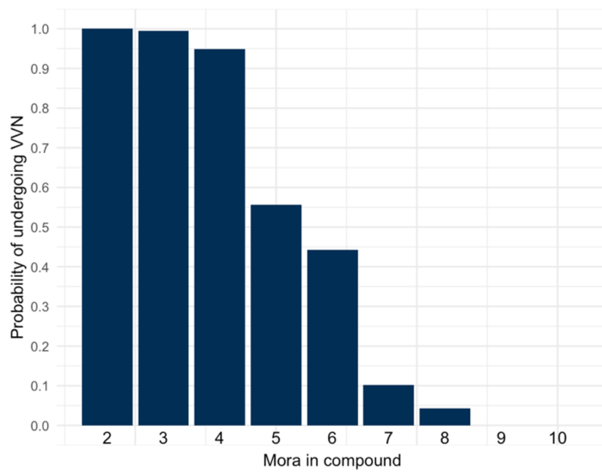

Figure 1. Probability of undergoing VVN for words of different mora lengths
(Excluding words with free members)

The figure above shows that at lengths 2 through 4, almost all words have the option to undergo VVN; at lengths 5 through 6, nearly half of the words optionally undergo VVN; at lengths 7 and above, almost all words never undergo VVN. There are two dramatic drops between 4 and 5, and 6 and 7. This is a typologically strange phenomenon, because speakers should at least have information about the mora length of a given word if they are really aware[5] of the generalization here and thus able to apply VVN based on the mora length. Before the specifics are explained in Subsection 1.2, I will briefly discuss the two other factors reported by Breiss et al. (2022) that affect words' probability of undergoing VVN.

In addition to mora length, Breiss et al. (2022) also report that relative M1 or M2 frequency (throughout this essay, M stands for a member in the compound word) and the nasality of the

[3] In Japanese, a mora-timing language, each mora takes almost identical amount of time (Warner & Arai 2001). Japanese linguistics uses the concept of moras more often than that of syllables.

[4] Both Figure 1 in this essay and Breiss et al. (2022)'s Figure 3 depict the frequency of compounds that undergo VVN under different mora lengths, i.e., the average probability of undergoing VVN. The former only includes compound words whose second member is free (i.e., it can also be used as an independent word), while Breiss et al. (2022)'s Figure 3 includes all words. The trends reflected in the two datasets look similar. The penultimate paragraph in this section explains why I would use this dataset with a smaller range.

[5] See Subsection 1.2 for more details.

preceding segment (being a nasal vs. a vowel) significantly affect the probability of undergoing VVN.

Relative frequency (cf. Hay 2003 and Hay and Baayen 2005, which demonstrate the utility of this measure) of members is defined as the log-frequency of a member (M1 or M2) minus the log-frequency of the compound per se. This variable reflects the tendency to maintain identity between different types of members (free vs. bound) and the compound they are in. It is explained by Breiss et al. (2021) as a means to enforce paradigm uniformity (output-to-output identity). When the M2 is bound and has zero frequency, it has a relative frequency of negative infinity, and according to the corpus data from Breiss et al. (2022), all the compounds with such a bound M2 never undergo VVN and thus exhibit no variations regarding the relative frequency of M2 so they are irrelevant to the factor of mora length investigated here and thus excluded in view of the considerations in this essay. This is why I chose not to present the data with all words in Figure 1 in this essay and only focused on the compound words with free members.

The nasality of the preceding segment is a natural predictor of the VVN alternation, representing local harmony/assimilation in nasality.

1.2 Naturalness of the three factors and the non-internalized regularity about mora length

Unlike the two factors, relative frequency and preceding nasality, the mora length of the compound, briefly discussed previously, is neither a natural nor a typologically common determinant of the allophone in the alternation. I will scrutinize all three factors in the following two paragraphs.

Relative frequency can be a part of lexical characteristics that can explain the variability of a process involving reference to surface forms of the paradigm of a word (see Steriade & Stanton 2020, Breiss 2021a, b) and this is a valid phonological ground to account for the relative-frequency-based VVN variations.

I use the term 'natural' for the other two factors in the sense of Peperkamp, Skoruppa & Dupoux (2006): a rule is natural when it satisfies the standard of (i.) Phonetic Proximity, (ii.) Contextual Relevance and (iii.) Markedness Reduction. The nasality preceding segment is a natural factor since it satisfies all three standards; the mora length is not a natural one as it greatly violates (ii.) Contextual Relevance which means 'the set of triggers/context is relevant to the change in that it tends to be homogeneous with respect to the feature(s) that undergo(es) the change' because the target changed from [ɡ] to [ŋ] becomes neither more similar nor dissimilar to its context in shorter words.

Naturalness of a phonological pattern can enhance its learnability, as shown by many experiments (e.g., Pater and Tessier 2003, Wilson 2003, Peperkamp, Skoruppa & Dupoux 2006, Hayes et al. 2009). In the case of unnatural phonological generalizations, very often there is a learning bias against them, despite evidence to the contrary. For a review on the learnability on unnatural patterns, see Hayes & White (2013).

Additionally, from another perspective, the effect of mora length on whether to nasalize a stop is unnatural because it involves second-order phonotactics (Walker & Dell 2006, Becker et al. 2011), in that the alternation is dependent on some aspect of the syllable or the word (here, the number of moras in the entire compound word) other than the position it fills. There is evidence suggesting that second-order phonotactics, if not unlearnable at all, are acquired at a slower pace and with less robustness in certain instances (for a review, see Walker & Dell 2006).

Progressive assimilation spotted in the pattern of Japanese VVN is not typologically rare, but mora length affecting nasality is. I will discuss this in more detail in subsection 1.3. In short, very few languages exhibit a pattern that involves counting the phonological units in a domain to more than 2 as a condition of the rule under any analysis, especially when it conditions segmental features like [nasal].

In fact, a previous study of mine (Jiang 2023) shows by conducting a wug test which controls for the relative frequency of M1 and M2 that the strange pattern of mora length is not internalized by native speakers at all, while the regularity of preceding nasality is highly internalized. During the experiment of this study, I created a series of Japanese nonce compound words by manipulating the factors of mora length and preceding nasality. I firstly showed participants the two members of the compound words as free words used in sentences and asked them to compound each pair. Finally I had them rate the relative acceptability or ‘naturalness’ of two potential options of the compound words: the original version and nasalized version (e.g., temi-gemo vs. temi-ŋemo), to find that in terms of the mora length, almost no pattern is internalized by native speakers, as the probability of undergoing VVN in nonce words is always close to 50%, the chance level (Figure 2.2), and this is totally different from the regularity in the real lexicon (Figure 2.1, copied from Figure 1) while in terms of preceding nasality, one thing is learned for nonce words: nasals give rise to slightly more nasalizations than vowels (Figure 2.4), and this is consistent with the statistics in the real lexicon (Figure 2.3). Figures 2.1 through 2.4 are taken from Jiang (2023).

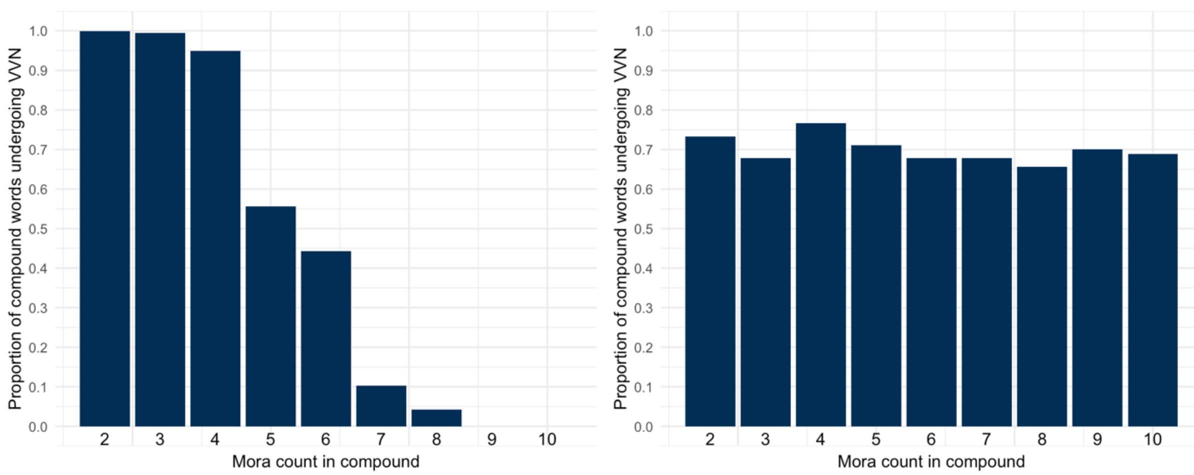

Figure 2.1 & 2.2. The statistics in the real lexicon (left) and the wug test (right)
(Predictor: mora length)

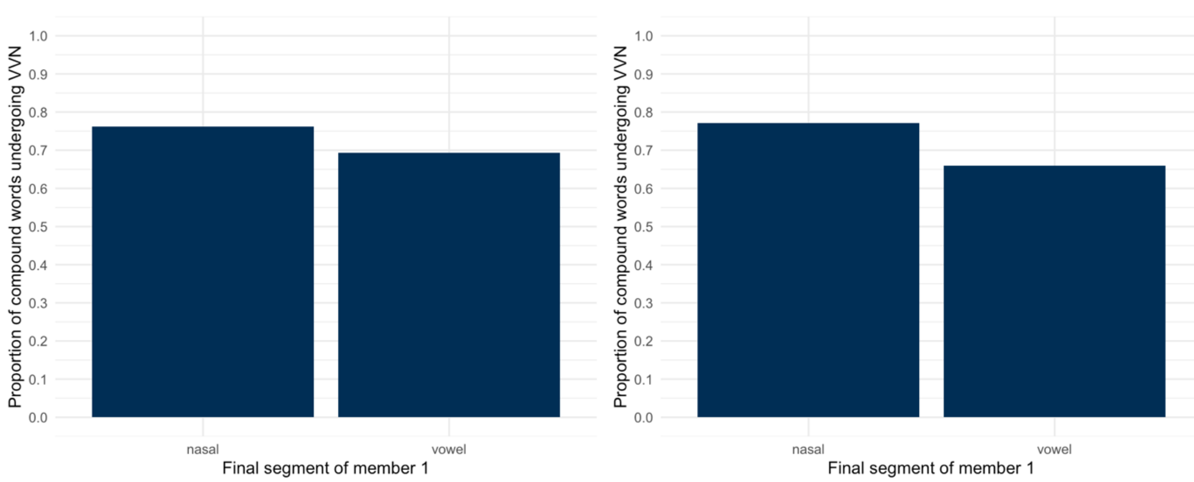

Figure 2.3 & 2.4. The statistics in real lexicon (left) and wug test (right)
(Predictor: preceding nasality)

If we assume that there is no factor other than the three above that can account for the unusual non-internalized pattern (although one likely factor is the frequency of words in that length)[7], then we can propose that the phenomenon of alternation which involves the information on word length (as shown in Figure 1) is an accident, and this is another example of the 'surfeit of the stimulus' effect which describes a phenomenon where the input to a particular linguistic structure may be not fully internalized as a part of the language knowledge learned during acquisition, as in Becker et al. (2011, 2012).

1.3 Length-referencing, counting and their typological rarity

How can the grammar have the information on word length? That is to say, how does phonology get access to the exact number of phonological units (here, moras)?

If we assume that the number of moras is not a part of the lexical information and that the mora length is the factor that is relevant and indeed playing a role in determining a word's probability of undergoing VVN, one possibility is that we can only access it by counting. Counting has been an unusual thing in phonology. There are many phonologists claiming in various ways the idea that phonological generalizations (rules or constraints) cannot 'count' to 3 or more (see Paster 2019 for a full list). For example, McCarthy & Prince (1999) argues that considering of locality, a count in grammar can run up to two: a rule may fix on one specified element and examine a structurally adjacent element and no other; Rose (1999) argues that the structure of binary branching is not a sort of counting in that 'the number 2 is an artefact of the hypothesis that languages select binary versus unary constituents … if arithmetic counting were allowed in representations, we would also expect ternary or quaternary constituents, a possibility which is arguably not required.' There are some languages in which phonological generalizations are analyzed to be counting (to 3 or more), but under re-analyses, they can be perfectly explained by binary structures (i.e., counting within 2) (see Paster 2019 for a review). Paster (2019) also proposes an idea that some phenomena, such as H tone assignment in Kuria, cannot be analyzed with any approach where counting is limited within 2. Whether or not currently we are able to formalize the counting pattern in UG at all, it is certain that the pattern of counting to 3 or more is typologically rare (Paster 2019). Moreover, the reported counting phenomena all focus on prosody (i.e., stress or tone), as an example of a counting phenomenon involving segmental features is missing (Paster 2019), thus it is very unusual that counting can regulate nasality in Japanese VVN.

Whilst counting has different senses in different literatures, there are not many phonological generalizations with regard to the so-called 'counting' to the quantity of 3 and above, let alone any phonological alternation about length which at least 'counts' to 7, as Japanese seems to do. To the best of my knowledge, no one reports any length-referencing alternation which counts to at least 7 in any other language. A length-referencing alternation is a phonological alternation where the application of a rule is conditioned by the length of a particular phonological units with in a domain (e.g., the number of syllables in a word). The typological gap of it, together with the non-internalized pattern about mora length in Japanese, leads me to ask whether there is an learning bias against referencing the length information (via counting)? To what extent can such a learning

[7] I will discuss the assumption and this factor in detail in Section 4, as this is one of the neglected factors that are not covered in Breiss (2022).

bias account for the typological rarity and thus possibly accounting for the non-internalization of length-conditioned VVN in Japanese?

1.4 Overview of the experiment

To summarize the findings of previous studies, counting to 2 is considered normal; counting to 3 or 4 is unacceptable from the perspective of many phonologists and is rare across-linguistically (but still attested); counting to 5 or more is never reported, with Japanese VVN potentially being an exception, but still it is not internalized.

Based on the above questions, the hypothesis to be tested in this essay is whether an alternation can be learned if it is conditioned by a length in phonological units in a domain, and if so, whether there is any difference in the learning results if different numbers matter as a conditioning length. I designed and conducted an artificial language learning experiment. See Section 2 for the advantage of such a paradigm.

As a preview, in the experiment of this study, I had the participants learn a length-referencing alternation in a novel language. They were to select a correct form of a compound given two single free morphemes.

There was a length boundary (a specific number of syllables in a word), and only below the boundary the length-referencing rule were able to apply. The length boundaries varied across groups because I wanted to investigate at which lengths the potential learning bias begins to exert influence. I set up seven groups - Groups I through VII - each of them were to learn a slightly different artificial language, with a unique length boundary (which is equal to the group number), and this rule applied only when the length (here, in syllables) of a stimulus word fell below this length boundary. In short, participants in Group *n* were to learn phonological alternations where the rule applied only when the length of the word was below the length boundary *n* (that is, in this case, when there were 1 through *n* syllables); the rule did not apply when the length of the word was above the boundary (that is, when the number of syllables was greater than *n*). The 70 participants were randomly assigned to one of the seven groups.

For example, Group IV learned the phonological alternation where a rule applied when the number of syllables was 1 to 4 and did not apply when the number of syllables was 5 to 8. This alternation is illustrated by the Figure 3. The specific alternating patterns to be learned by each group are placed in Appendix 1.

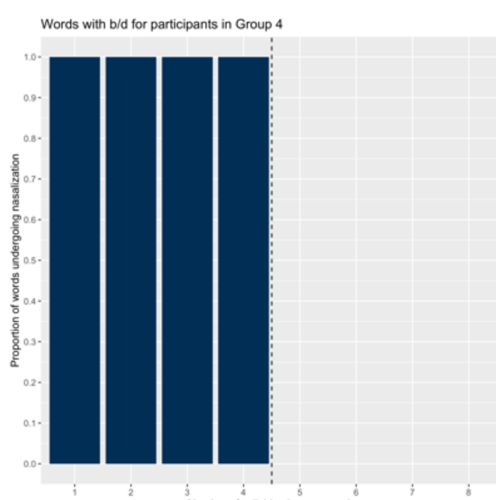

Figure 3. The alternating pattern to be learned by participants in Group IV

The above design was intended to roughly simulate the length-conditioned trend in the Japanese corpus (see Figure 4). The similarity between the artificial languages designed here and Japanese is that the rule does not apply when the word is short in length and applies when the word is long enough in length. The simplification I made is that the stimuli I designed follow a two-value distribution while the Japanese data follows a sigmoid-like distribution, and that the nasalization is a must-apply rule when the condition is satisfied, rather than an optional rule.

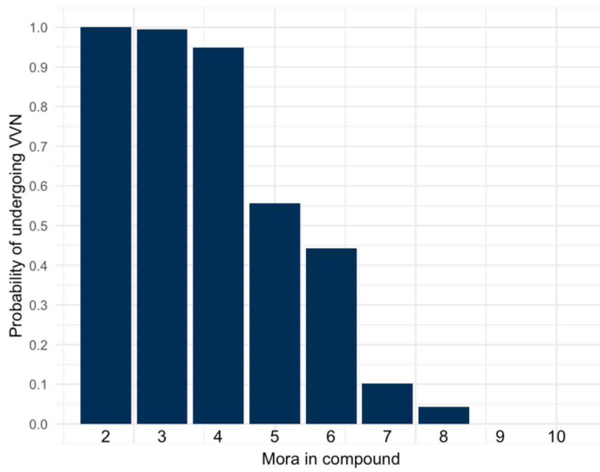

Figure 4. The trend about mora length in Japanese (Copied from Figure 1)

As a preview, the rule to be learned was a compounding rule, just like VVN in Japanese. When two members, in this case, two free morphemes (also free words) concatenated to form a compound, the voiced oral stops in the initial positions of the second morpheme might undergo nasalization (depending on the length of the word and the length boundary in each group), but instead of /g/ to [ŋ] in Japanese, it was /b/ to [m] and /d/ to [n] in the artificial language. For example, a quadrisyllabic compound /vupa-diwu/ was realized as [vupa-niwu] for participants in group V through VII as the word length 4 fell below their groups' boundaries, but remained unchanged as [vupa-diwu] for participants in group I through IIV as 4 is not greater than 1, 2, 3 or 4. Detailed materials will be given in Subsection 2.2.

The rest of this essay is organized as follows: Section 2 is an overview of the experiment's design; Section 3 analyzes the results from the experiment; Section 4 explores related topics and Section 5 concludes.

2 Method

My experiment adopts the paradigm of artificial language learning (ALL), which is widely used in studying the learnability of phonological generalizations as in Finley (2017), White (2014), Wilson (2006), among many others. This paradigm has a design specific to the research question that allows for the manipulation of structures to which learners are exposed in the lab, and by comparing the learning results of different structures, researchers are able to assess how participants generalize and acquire knowledge about the language and to what degree their learning results are biased.

2.1 Participants

I recruited 70 native English monolingual speakers aged between 20 and 66 averaged 39.04, with 30 of them being male and 39 of them being female (1 person refused to report gender), without mastery of any language other than English, from Prolific. To further control for language background, they had to hail from the United States and not have spent more than 6 months abroad.

Participants were randomly placed into seven groups, each consisting of ten individuals. The basis for the groupings is described in Subsection 2.2.

To ensure data quality, I included an attention check among the testing trials. Participants who failed the attention check were excluded from the data (n=21). Finally, 49 of them passed. The breakdown of eligible participants in each group are shown in the table below.

Table 1: The numbers of eligible participants in each group

group	number of subjects
I: 1-syll	8
II: 2-syll	8
III: 3-syll	5
IV: 4-syll	9
V: 5-syll	7
VI: 6-syll	6
VII: 7-syll	6

2.2 Materials

All stimuli of a compound word consisted of two free morphemes. For simplification, the morpheme boundary for all compound words had to be in the middle of it[8].

There were compound words of 8 different lengths 1[9] through 8, and each length had 20 stimuli. Ten of them had a b/d-initial second morpheme; and the other ten had a non-b/d-initial second morpheme. Thus, there were 8*20 = 160 compound words in total.

For stimuli of the length 2 through 8, all syllables were always shaped as CV. The first syllables of b/d-initial second morphemes, which were DV where D represented /b/ or /d/. D might undergo nasalization, becoming a nasal with the same place of articulation /m/ or /n/. All other CV syllables never started with D. V represented one of the three most common vowel phonemes, /a, i, u/, utilized in the test and all of them were short. C was a consonant from the English consonant inventory /p, f, v, t, s, z, l, tʃ, dʒ, ʃ, ʒ, j, k, g, w, h/. There were 16 of them in total. I did not include /θ, ð, r/ because they made the language English-like; /ŋ/ cannot be syllable-initial in English so was excluded. /b, d, m, n/ were also excluded for C because they should be in the D positions. For example, the template for quadrisyllabic compounds were CVCV-CVCV for words with a non-b/d-initial second morpheme (there were 10 of such words) and CVCV-DVCV for the remaining 10 words of that length.

[8] For example, for a quadrisyllabic word σσσσ, the division between two morphemes is always σσ-σσ; for a trisyllabic word σσσ, the morpheme division is always σ-σσ (for odd-syllable compounds, the first morpheme is always one syllable more than the second).

[9] Note that the length 2 through 8 represents lengths (the number of syllables), while the ‘length’ 1 represents a certain structure CVD-V or CVC-V so it is not length that is relevant for stimuli labelled as ‘1’. Strictly speaking, CVD-V or CVC-V is a disyllabic word, but it has the D at issue (/d/ or /b/) in the first syllable. The number ‘1’ here is just a label for ‘the first syllable’. Through this approach, it is possible to investigate its impact of a phonological rule on learnability when it functions in the first syllable. In a word CVD-V, the D which is to undergo nasalization is not in the initial position of the second morpheme, unlike words of other lengths. The meanings of C, V and D will be elucidated in the following paragraph.

For those words containing a D (/d/ or /b/), a potential alternation of nasalization occurred when a compound was formed by concatenating two free morphemes where a D (/d/ or /b/) became an N (/n/ or /m/), as stated in (3). Whether this change applied or not depended on the word length and the condition of the group where length boundary differed.

(3) The nasalization rule of compounding

$[b, d] \rightarrow [m, n] / [_{\omega}\ \mu\ [_{\mu}\ \_\_\ X_0\ ]_{\mu}\ ]_{\omega}$

only when $|\omega_{\mu\text{-}\mu}| \leqslant |\text{boundary}|$

Where μ denotes a free morpheme and ω a compound word; $\omega_{\mu\text{-}\mu}$ denotes a bimorphemic compound; $|\omega_{\mu\text{-}\mu}|$ denotes the number of syllables in the compound and |boundary| is the cut-off point for whether or not the rule applies, as described in Subsection 2.2, which is also the number of the group.

In plain English, for participants in Group I, this alternation only occurred to monosyllabic compound words with a D in them; for participants in Group X, this alternation occurred to '1', 2, 3 … through X-syllabic compound words with a D.

Four recordings for each bimorphemic compound were recorded: the first free morpheme, the second free morpheme, the non-nasalized compound and the nasalized compound with an N (/n/ or /m/). Note that a nasalized version for all compounds was recorded, no matter whether it contained a D (/d/ or /b/) or not. For those words with a D, the correct answer was predicted by the rule in (3); for those words without a D, the correct answer was always the non-nasalized version. I included words without D as stimuli which was not nasalized at all to prevent participants from learning a rule like 'short compounds contain nasals.' /d/ and /b/ changed into /n/ and /m/, respectively; whether other consonants changed into /m/[+labial] or /n/[-labial] in the initial syllable of the second morpheme in the nasalized recording depended on the labial feature of them. For example, if it was CVCV-lVCV, the incorrect nasalized answer contained /n/ as /l/ is [-labial]; if a compound word was CVCV-wVCV, the incorrect nasalized answer contained /m/ because /w/ is [+labial]. Eight of the C consonants corresponded to /m/ and the other eight corresponded to /n/. The correspondence is shown in Table 2.

All the of stimuli had an initial stress. The stress of a compound was located as the first syllable as well, which acted as a rule applying by default when a compound was formed. The stimuli were pronounced by a native American English speaker.

Table 2. The correspondence between C and its nasalized version N in the incorrect answers

C	N	C	N
p	m	t	n
f	m	s	n
v	m	z	n
tʃ[10]	m	l	n
dʒ	m	j	n
ʃ	m	k	n
ʒ	m	g	n
w	m	h	n

[10] The four segments [tʃ, dʒ, ʃ, ʒ] in English are heavily labialized as a secondary articulation.

Each free morpheme (also free words) had a unique meaning that was fixed across groups and across trials. When two free morphemes M1 and M2 were used to form a compound word M1-M2, the meaning of the compound was the combination of the two morphemes, i.e., an M2 with the characteristics of M1. e.g., if M1 meant ‘pineapple’, and M2 meant ‘clock’, then M1-M2 meant ‘pineapple-shaped clock’; if M1 meant ‘pear’, and M2 meant ‘lighter’, then M1-M2 meant ‘pear-patterned lighter’. The meanings of each free morpheme and formed compound word were shown by pictures. All pictures were automatically generated by an AI called Midjourney based on descriptions. For example, the pictures of the above examples are given in Figure 5.1 through 5.6.

Figure 5.1, 5.2, 5.3. A pineapple, a clock, and a pineapple-shaped clock

Figure 5.4, 5.5, 5.6. A pear, a lighter, and a pear-patterned lighter

At the beginning there was a sentence for participants to test the speaker of the device, saying ‘I speak English’. This was synthesized by Apple Voices.

2.3 Procedure

The experiment was run on Gorilla, where participants were expected to spend around 30 minutes to finish it.

At the beginning of the experiment, participants were asked to read an overview of the experiment where they were told that the purpose of it was to find out how people behave when they learn a new language. By clicking a consent button, they agreed and proceeded to participate. After that, participants were taken to a page to test their device, where they listened to an English sentence and were asked to enter this sentence into a textbox while adjusting the volume of the device to a comfortable level. This was to make sure the audio worked in good quality. Afterwards, they were told in detail about the procedure of this experiment and what was required of them.

They were all told on the page of procedure instruction that there would be an attention check asking what day of the week it was but they were supposed to type ‘yes’ instead of the day of the week, to make sure they had read the instructions carefully.

Then they were randomly assigned to one of the seven groups without knowing it. They encountered different artificial languages but were unaware that there existed any difference among the languages to be learned by themselves and other participants.

The main body of the experiment consisted of two phases, a training phase and a testing phase. The training phase commenced initially, with participants being directed to an exemplification page which elucidated the procedural steps with a specific illustrative example. Half of the stimuli were used in this phase, that is, 80 compounds in total. Each page corresponded to a unique compound, so each participant encountered 80 pages in total. Participants in 8 groups encountered the same 80 stimuli (although words below different length boundaries were nasalized in different groups).

The stimuli were presented in a self-paced forced-choice task. On each page, there were a picture of M1 on the upper left corner with its audio automatically played once; after a short while, the picture of M2 showed up on the upper right corner with its audio automatically played once. Then a picture of the compound M1-M2 was demonstrated, with a question besides it saying ‘How do you call this item?’. Then there was a pause for participants to manually play the two audios above for any times they wanted until they were ready to proceed and clicked the button saying ‘Show me the options.’ Then a button ‘Option 1’ was presented with the audio of one possible answer automatically played once; after it was finished, it disappeared and the button ‘Option 2’ and an audio of another possible answer were presented in the same way. The two options were a non-nasalized normal concatenation of M1-M2 and its nasalized version, but the order of them were randomized. Afterwards, the two option buttons appeared again with two audio buttons besides them for participants to click for unlimited times in case that they needed to listen to the two options again.

After they made a choice, feedback would be provided. If they chose the correct answer for their particular group, a ‘Correct’ sign would be shown and they would be automatically directed to the page of the next trial. If they hit the incorrect answer, a sign would be shown saying that ‘Incorrect! Try the other option’. Only by clicking the correct option manually were they able to proceed.

Note that during the entire experiment, no orthography of the stimuli was provided, to prevent the participants from using the non-linguistic strategy of ‘explicit counting’ consciously.

The page of attention check might appear between any two trials at this stage.

After participants finished all the training trials, it was the testing phase and they were presented with a short instruction saying there would no longer be any feedback and required to simply apply what they had learned before to infer the correct name of the third picture on each page. Then there came the 80 testing trials, all things being identical except there was no feedback and they could simply click any option to proceed to the next page.

Finally, a survey was made on the language background of participants and if they found any specific pattern in the test. After this step, a completion code will be provided, which is essential for me to determine the data’s usability.

3 Results

The main purpose of this essay is to explore whether learners can learn any length-referencing alternations at any length at all, and whether there is a difference at different lengths.

First, I removed all the data of participants who failed the attention check (n=21). I only used data collected from the testing phase, excluding training data.

3.1 Focus words

I call the words with b/d focus words because these words might nasalize, depending on their lengths. The remaining words served as filler words in the experiment, which never nasalized.

Figures 6.1 through 6.7 shows the percentage of nasalized responses for focus words, according to *group* (I through VII) and the compound length which reflects the boundary *side* (below or above). Recall that the number of a group is also the length boundary of whether the compounding nasalization rule applied in that group. In each figure, the length boundary is represented by a dashed line, only below which the compounding nasalization rule applied.

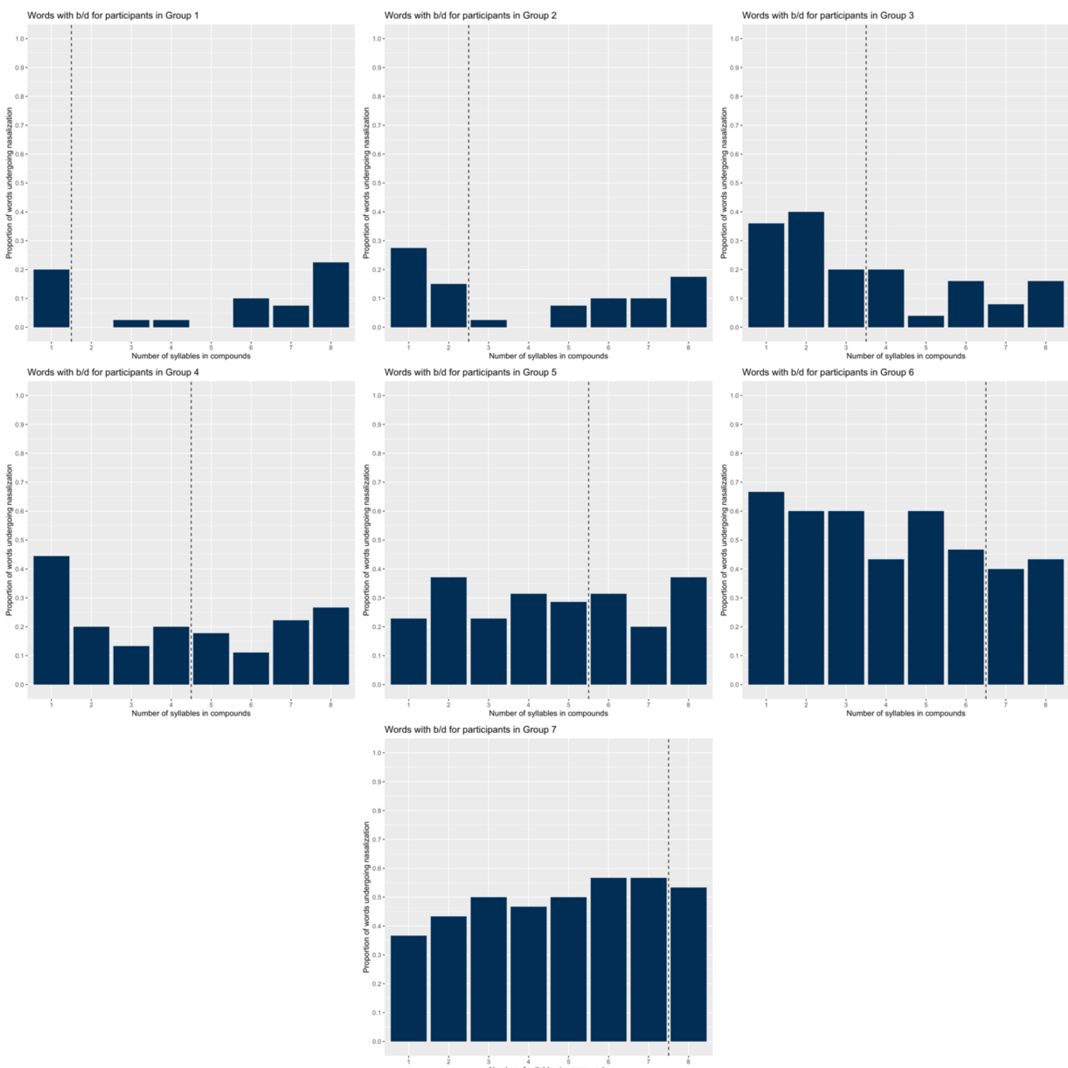

Figure 6.1 through 6.7. Proportions of focus words with /b/ or /d/ undergoing nasalization for participants in each of the 7 groups

The difference in percentage between the two sides of the dashed line in each figure is not considerable, but the differences are great across different figures representing different groups.

I analyzed the data with a mixed-effect multi-variable logistic regression model with maximal random effects (following Barr et al. 2013), implemented in R. The data set is only limited to those focus words, as others should not undergo any nasalization during the process of compounding. The dependent variable is whether each response underwent nasalization or not.

To begin with, in the initial model, the fixed-effect predictors are the *group* which a participant is in (a categorical variable), the condition of boundary *side* (below or above, = 1 when below) and their interaction. I allowed maximal random effects, i.e., by-item random slopes for all variables and a by-item intercept, as well as a by-subject random slope for *side* only (because for each participant, their *group* is fixed) and a by-subject random intercept. The summary of fixed effects for the initial model is presented in Table 3, with the R code in its footnote.

Table 3. Summary of the fixed-effect variables for the initial model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.5377	0.6208	-5.699	0.000000 ***
group2	0.3328	0.7790	0.427	0.669176
group3	1.1776	0.8474	1.390	0.164632
group4	1.6548	0.7629	2.169	0.030089 *
group5	2.5007	0.7764	3.221	0.001279 **
group6	3.2841	0.8610	3.814	0.000137 ***
group7	3.9188	1.0589	3.701	0.000215 ***
side	0.9083	1.4353	0.633	0.526851
group2:side	-0.0331	1.7138	-0.019	0.984591
group3:side	-0.2837	1.9067	-0.149	0.881705
group4:side	-1.1087	1.6878	-0.657	0.511261
group5:side	-1.2048	1.7174	-0.701	0.482994
group6:side	-0.3524	1.7811	-0.198	0.843157
group7:side	-1.4327	1.8803	-0.762	0.446090

R code: glmer(Nasalized_Response ~ group * side + (1 + side | subject) + (1 + group * side | item), data = focus_data, family = binomial)

I compared this initial model with some simpler ones, using backward stepwise comparison by way of a likelihood ratio test implemented in R with *anova(model_1, model_2, test="Chisq")*. The criterion was that a fixed-effect variable should be excluded unless it could improve the model, that is, a significant p-value for the log-likelihood statistics, a smaller AIC and a smaller BIC. Due to the large p-value (Pr(>|z|)) of the interaction term, I excluded the interaction term in the first step, resulting in Model 2. Compared to the initial full model, Model 2 decreased the AIC (from 1929.303 to 1786.148) and BIC (from 2610.149 to 2048.441), and Pr(>Chisq) $\approx$ 1, indicating we could exclude the interaction term to get a better model.

Then I attempted to exclude the remaining factors. I removed the factor *side* first to get Model 3. The exclusion had both AIC and BIC increase (AIC: from 1786.148 to 1885.948; BIC: from 2048.441 to 2086.853) and caused a significant change to the model (Pr(>Chisq) = 0.000). Therefore, the factor of *side* should be remained.

Similarly, I excluded the factor of *group* from Model 2, to get Model 4. Model 2 and 4 have an AIC of 1786.148 and 1760.755, a BIC of 2048.441 and 1805.401, respectively. Based on the likelihood ratio test, the p-value Pr(>Chisq) = 0.071. This value is greater than the typical significance level of 0.05, but considering that we only compared each group with one group and the sample size was small, the p-value, which was less than 0.1 and very close to 0.5, could also be considered significant. The AIC of Model 4 (1760.755) were just slightly lower than that of Model 2 (1786.148); such differences were not substantial. In addition, the effects when *group* = 4, 5, 6 or 7 were significant (compared to *group* = 2). Thus, there is reason to retain the *group* variable.

Therefore, the final model was Model 2, with both *group* and *side* included. The fixed effects of Model 2 are reported in Table 4 below.

Table 4. Summary of the fixed-effect variables for Model 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.4908	0.5474	-6.378	0.00000 ***
group2	0.4042	0.6664	0.606	0.54419
group3	1.1818	0.7569	1.561	0.11845
group4	1.4204	0.6549	2.169	0.03008 *
group5	2.3667	0.6915	3.423	0.00062 ***
group6	3.2948	0.7259	4.539	0.00000 ***
group7	3.2008	0.8322	3.846	0.00012 ***
side	0.2676	0.4457	0.600	0.54821

R code: glmer(Nasalized_Response ~ group + side + (1 + side | subject) + (1 + group + side | item), data = focus_data, family = binomial)

Overall, there was a significant effect of *group* when it took values greater than 3. For *group* = 2 or 3, the differences in mean values did exist (when group = 1, 2, 3, the proportion of nasalized responses was 8.13% ,11.25%, 20.00%, respectively) but were not significant due to the small sample size. The effect of *side* (below or above) was not significant (Pr(>|z|)=0.54821), although including it significantly improved the model. I interpreted the results as indicating that participants were unable to differentiate the context of below-boundary length and above-boundary length thus failed to learn the length-referencing alternation. The interaction term was not included to improve the model, indicating that groups with different length boundaries *all* failed to learn their particular length-referencing alternations equally.

3.2 Filler words

I also examined whether participants overapplied the nasalization rule to the filler words without /d/ or /b/ that might be subject to nasalization. I made the same figures as in Figure 6.1 through 6.7 with another dataset which excluded all the focus words with /d/ or /b/.

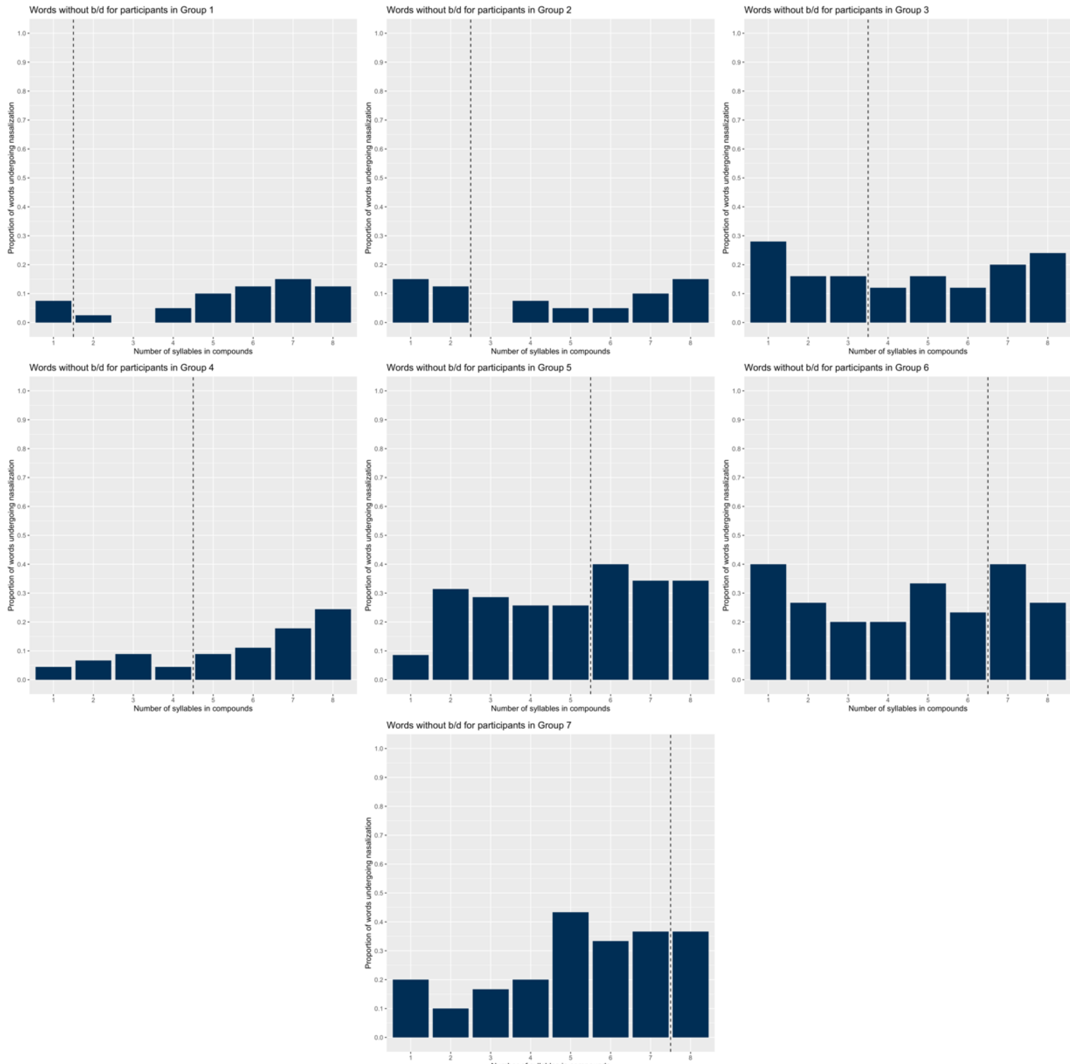

Figure 7.1 through 7.7. Proportions of filler words without /b/ or /d/ undergoing nasalization for participants in each of the 7 groups

The above figures show that in all groups there was a similar nasalization rate on both sides of the dashed line.

I conducted the same analysis as the previous one with only dataset different. The results were almost identical: I ended up with the same optimal model with the remaining factors of *group* and *side*. Still, the effect of *side* was not significant (Pr(>|z|) = 0.29653). The results indicated that applicants overapplied the nasalization process to filler words, no matter the filler word is longer than the boundary length or not.

Table 5. Summary of the fixed-effect variables for Model 2 (for filler words only)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.2449	0.5558	-5.838	0.00000 ***
group2	0.1031	0.7411	0.139	0.88936
group3	1.2685	0.7957	1.594	0.11089
group4	0.6084	0.7403	0.822	0.41113

group5	2.2576	0.7219	3.127	0.00176 **
group6	2.3137	0.7704	3.003	0.00267 **
group7	2.2923	0.8234	2.784	0.00537 **
side	-0.3816	0.3655	-1.044	0.29653

R code: glmer(Nasalized_Response ~ group + side + (1 + side | subject) + (1 + group + side | item), data = filler_data, family = binomial)

3.3 Word types

Finally, I compared the two datasets, one of which contained focus words with b/d only and the other contained the filler words only. I aimed to investigate whether the learning results were different in different datasets. I ran a regression with three concerning variables and all their possible interactions, *wordtype* * *group* * *side* (*wordtype*=1 if only if the word is a focus word). This was not only to examine the effect of *wordtype*, but also to allow for variations in such an effect under the conditions of different *group* and *side*. This model also included maximal random effects (note that *wordtype* should not have a by-item random slope because each item had only one fixed *wordtype* – it either contained b/d or did not contain b/d). I reduced this model using the method of backward stepwise comparison again, and below table summarizes the final model, with all three factors having main effects and a two-way interaction of *side* and *wordtype*. I did not attempt to remove the insignificant factors *side* or *wordtype* because their interaction was significant (Pr(>|z|) = 0.0085) and could significantly improve the model (Pr(>Chisq) = 0.000).

Table 6. Summary of the fixed-effect variables for the dataset-comparison model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.46162	0.46041	-7.519	0.0000 ***
group2	0.16418	0.57476	0.286	0.7751
group3	1.30096	0.61934	2.101	0.0357 *
group4	1.06046	0.56159	1.888	0.0590 .
group5	2.29214	0.56600	4.050	0.0000 ***
group6	2.90064	0.59698	4.859	0.0000 ***
group7	2.83847	0.64302	4.414	0.0000 ***
side	-0.52396	0.33505	-1.564	0.1179
wordtype	0.08027	0.23458	0.342	0.7322
side:wordtype	0.95897	0.36442	2.631	0.0085 **

R code: glmer(Nasalized_Response ~ group + side + wordtype + side:wordtype + (1 + side * wordtype | subject) + (1+ group + side | item), data = all_data, family = binomial)

The effect of *group* and *side* was like the case where two smaller datasets were examined separately. The individual effect of *wordtype* alone was not significant, but the interaction term *side:wordtype* was significant at the 0.01 level, indicating that although different types of words (i.e., focus words or filler words) induced similar rates of nasalization, the difference between words above and below the boundary for focus words (with b/d) was greater than that for filler words (without b/d). This shows that although generally the participants were unlikely to discern whether focus words behaved differently from filler words in generally, they were able to distinguish the discrepancy between both sides of the length boundaries, which were different between the two word types, that is, they were better at doing this for focus words than filler words. This can be easily explained through experimental design - during the training phase, for filler words, the nasalization rate was equal on both sides of the boundary (both were 0%); for focus words, the nasalization rate on the left side of the boundary was 100%, but on the right side, it was 0%, making the distinction extremely clear. Therefore, the substantial difference in input under different word-type conditions could account for the significant effect of the interaction term.

To conclude this subsection, for all words, participants failed to locate the *target* of the rule (*wordtype* was not significant) nor notice the *context* of the rule (*side* was not significant), but were able to integrate information about *wordtype* and *side*, thereby utilizing this interactive cue to determine the nasalization rate. They learned the discrepancy of nasalization rate between two sides of the length boundaries was larger in focus words than in filler words, which indicated that they noticed the different behaviors of two types of words.

4 Discussion

The results show no evidence to support the hypothesis that I proposed earlier. In the test, a length-referencing alternation was always underlearned, no matter what exact numbers were used as a conditioning length boundary in the context of the rule. The results of this essay favor the argument that phonology cannot count, as there is no evidence in this experiment supporting that length can play any rule in the alternations.

Some other related discussions are as follows.

4.1 The typological gap and the bias against length-referencing alternations

The results indicate that the length-referencing alternations is not learned, irrespective of the exact length specified in the rule's context. The length conditions play no role in regulating the alternation during the learning process.

I argue that a synchronic learning bias can lead to the underlearnability of such a length-involved alternation, and thus it can account for the typological observation that few languages (besides Japanese) exhibit similar length-referencing alternation when it counts to more than two (Paster 2019). This argument is familiar in recent years, as a growing number of literatures claim that the typological gap of a phonological pattern can be explained by its underlearnability (e.g., Stanton 2016, Yin & White 2018).

4.2 More on Japanese VVN

In my study on Japanese VVN (Jiang 2023), I proposed that a learning bias can account for the non-internalized pattern of it, and the current study is consistent with that argument because this learning bias was shown to be found in speakers of other languages, thus likely to be universal.

Meanwhile, Jiang (2023) proposed that another factor can account for the non-internalized Japanese VVN pattern, which is token frequency of a word, because token frequency is positively correlated with word length. If token frequency could *completely* account for the variation of VVN in Japanese, no counting is required in the context of this alternation, and the argument that there is a learning bias against counting would be weakened. While further study on the effect of token frequency on Japanese VVN is needed, in the current study I show the existence of the learning bias against counting, which *more or less* shapes Japanese speakers' learning process of the VVN pattern, making the number of moras an inaccessible cue to decide nasalization of /g/.

Even if token frequency could *completely* account for the variation of VVN in Japanese as a synchronic factor, the evidence given in this essay supporting the existence of the learning bias against length-referencing alternation is consistent with the possibility that, at some point in history, mora length played a role in regulating VVNs. However, because the factor of mora length was difficult to learn, it was overshadowed by the factor of token frequency that came along with it in the course of the evolution of the language, resulting in a situation where it does not play a role at present. This theoretical possibility is to be verified or falsified in future studies.

4.3 The frequency of words exhibiting the rule to be learned

In the regression model presented in Table 6, the effect of *group* (representing 'length boundary') was significant. However, while the present experiment controlled for the number of words of different lengths learned by each group, it did not control for the number of words exhibiting the nasalization rule learned by each group – this control was impossible under the current design[12] because the length boundary varied from group to group. This lead to the fact that the participants of Group I encountered only five words to be nasalized in the training phase, all of whose lengths were 1, and which accounted for only 5/80 out of all stimuli, but the participants of Group VII encountered 35 words to be nasalized, whose lengths ranged from 1 to 7, and which accounted for as much as 35/80 out of all stimuli. The occurrence of the nasalization rule in this experiment was actually collinear with the length boundary (*occurrence* = 5 × *group* number), so the variable of *group* in the test actually reflected not only the location of the length boundary but also the occurrence of nasalized words. It is possible that occurence of words exhibiting the nasalization rule could contribute to learning, independent of 'length boundary.' Possible evidence lies in the fact that the mean value of the proportions of nasalized words in the responses closely approximated the frequency of words exhibiting the nasalization rule that was exposed to the participants during the training phase. See the table below for details.

Table 7. Proportions of words to be nasalized exposed to participants and in the response of them

length boundary	frequency of nasalized words in the training	mean value of proportion of nasalized words in responses		
		focus words	filler words	overall
1	5/80=6.25%	8.13%	8.13%	8.13%
2	10/80=12.50%	11.25%	8.75%	10.00%
3	15/80=18.75%	20.00%	18.00%	19.00%
4	20/80=25.00%	21.94%	10.83%	16.39%
5	25/80=31.25%	28.93%	28.57%	28.75%
6	30/80=37.50%	52.50%	28.75%	40.63%
7	35/80=43.75%	49.17%	27.08%	38.13%

Although it comes with some degree of imprecision, generally, the pattern in the responses quantitively obeys the Law of Frequency Matching (Hayes et al. 2009, p. 826) arguing that 'speakers of languages with variable lexical patterns respond stochastically when tested on such patterns. Their responses aggregately match the lexical frequencies'. The results indicate that although speakers failed to learn the nasalization rule conditioned by the *side* of the length boundary, there is possibility that they tacitly obeyed the Law of Frequency Matching.

4.4 Product-oriented learning and a future direction

Considering both the observations of frequency matching and failure in targeting b/d as correct segments to nasalize, the results show that what the learners really learned as a

[12] It is possible to design a control for the number of nasalized inputs, but in doing so, participants would not get identical stimuli across the groups, that is, participants in the lower groups encounter more lower-syllable stimuli so they are able to be exposed to an equal number of nasalized inputs as higher groups.

generalization might not be a context-conditioned rule which only applied to focus words but a specification governing the wordforms that some compounds should have a nasal immediately after the morpheme boundary, regardless of the input segments. The results clearly indicate frequency matched product-oriented generalizations (Bybee, 1995), that is, they learned something on the level of the output forms. They did not learn the phonological rule for the filler words, which was a mapping from input to output, as is called source-oriented generalizations in the terminology of Bybee (1995).

Kapatsinski (2012) argues that presentation conditions may bias a learner in favor of source-oriented or product-oriented generalizations: 'The learner in product-oriented training is presented with one word-form from a paradigm at a time, while the learner in source-oriented training is presented with pairs of words that share the stem.' (p. 70) According to this criterion, my experimental design falls under the category of source-oriented training, as I presented on every page a pair of simple words and a compound word – a concatenated form of the two simple words, indicating a clear compounding rule. My results show that, even when the presentation condition is favorable for extracting rules, participants showed a strong preference for product-oriented generalizations. This aligns with Kapatsinski's (2012) experimental results.

Here I explored the reason for the rules' overgeneralization to filler words in my study. There were actually three sorts of words: short focus words (whose length fell below the length boundary), long focus words, and filler words. The last two sorts of words had one thing in common: they both never nasalized. They added up to more than 50% of total words in any participant group (more specifically, 56.25% to 93.75%, depending on participant groups). Learners in my test sporadically encountered scattered short focus words where the nasalization rule applied, and the learning process of the rule was disrupted by both filler words and long focus words. Nasalization rule was never exemplified by a sequential bunch of short focus words.

I argue that although the current study show that the *side* factor does not play a role, this result is biased to a certain extent by the tendency to draw product-oriented generalizations. Thus, the participants learned the rule only to a limited extent. They did learn something about the rule, i.e., there was an interactive effect of *wordtype* (focus vs. filler) and *side* (below or above), the former of which must be based on the input as a part of the rule. However, they did not learn everything about the rule: they could not directly locate the target of the rule, which had to be the focus words. Thus, we still have no idea whether participants are in fact able to learn the length-referencing alternations in a *more* source-oriented paradigm which draw their attention to the mapping from input to output where length plays a role.

For a future study, the experimental design can focus on promoting source-oriented learning. Two compound words which share the first simple word can be presented sequentially, one being a focus word (with b/d to nasalize), and the other being a filler word (without b/d). For example, on the first page in the training phase, there can be: 'taki' tree, 'boli' cookie; 'taki-moli' tree-shaped cookie; on the second page, there can be 'taki' tree, 'seko' cake; 'taki-seko' tree-shaped cake. This can leave an impression of contrast between focus words and filler words in the ***input*** to participants, and thus they are more likely to eliminate the interference of filler words and focus on the length condition which regulates the application of a rule. In such an experimental design, participants might have better learned the rules. If, under these conditions, they still cannot differentiate the factor of *side*, it provides stronger evidence for the underlearnability of length-referencing alternations.

5 Conclusion

Through an artificial language learning test, this study demonstrates that language learners have a bias against length-referencing alternations, no matter what the length boundary is stated in the context of the rule. Participants in the test could extend the indifference of nasalization rate on both sides of the length boundaries to filler words without target sounds, which should never nasalize, as they failed to distinguish the targets of the length-referencing rule from the non-targets. This learning bias may be the reason why length-referencing alternations are rare in human languages.

Reference:

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255-278.

Becker, M., & Gouskova, M. (2016). Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic inquiry*, 47(3), 391-425.

Becker, M., Ketrez, N., & Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, 84-125.

Becker, M., Nevins, A., & Levine, J. (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language*, 231-268.

Breiss, C. (2021a). Base effects are probabilistic: a case study in lexical conservatism. In *Talk at the 95th Annual Meeting of the LSA*.

Breiss, C. (2021b). *Lexical Conservatism in phonology: Theory, experiments, and computational modeling*. University of California, Los Angeles.

Breiss, C., Katsuda, H., & Kawahara, S. (2021). Paradigm uniformity is probabilistic: Evidence from velar nasalization in Japanese. In *Proceedings of WCCFL* (Vol. 39).

Breiss, C., Katsuda, H., & Kawahara, S. (2022). A quantitative study of voiced velar nasalization in Japanese. *University of Pennsylvania Working Papers in Linguistics, 28*(1), 4.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, 10(5), 425-455.

Carr, P. (2006). Universal grammar and syntax/phonology parallelisms. *Lingua*, 116(5), 634-656.

Finley, S. (2017). Learning metathesis: Evidence for syllable structure constraints. *Journal of Memory and Language*, 92, 142–157.

Hay, J. (2004). *Causes and consequences of word structure*. Routledge.

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in cognitive sciences*, 9(7), 342-348.

Hayes, B., Siptár, P., Zuraw, K., & Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 822-863.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic inquiry*, 44(1), 45-75.

Itô, J., & Mester, A. (1996). Correspondence and compositionality: The ga-gyō variation in Japanese phonology.

Jiang, H. (2023). Learnability of the mora-counting alternation of /g/ nasalization in Japanese compounds. Manuscript. [Available here.]

Kapatsinski, V. (2012). What statistics do learners track? Rules, constraints or schemas in (artificial) grammar learning. *Frequency effects in language learning and processing*, 1, 53-82.

McCarthy, J. J., & Prince, A. (1999). Prosodic morphology 1986. *Phonological theory: the essential readings*, 238-288.

Paster, M. (2019). Phonology counts. *Radical: A Journal of Phonology*, 1, 1-61.

Pater, J., & Tessier, A. M. (2003, August). Phonotactic knowledge and the acquisition of alternations. In *Proceedings of the 15th International Congress on Phonetic Sciences* (Vol. 1180, pp. 1177-1180).

Peperkamp, S., Skoruppa, K., & Dupoux, E. (2006). The role of phonetic naturalness in phonological rule acquisition.

Rose, Y. (1999) 'A structural account of Root node deletion in loanword phonology', *Canadian Journal of Linguistics*, 44, pp. 359–404.

Stanton, J. (2016). Learnability shapes typology: the case of the midpoint pathology. *Language*, 753-791.

Steriade, D., & Stanton, J. (2020). Productive pseudo-cyclicity and its significance. *Talk at LabPhon 17*. [Available here.]

Warner, N., & Arai, T. (2001). Japanese mora-timing: A review. *Phonetica*, 58(1-2), 1-25.

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. Cognition, 130(1), 96–115.

Wilson, C. (2003, March). Experimental investigation of phonological naturalness. In *Proceedings of the 22nd west coast conference on formal linguistics* (Vol. 22, pp. 533-546). Somerville, MA: Cascadilla Press.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5), 945–982.

Yin, S. H., & White, J. (2018). Neutralization and homophony avoidance in phonological learning. Cognition, 179, 89-101.

Appendix 1: Rules on words with b/d to be learned for participants in each of the 7 groups