MaxEnt and sound symbolism in Pokémon names III: Intersecting the effects of mora counts and vowel quality

Abstract

In recent years, we witness a rise of interest in accounting for probabilistic generalizations in linguistic patterns using formal grammatical theories. In this context, Maximum Entropy Harmonic Grammar (MaxEnt HG) has been shown to be a useful analytical tool in modeling probabilistic patterns in various linguistic domains. Some recent studies (Kawahara 2020c, 2021) have proposed to extend the scope of MaxEnt HG by applying it to the analysis of sound symbolism, systematic associations between sounds and meanings. These studies examine what Hayes (2020) refers to as the quantitative signature of MaxEnt HG, the set of typical probabilistic patterns that MaxEnt HG is predicted to generate. The particular quantitative signature that these studies found is wug-shaped curves, which consist of multiple sigmoid curves. Inspired by these studies, the current experiment examined whether we can identify yet another instance of wug-shaped curves in sound symbolism, with the empirical target being the judgment of Pokémon characters' evolution status by native speakers of Japanese. The current experiment shows that as name length increases, the post-evolution responses increase in a sigmoidal manner, and that this sigmoidal curve is shifted depending on the vowel quality of the stimuli, resulting in a wug-shaped curve. To model the results, an MaxEnt HG analysis, equipped with OT-style constraints, is developed as an example of a generative phonological analysis of sound symbolism.

1 Introduction

1.1 General theoretical background

In recent years, we witness a rise of interest in accounting for probabilistic generalizations in linguistic patterns using formal grammatical theories. In this context, Maximum Entropy Harmonic Grammar (MaxEnt HG) has been shown to be a useful analytical tool in modeling probabilistic generalizations in various linguistic domains (Goldwater & Johnson 2003; Hayes 2020). Some recent studies (Kawahara 2020c, 2021) have proposed to extend the scope of MaxEnt HG by applying it to a hitherto understudied domain; namely, the analysis of sound symbolic connections, systematic associations between sounds and meanings (Hinton et al. 1994). These studies examine what Hayes (2020) refers to as the quantitative signature of MaxEnt HG, the set of typical probabilistic patterns that MaxEnt HG is predicted to generate. The particular quantitative signatures that these studies found are wug-shaped curves, which consist of multiple sigmoid curves. Building upon these studies, the current experiment examined whether we can identify yet another instance of a wug-shaped curve in sound symbolism.

MaxEnt HG is an application of a multinomial logistic regression model—which is a general statistical tool—to linguistic analyses (Jurafsky & Martin 2019). It is also possible to understand MaxEnt HG as a stochastic extension of Optimality Theory (OT: Prince & Smolensky 1993/2004), the latter of which has been used, since its initial proposal in the early 1990's, as one of the dominant analytical frameworks among the linguistic community, especially for phonological analyses (McCarthy 2002, 2008).

Just like OT and many other grammatical models in generative grammar, MaxEnt HG takes one level of representation (e.g. underlying representation) and maps it to another level of representation (e.g. surface representation). This mapping between these two representations is regulated by the set of violable constraints, again just like in OT. One major difference between MaxEnt HG and OT is that the constraints are weighted in the former, whereas they are ranked in the latter. Thus, all pieces of information from the entire constraint set are taken into account in MaxEnt HG, whereas OT resorts to only the highest ranked relevant constraint to distinguish between two candidates.[1] Another major difference is that MaxEnt HG assigns a probability distribution over a candidate set, rather than deterministically choosing one output form, as OT does.

Stepping back and viewing it from a more general perspective, MaxEnt HG can be considered as a general tool which takes various sorts of information (=constraints in the parlance of OT) and assigns a probability distribution over several possible outcomes. Because of its generality, this tool, or something akin to it, has been shown to be useful in modeling various patterns in different areas of linguistic inquiry, including speech production (Lefkowitz 2005), categorical speech perception (Hayes 2020; Kluender et al. 1988), probabilistic phonological alternations (Zuraw & Hayes 2017), diachronic phonological changes (Harrison et al. 2002), metric patterns (Hayes et al. 2012), syntactic generalizations (Bresnan et al. 2007), semantics/pragmatics-related judgments (AnderBois et al. 2012), historical changes in syntax (Kroch 1989; Zimmermann 2017), as well as sociolinguistic variation patterns (Rousseau & Sankoff 1978).

[1]From the perspective of more general decision-making strategies, the former corresponds to a familiar regression-based model in which all the pieces of information are taken into consideration, whereas the latter corresponds to a fast-and-frugal heuristic decision making approach, in which decision criteria are ordered in a lexicographic manner (Gigerenzer & Gaissmaier 2011). See Kawahara & Breiss (2021) and Tesar (2007) for further discussion on this point.

Building upon this growing body of studies, two recent studies (Kawahara 2020c, 2021) propose that we can even further extend the scope of MaxEnt HG by applying it to the modeling of sound symbolic patterns, systematic associations between sounds and meanings (see also Kawahara et al. 2019; Kumagai & Kawahara 2019). The present paper reports a follow-up study of this continuing effort. See §1.4 for the full justification why we find it important to continue this line of study.

The approach that we take in this paper is heavily inspired by Hayes (2020), who proposes to take a top-down approach. He points out that as we study probabilistic linguistic patterns, we find specific quantitative patterns that are recurrently observed across different linguistic domains. This general observation can be explained if the same mechanism, e.g. MaxEnt HG, governs these various aspects of linguistic knowledge. In turn then, we can study the mathematical predictions of MaxEnt HG, and examine how these predictions pan out in actual linguistic patterns, especially in domains that have been understudied, such as sound symbolic mappings. The specific mathematical predictions that Hayes (2020) proposes to study are what he refers to as the **quantitative signatures**, which are the set of quantitative patterns that a particular theoretical framework is predicted to generate.

The specifics of the MaxEnt HG mathematics, as well as an actual MaxEnt-based analysis of sound symbolic patterns, are discussed in section 5, once the experimental results are described. To illustrate the main purpose of the experiment, however, we start with the illustration of the quantitative signature of MaxEnt HG in the next subsection.

1.2 The quantitative signature of MaxEnt HG

In this subsection we will review the quantitative signature of MaxEnt HG laid out by Hayes (2020).[2] The general observation that led Hayes (2020) to explore the quantitative signature of MaxEnt HG is that similar probabilistic patterns hold across different linguistic domains and that such probabilistic linguistic generalizations often exhibit multiple sigmoidal curves. From the mathematical point of view, this is a natural consequence of MaxEnt HG, as we illustrate below.

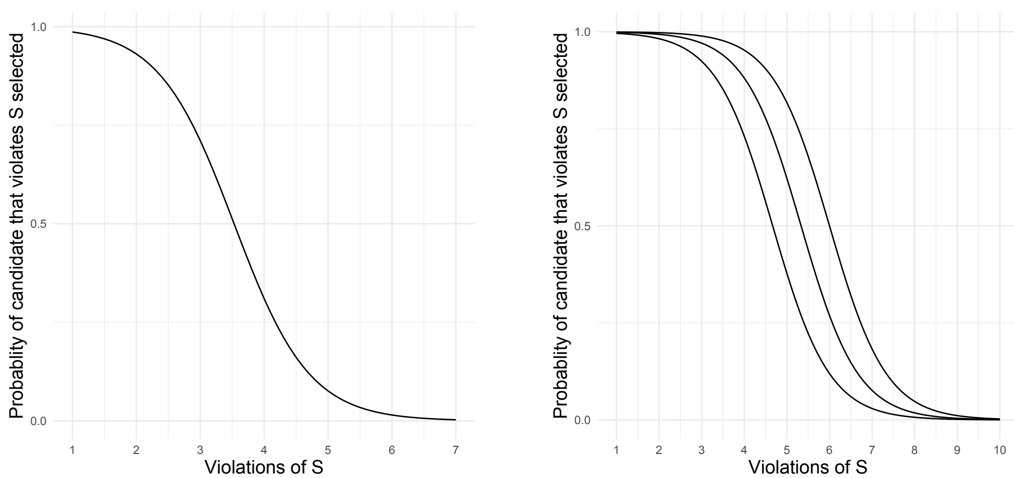
For the sake of illustration, let us posit a scalar constraint S, whose violation is assessed on a linear scale: i.e. 1, 2, 3....N. Let us posit another constraint, B, whose violation profile is accessed in a binary fashion.[3] When B and S conflict with each other, MaxEnt HG predicts that the relationship between the number of violations of S and the probability of the candidate that

[2]This subsection largely owes to Kawahara (2020c) and Kawahara (2021), as well as Hayes (2020). We also note here that Noisy Harmonic Grammar (Boersma & Pater 2016) can also yield the quantitative signatures that are very similar to those of MaxEnt HG, depending on how noise is added during the evaluation of the output candidates (Hayes 2017). Since the difference in prediction between MaxEnt HG and Noisy HG can be extremely subtle (Hayes 2017), we will not address this difference in this paper.

[3]In Hayes's (2020) terminology, S=VARIABLE and B=ON/OFF.

violates S being selected as a winner should manifest itself as a **sigmoid curve** (=S-shaped curve), as illustrated in Figure 1(a). The linear scale (i.e. the constraint violations of S) on the x-axis is converted to a sigmoidal curve in MaxEnt HG, because MaxEnt uses a logistic transformation ($\frac{1}{1+e^{-N}}$) as it calculates the probability distribution of output candidates. In this formula, $-N$ (the linear predictor of exponentiation) is linearly correlated with the number of violation marks assigned by S, and the weight of the constraint B serves as an intercept term for $-N$.

To the sigmoidal curve generated via the interaction between S and B, we can add the effects of another constraint P (for "Perturber"). The result is that this sigmoid curve is shifted horizontally, yielding another sigmoid curve. Hayes (2020) refers to the set of two sigmoid curves as a **wug-shaped curve**, as it looks like a wug, one of the best-known mascot characters in linguistics (Berko 1958). When P can be violated once or twice, it yields three sigmoid curves, which is shown in Figure 1(b). Together, the interaction between the three constraints—S, B, P—results in a **stripy wug-shaped curve**.



(a) (b)

Figure 1: (a) A sigmoid curve generated by the MaxEnt mathematics. The logistic function which generates this curve is $f(x) = \frac{1}{1+e^{-N}}$. While the changes in the x-axis, and hence those in the y-axis, should be discrete, these values are plotted continuously for the sake of illustration. (b) Multiple sigmoid curves, shifted by multiple violations of P, instantiating a stripy wug-shaped curve. The weights of the three constraints are: $S = 1.5, B = 7, P = 1$. These figures are adapted from Kawahara (2021).

In summary, a stripy wug-shaped curve is a quantitative signature of MaxEnt HG, which has three mathematical features: (1) it consists of more than two sigmoid curves, (2) the curves are separated from one another, and (3) the slopes of the sigmoid curves are identical. Two recent experiments (Kawahara 2020c, 2021) argued that we observe a (stripy) wug-shaped curve in sound symbolism. The current experiment was set out to examine whether we can identify another instance of a wug-shaped curve in sound symbolism.

1.3 A (stripy) wug-shaped curve in sound symbolism

Hayes (2020) argues that a stripy wug-shaped curve is commonly observed in probabilistic phonological alternation patterns (Ernestus & Baayen 2003; McPherson & Hayes 2016; Zuraw & Hayes 2017) and other linguistic domains (see also the website accompanying Hayes's paper, "A gallery of wug-shaped curves").[4] Kawahara (2020c, 2021) followed this general method and demonstrated that (stripy) wug-shaped curves are observed in the domain of sound symbolism, systematic connections between sounds and meanings (Hinton et al. 2006). The current experiment is a direct follow-up of these two experiments.

The experiments reported by Kawahara (2020c, 2021), as well as the current experiment, are situated within the research paradigm dubbed "Pokémonastics" (Shih et al. 2019), in which researchers use Pokémon character names to study the nature of sound symbolic patterns in natural languages (Kawahara et al. 2018b).[5] Kawahara & Breiss (2021) summarize the several research advantages of this research program for cross-linguistic studies of sound symbolism; for the purpose of the current paper, which focuses on the examination of MaxEnt as an analytical framework, it suffices to note that many Pokémon characters undergo evolution, and when they do so, they generally get larger and heavier and are also called by a different name. The first Pokémonastics study, which analyzed the existing Pokémon names in Japanese (Kawahara et al. 2018b), found that the names of evolved characters tend to be longer, and are more likely to contain voiced obstruents. For instance, *Ki-mo-ri* (3 moras) evolves into *Ju-pu-to-ru* (4 moras), and the latter contains a voiced obstruent [dʑ] (a voiced palato-alveolar affricate) name initially. Likewise, *ri-ri-i-ra* (4 mora) evolves into *yu-re-i-do-ru* (5 moras) and acquires a voiced obstruent [d] in its new name.

The effects of voiced obstruents arguably arise from the frequency code (Ohala 1994). Since voiced obstruents are characterized by low frequency energy during their constrictions and/or in adjacent vowels in terms of their low f0 and F1 (Kingston & Diehl 1994; Stevens & Blumstein 1981), this low frequency characteristics may be mapped on a large image, because a vibrator that omits low frequency sounds is *ceteris puribus* larger. The fact that longer means stronger can be attributed to the iconicity of quantity (Haiman 1985), in which a larger quantity is expressed by longer linguistic expressions (Dingemanse et al. 2015).

Kawahara (2020c) used these two sound symbolic associations to examine whether we would observe a wug-shaped curve in sound symbolism, by manipulating the mora counts and the pres-

[4] `https://linguistics.ucla.edu/people/hayes/GalleryOfWugShapedCurves/index.htm`

[5] For the importance of studying sound symbolism from the perspective of cognitive science and linguistics, see Dingemanse et al. (2015) and Kawahara (2020a), respectively. See also the references cited therein. There are now many overview articles on this topic, which are cited in Kawahara (2020a).

ence of a voiced obstruent. The experiment asked native speakers of Japanese whether each name was better suited for a pre-evolution Pokémon character or a post-evolution character. Kawahara (2020c) found that the increase in name length increases post-evolution responses in a sigmoidal manner, and that a name-initial voiced obstruent horizontally shifts the entire sigmoid curve, which together results in a wug-shaped curve. Kawahara (2021) built on this finding and varied the number of voiced obstruents, showing that we observe three sigmoid curves, separated from each other according to the number of voiced obstruents. The latter study thus found that a wug-shaped curve in sound symbolism can be stripy.

The current experiment continues this effort and explores whether we would observe another stripy wug-shaped curve. This experiment intersects mora counts, following the two previous experiments (Kawahara 2020c, 2021), with vowel quality, a new manipulation in the current study.

1.4 Motivating the current experiment

At this point we would like to clarify why we are running an experiment that is arguably similar to those reported by Kawahara (2020c, 2021). One general reason is that replication is an important practice that has been under-appreciated in linguistics (and psychology) (see e.g. Chambers 2017; Porte 2012; Roettger & Baer-Henney 2019). One aim of the current study is to examine whether we would obtain another instance of wug-shaped curve with a set of stimuli that is very different from the two previous studies. The second reason is that a wug-shaped entails that different curves should have the same slopes, and a Bayesian analysis is necessary to examine its aspect (for which see below for more). None of the patterns discussed by Hayes (2020) have been analyzed from this perspective. In fact, to the best of our knowledge, Kawahara (2021) is the only dataset instantiating wug-shaped curves that has been analyzed using a Bayesian method. Therefore, having more Bayesian analyses is informative in accessing whether MaxEnt HG is truly a suitable framework to model probabilistic patterns in linguistics. Third, at a descriptive level, we were interested in whether different vowel quality would impact the judgment of evolvedness in Pokémon names, and if so, which phonological dimension would be relevant. Finally, while sound symbolism is receiving a remarkable degree of attention from psychologists and cognitive scientists in recent years (Nielsen & Dingemanse 2020), it is not so much the case among the generative linguistics community (Alderete & Kochetov 2017; Kawahara 2020a). By further exploring a possible parallel between probabilistic phonological patterns and sound symbolic patterns, we would like to highlight the potential usefulness of analyzing sound symbolic patterns in the context of generative linguistic inquiry.

To reiterate, the new factor that is manipulated in this experiment is the vowel quality difference ([a] vs. [i] vs. [u]). While the main purpose of the experiment is the examination of MaxEnt HG as an analytical framework for linguistics, studying the effects of vowel quality is interesting from the perspective of sound symbolism research. On the one hand, vowel quality does not seem to play a crucial role in the sound symbolic patterns of the existing Japanese Pokémon names (Kawahara et al. 2018b). On the other hand, many studies in the sound symbolism literature have shown that low vowels tend to be judged to be bigger than high vowels (e.g. Jespersen 1922; Newman 1933; Sapir 1929; Ultan 1978), arguably because the oral aperture is wider for the former than for the latter. In addition, there is a general observation that back vowels may be judged to be larger than front vowels (Berlin 2006). This is because the second formant frequency is lower for back vowels than for front vowels, and the physics tells us that low frequency sounds are omitted from a large resonating chamber (Ohala 1994). If the vowel quality (vowel height and/or vowel backness) triggers size-related sound symbolic effects, it would not be surprising if we observe the effects of vowel quality in Pokémon names, because evolved characters are generally larger. However, which phonological dimension—height vs. backness—determines the sound symbolic effects of different vowels is still debated (see e.g. Dingemanse et al. 2015; Knoeferle et al. 2017; Shinohara & Kawahara 2016). The current experiment can be understood as offering a new contribution to this debate.

If both vowel height and backness matter in determining the size-related sound symbolism, the specific prediction is that speakers should judge [a] (low back) to be the larger than [u] (high back), which is in turn larger than [i] (high front). The current experiment aimed at examining whether we would observe this three-way distinction in the context of Pokémonastic studies. A previous Pokémonastic experiment by Kumagai & Kawahara (2019), which used a two-alternative-forced choice (2AFC) format, shows that Japanese speakers find names with [a] to be more suitable for post-evolution characters than names with [i] and those with [u], although that experiment did not directly compare [i] and [u]. Another 2AFC experiment targeting English speakers found that they judge names with [u] to be more suitable for post-evolution characters than those with [i] (Kawahara & Moore 2021). In the current experiment we were interested in re-examining these results, because the 2AFC format, in which two names are presented as a pair, may overestimate the effect size of sound symbolic connections (Kawahara et al. 2021; Nielsen & Rendall 2011, 2013; Westbury et al. 2018). Therefore, there is a general need in the sound symbolism research to reexamine sound symbolic effects in a more conservative task in which stimuli are presented in isolation (see in particular Westbury et al. 2018). This reexamination seemed necessary, partly because Kawahara et al. (2018a) found that Japanese pre-schoolers did not judge names with [a] to be more suitable for post-evolution characters than those names with [u]. In general, no Pokémonastic experiments have examined a tripartite vocalic distinction within the same group of speakers. The current experiment therefore attempts to fill these gaps in the literature.

2 Methods

2.1 Stimuli

Table 1 shows the set of stimuli used in the current experiment in IPA. The mora counts were varied from 2-moras to 6-moras, which each correspond to the minimum and maximum name lengths in the real Pokémon names. All names consist of open CV syllables so that syllable boundaries and mora boundaries coincided with one another (Kawahara 2016). In the current experiment, the names had the same vowel, either [a], [i] or [u], in all the syllables.[6] No voiced obstruents appeared in the stimuli, because they have clear sound symbolic values for Japanese speakers (Hamano 1998). We also avoided using [p] for the same reason (Kumagai 2019).

2.2 Procedure

In the instructions, the participants were reminded that Pokémon characters often undergo evolution, and that when they do, they generally tend to get heavier, larger and stronger. In the main trial of this experiment, participants were provided with one nonce name per trial, and were asked to judge whether each name was better suited for a pre-evolution character or a post-evolution character. The order of the stimuli was randomized for each participant. The stimuli were presented in the *katakana* Japanese orthography, which is used for the real Pokémon names. The participants were asked to provide their responses based on their intuition, rather than thinking about right or wrong answers. They were also asked to silently read each stimulus before making their decisions, so that they would use their auditory impression as they provided their responses.

2.3 Participants

The experiment was administered online using SurveyMonkey. There were no compensations, monetary or otherwise, for participating in the experiment. The current experiment was advertised on a Pokémon fan blog, and a total of 507 people completed the experiment over a single weekend.[7] Eight speakers reported that they were non-native speakers of Japanese. As many as 101 participants reported that they took part in a Pokémonastics experiment before (which is unsurprising because a number of Pokémonastics experiments had been advertised on this blog).

[6] Kawahara (2021) states that an attempt to use the same vowels in long names resulted in artificial-sounding names. We submit that Kawahara (2021) was not creative enough in this regard.

[7] `http://pokemon-matome.net`. We thank the blog administrator for their help with the participant recruitment. We would like to take this opportunity to make the point that being able to collect as many as 500 participants over a single weekend without monetary compensation is a distinct forte of using Pokémon names as a topic of exploration in sound symbolism research. See Kawahara (2020b) for potential applications of the Pokémonastic research for teaching and public outreach.

Table 1: The stimuli in IPA.

	[a]	[i]	[u]
2 moras	[ha.sa]	[çi.ɕi]	[ɸu.su]
	[ɾa.ja]	[ɾi.mi]	[ɾu.ju]
	[ka.ja]	[ki.ni]	[ku.ju]
	[ta.sa]	[tɕi.ni]	[tsu.su]
	[wa.ma]	[mi.çi]	[nu.ɸu]
3 moras	[ha.sa.ɾa]	[çi.ɕi.ki]	[ɸu.su.ɾu]
	[ɾa.wa.ja]	[ɾi.mi.ki]	[ɾu.ju.mu]
	[ka.ja.wa]	[ki.ni.ɾi]	[ku.ju.nu]
	[ta.sa.ma]	[tɕi.ni.mi]	[tsu.su.mu]
	[wa.ma.ɾa]	[mi.çi.ɾi]	[ɸu.mu.ɾu]
4 moras	[ha.sa.ɾa.na]	[çi.ɕi.ki.mi]	[ɸu.su.ɾu.mu]
	[ɾa.wa.ja.na]	[ɾi.mi.ki.ni]	[ɾu.ju.nu.ku]
	[ka.ja.ta.ɾa]	[ki.ni.ɾi.çi]	[ku.ju.ɾu.nu]
	[ta.sa.ma.ja]	[tɕi.ni.mi.ɾi]	[tsu.su.mu.ɾu]
	[wa.ma.ɾa.na]	[mi.çi.ɾi.ni]	[mu.ɸu.ɾu.ku]
5 moras	[ha.sa.ɾa.na.ja]	[çi.ɕi.ni.ki.mi]	[ɸu.su.tsu.mu.ɾu]
	[ɾa.wa.ta.ja.na]	[ɾi.mi.ki.tɕi.ni]	[ɾu.ju.ku.nu.mu]
	[ka.ja.na.ta.ɾa]	[ki.ni.ɾi.çi.ɾi]	[ku.ju.ɸu.ɾu.nu]
	[ta.sa.ma.na.ja]	[tɕi.ni.mi.ɾi.çi]	[tsu.su.ju.mu.ku]
	[wa.ma.sa.ɾa.na]	[mi.çi.tɕi.ɾi.ni]	[mu.ɸu.su.ɾu.nu]
6 moras	[ha.sa.ɾa.ta.na.ja]	[çi.ɕi.ɾi.ni.ki.mi]	[ɸu.su.nu.tsu.mu.ɾu]
	[ɾa.wa.ta.ma.ja.na]	[ɾi.mi.ki.tɕi.ni.mi]	[ɾu.ju.tsu.ɸu.nu.mu]
	[ka.ta.ɾa.na.ta.ma]	[ki.ni.ɾi.mi.çi.ɾi]	[ku.ju.ɸu.ɾu.nu.tsu]
	[ta.sa.ma.na.ɾa.ja]	[tɕi.ki.ni.mi.ɾi.ki]	[tsu.su.mu.ju.ku.ɾu]
	[wa.ma.sa.ɾa.na.ta]	[mi.çi.ki.tɕi.ɾi.ni]	[mu.ɸu.su.ɾu.nu.ku]

The data from these participants were excluded, and as a result, the data from the remaining 398 participants entered into the following statistical analysis.

2.4 Statistics: Bayesian regression analyses

The results were analyzed using a Bayesian mixed effects logistic regression model. While it is impossible to provide a full review of the advantages of Bayesian analyses, we provide a very brief review in this subsection (see e.g. Kruschke 2014; Kruschke & Liddell 2018 for accessible introductions). Bayesian analyses take prior information and the experimental data into consideration to produce a range of possible values for each estimated parameter, which are referred to as posterior distributions. Unlike a more traditional frequentist analysis, we can interpret these posterior distributions as directly reflecting our uncertainty about the estimates. People often interpret

95% confidence intervals calculated in a frequentist analysis as if they directly reflect the ranges of possible values that the estimates can take, but this is a misinterpretation (e.g. Kruschke & Liddell 2018). Being able to provide a more intuitive interval estimate for a parameter of interest is one virtue of Bayesian modeling. As a useful heuristic, we can examine the middle 95% of the posterior distribution, known as 95% Credible Interval (95% CI) or 95% High Density Interval. If the 95% CI does not include 0, then we can take that effect to be meaningful.

One important prediction that MaxEnt HG makes is that the interaction terms between the two crucial factors—i.e. the effects of mora counts and vowel quality in the current experiment—should be null, because the slopes of the different curves in wug-shaped curves should be identical. Meaningful interaction terms, on the other hand, would indicate that the slopes are different, as interaction terms function as slope adjustment terms (Winter 2019). In order to access the null effects of the estimates, which is possible only in Bayesian analyses but not in frequentist analyses (Gallistel 2009), we can resort to an analysis using ROPE (Region of Practical Equivalence: e.g. Kruschke & Liddell 2018; Makowski et al. 2019). The basic idea is that we define a range that is "practically equivalent" to a point estimate, which in this case is $\beta = 0$. In principle, each researcher can define the width of the range of what it means to be "practically equivalent," but here we follow Makowski et al. (2019) and take a standardized effect size of 0.1 to define that range.[8] Since a standardized effect size of logistic regression is $\frac{\pi}{\sqrt{3}} = 1.8$, the ROPE ranges from [-0.18, 0.18]. The `bayestestR` package (Makowski et al. 2020) was used to calculate how many posterior samples for the coefficients of the interaction terms were included in this ROPE.

2.5 Actual implementation

Following the open science initiative in linguistics and psychology (Chambers 2017; Garellek et al. 2020; Winter 2019), the raw data, the R markdown file with analysis codes, and the Bayesian posterior samples are all made available at Open Science Framework (osf) repository.[9] The R markdown file also contains additional analyses, such as conditional effects and a posterior predictive check, which are not reported in the paper. Interested readers are welcome to further examine the data.

The actual analysis was implemented using the `brms` package (Bürkner 2017) and R (R Development Core Team 1993–). The dependent variable was the binary-coded responses (0 = pre-evolution; 1 = post-evolution). The fixed predictor variables were mora counts, vowel quality and their interaction terms. The mora count was centered because it is a numeric variable (Winter 2019). The random factors included free-varying intercepts for items and participants, as well as free-varying slopes for participants for the two fixed factors and their interaction terms. Being able to fit a model with a complex random structure without convergence issues is a yet another virtue

[8] This effect size corresponds to a negligible effect size in Cohen's (1988) widely used proposal.

[9] `https://osf.io/b6s83/?view_only=a29df2c023f246e399124958e74f9ccc`

of Bayesian analyses (see e.g. Eager & Joseph 2017).

The weakly informative priors, the default in `brms`, were used. Four chains were run with 2,000 iterations. The first 1,000 iterations were disregarded as warmups. All the $\hat{R}$-values were 1.00 and there were no divergent transitions, indicating that the chains mixed successfully. The R markdown file available at the osf repository shows complete details of this analysis.

3 Results

Figure 2 plots the post-evolution response ratios for each item, averaged over all the participants, with each panel representing different vowel conditions. For each vowel condition, the `ggplot2` package (Wickham 2016) was used to superimpose a logistic curve. We observe a steady increase in post-evolution responses as the name lengths increase, going from left to right in Figure 2. Moreover, it appears that each vowel condition instantiates a sigmoidal (=S-shaped) curve. Note, however, that we are telling `ggplot2` to fit a logistic/sigmoid curve, and therefore, any pattern can in principle look as if it could be modeled with a sigmoid curve.

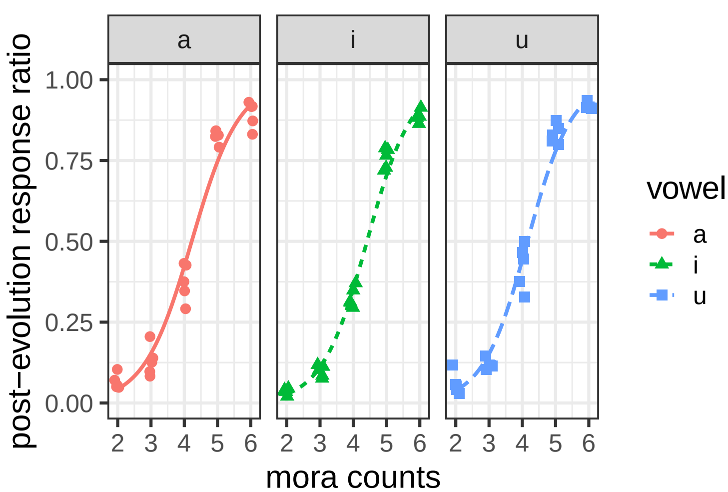


Figure 2: The post-evolution response for each item averaged over all participants for the three vowel conditions. The points are horizontally jittered by 0.1. Logistic curves are superimposed using `ggplot2` for each vowel condition.

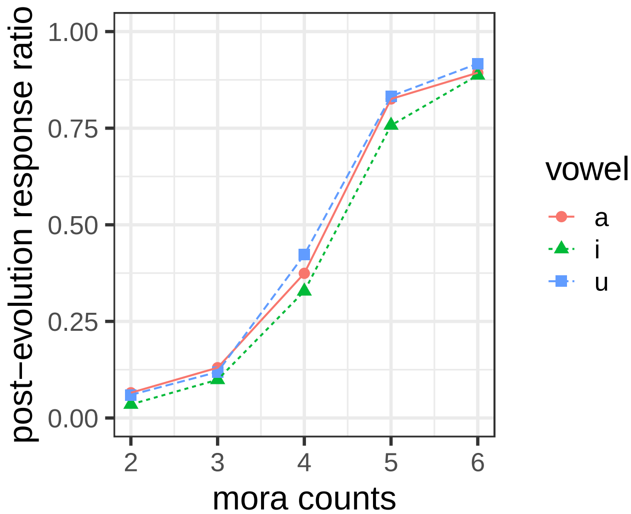


Figure 3: The line-plot with grand averages.

To address this concern, Figure 3 presents a line-plot of grand averages for each condition. This analysis, unlike Figure 2, does not presuppose that sigmoid curves would fit the observed data points well. Nevertheless, each curve does appear to instantiate a sigmoidal curve in that the slopes are rather steep in the middle range (i.e. between the 3-mora condition and 5-mora condition), whereas the change at the left and right edge of the x-axis continuum does not substantially impact the judgment. Having a steep change in the middle of the x-axis continuum is a characteristic of sigmoid curves, as we illustrated in Figure 1(a). This result is also in line with the two previous studies (Kawahara 2020c, 2021) which manipulated mora counts in a way that is similar to the current experiment (although the actual stimuli used in the current experiment are very different).

Table 2 shows the model summary of the Bayesian mixed effects logistic regression analysis. First, the intercept is negative. Since the mora count is centered and the baseline for the vowel quality is [a], this negative intercept indicates that 4-mora long names with [a] induced the post-evolution responses less than 50% of the time (the model prediction is $\frac{1}{1+e^{0.46}} = 0.39$). The β-coefficient for the effects of mora count is positive and its 95% CI does not include 0, which shows that an increase in mora counts credibly increased the post-evolution responses.

The β-coefficient for the difference between [a] and [i] is negative, and its upper bound of the 95% CI is lower than 0. This indicates that [i] meaningfully lowered the post-evolution responses with respect to [a]. The 95% CI for the β-coefficient for the difference between [a] and [u], on the other hand, includes 0, suggesting that [a] and [u] do not meaningfully differ from one another. The general conclusion we can draw from these results is that at least for the current case at hand, it is vowel backness, not vowel height, which impacted the post-evolution responses, supporting the proposal that it is vowel backness—or second formant frequency—that is relevant for size-related vocalic sound symbolism (Berlin 2006; Ohala 1994). At least in the current experimental

Table 2: Summary of the Bayesian mixed effects logistic regression model.

	β	error	95% CI
intercept	-0.46	0.10	[-0.65, -0.27]
mora count	2.11	0.09	[1.93, 2.29]
[a] vs. [i]	-0.38	0.12	[-0.62, -0.15]
[a] vs. [u]	0.12	0.12	[-0.11, 0.35]
mora count $\times$ [a] vs. [i]	0.13	0.09	[-0.05, 0.31]
mora count $\times$ [a] vs. [u]	0.09	0.10	[-0.10, 0.28]

setting, vowel height (=the distinction between [a] vs. [u]) did not seem to credibly impact the post-evolution responses.[10]

The 95% CI for the two interaction terms includes 0, suggesting that the current data do not offer convincing evidence that the slopes between the three curves meaningfully differ from one another, as predicted by MaxEnt HG. However, the 95% CIs are not fully contained in the ROPE (=[-0.18, 0.18]), and therefore, we were unable to fully accept the null effect for these two interaction terms. Further examination of the posterior distributions show that 73.7% ([a] vs. [i]) and 84.1% ([a] vs. [u]) of the 95% CIs are contained in this ROPE. When we examine the whole posterior samples (Makowski et al. 2019), 77.0% and 86.0% of them are contained in the ROPE. We are therefore about 74%$\sim$85% confident that the slopes between the three curves are identical.

4 Discussion

4.1 Are the results (stripy) wug-shaped curves?

Let us first discuss whether the current experimental result in Figure 3 supports the prediction of MaxEnt HG, instantiating a stripy wug-shaped curve. To repeat its mathematical definition, a wug-shaped curve, as predicted by MaxEnt HG, has three defining mathematical features (Hayes 2020): (1) it consists of multiple sigmoid curves, (2) the curves are separated from one another, and (3) the slopes of the sigmoid curves are identical.[11]

[10]Since the 95% CI is not fully contained in the ROPE, we cannot accept the hypothesis that this effect is indeed null (Kruschke & Liddell 2018). The current results do not allow us to make a conclusive statement about the effects of vowel height, and whether we can conclude the true null effect has to be explored in future studies. In addition, the lack of credible effects of vowel height in the current experiment does not preclude the possibility that vowel height can be relevant for other sound symbolic meanings. See Dingemanse et al. (2015) for a summary of sound symbolic meanings arising from differences in vowel height and vowel backness.

[11]It is possible to generate a wug-shaped curve with different slopes, as long as we admit a locally conjoined constraint within MaxEnt HG, as in fact proposed by Shih (2017). For the sake of restrictiveness of the theory, we proceed with the assumption that wug-shaped curves should have the same (or comparable) slopes (Hayes 2020).

Starting with the third defining feature, the ROPE analyses of the two interaction terms show that we can be 70%∼85% confident that the slopes between the three curves can be considered as identical for practical purposes. The results were less clear-cut than those of Kawahara (2021), whose interaction term was fully contained in the ROPE. Instead of a binary yes-significant vs. not-significant dichotomy embraced in frequentist statistical tests, Bayesian analyses can provide a quantifiable measure of how much certainty we can accept the (null) hypothesis. Despite the fact that the current results are not as clear-cut, we nevertheless submit that they are encouraging.

More generally speaking, the current results highlight the importance of a Bayesian approach in accessing wug-shaped curves in linguistic patterns, as it tells us with how much certainty we can conclude that the slopes are practically identical. More Bayesian analyses are warranted to examine to what extent we can conclude that other putative examples of wug-shaped curves—particularly those discussed by Hayes (2020)—show practically identical slopes.

As for the second requirement of the stripy wug-shaped curves, the current experiment revealed that there are two sigmoidal curves: one for [i] and another curve for two back vowels. The result therefore is a wug-shaped curve, but (unfortunately) not a stripy one.

As noted in the two previous studies (Kawahara 2020c, 2021), the first defining characteristic of a wug-shaped curve is hardest to defend, and it is actually impossible to be certain that a sigmoid curve is the best function to model the current data. This is because linguistic data always involve some good degree of natural variability, and no linguistic data would perfectly fit the mathematical definition of sigmoids. Moreover, there are countless numbers of mathematical functions that can be potentially fit to the data, and therefore we need to resort to informed guesses based on cross-linguistic considerations. We can, for example, posit a neural network consisting of multiple nodes, each of which is activated via a logistic function. Such a neural network may be able to better fit the data, but it may have excessive expressive power for linguistic theorization (although we do not wish to imply here that neural networks are not suited for modeling of linguistic behavior: see e.g. Manning et al. (2020) and Linzen & Baroni (2021).[12]

For the current experiment, we maintain that it is a reasonable conjecture that sigmoid functions fit the current data well, since there is a steeper increase in the middle range, compared to the low and high ends of the x-axis continuum. This is an aspect of sigmoidal curve that Hayes (2020) emphasizes, under the following slogan: "certainty is evidentially expensive" (p.6). The next section shows that indeed, MaxEnt HG, which predicts these curves to be sigmoidal, models the experimental data very well.

[12]The analysis in §5 shows that the current dataset at least does not require an expressive power beyond that of MaxEnt HG (i.e. logistic regression).

4.2 On the effects of vowel quality

Before we proceed to the MaxEn HG analysis, we would like to briefly discuss a few other topics, starting with what we found about the effects of the vowels. In the current experiment, the crucial distinction seems to be [a]/[u] vs. [i], that is, back vowels vs. front vowels. This result is in line with the proposal by Berlin (2006), who argues that vowel backness plays a crucial role in determining size-related sound symbolism (see also Knoeferle et al. 2017). The current result, however, is at first blush at odds with the finding by Kumagai & Kawahara (2019), who showed that when presented with a pair of names with [a] and those with [u], Japanese speakers find the former to be more suitable for post-evolution Pokémon characters (cf. Kawahara et al. 2018a). We suspect that this may be a case in which sound symbolic effects were overestimated in a 2AFC experimental format (Kawahara et al. 2021; Nielsen & Rendall 2011, 2013; Westbury et al. 2018). The current result shows that the distinction between [a] and [u] is not robust enough to be clearly observed when each stimulus is presented in isolation rather than in pairs. This result highlights the importance of examining sound symbolic experiments in an experimental format that is more conservative than a 2AFC format (Westbury et al. 2018).

While the current experiment has revealed a credible effect of the vowel backness difference, the effect appears to be not as strong as that of voiced obstruents found in Kawahara's (2021) study. The separation between the three curves due to the different numbers of voiced obstruents in Kawahara (2021) seems more substantial than the wug-shaped curve obtained in the current experiment; i.e. the current wug is skinner, despite the fact that the vowel quality difference is realized in all the syllables in the current stimuli. The logistic regression coefficient for the effects of voiced obstruents in Kawahara's (2021) experiment is 0.49, compared to the regression coefficient for the [a] vs. [i] difference in the current experiment, which is 0.38. This difference may arise from the fact that in the Japanese mimetic system, a voiced obstruent is actively deployed to signal certain sound symbolic meanings (e.g. [ton-ton] "knocking" vs. [don-don] 'knocking hard") (Hamano 1998). How a phonological characteristic of a particular language affects the sound symbolic judgments of its speakers is an interesting topic that is worthy of further exploation.

4.3 Cumulativity and different decision making strategies

The second aspect of the data that is worth emphasizing is the fact that both vowel backness and mora counts cumulatively affected the post-evolution responses. This result is an expected one, as long as a mechanism like MaxEnt HG governs the sound symbolic mappings, because the MaxEnt math takes into account all the pieces of information that are available.

On the other hand, this general result is not expected if a mechanism like Optimality Theory (Prince & Smolensky 1993/2004) is responsible for the sound symbolic mappings. This is because

OT takes only the highest ranked constraint into account when deciding between two candidates, as achieved by the strict domination of constraint rankings. In the current context, an OT-like mechanism predicts that it is either mora counts or vowel quality, whichever is more important, that would determine the participants' responses, but that is not what the current data seems to suggest.

We acknowledge that nobody has attempted to apply OT to model sound symbolic patterns, but the point can be more general. The current experiment shows that people take into account both mora counts and vowel quality when making a decision about evolvedness of Pokémon characters. A fast-and-frugal heuristic decision making strategy (Gigerenzer & Gaissmaier 2011), of which OT is an example, would be unable to model the sound symbolic judgment pattern (see Kawahara & Breiss 2021 who reached a similar conclusion).

5 A MaxEnt HG analysis

This section develops a generative analysis of sound symbolism to model the sound symbolic connections obtained in the experiment. There are a couple of notable features of this analysis (Kawahara et al. 2019; Kawahara 2020c, 2021). First, just like many generative phonological analyses, we are developing a model that maps one representation to another representation. In "standard" phonological analyses, the mapping that is modeled is usually between underlying representation to surface representations. In the analysis developed below, the mapping is from sounds to meanings. Second, since the sound symbolic mappings are inherently stochastic (Dingemanse 2018; Kawahara et al. 2019), we need a model that captures the stochastic nature, and as we will see below, MaxEnt HG is a useful tool for that purpose.[13] Third, in order to make clear that our analysis is an extension of standard phonological analyses, we deploy the sort of constraints that have been used in the OT research tradition (Prince & Smolensky 1993/2004). More specifically, we use the constraint schemata of McCarthy (2003) to highlight the parallel between formal phonological analyses and the generative analyses of sound symbolic patterns.

As discussed at the outset of the paper, MaxEnt HG is mathematically equivalent to multinomial logistic regression (Jurafsky & Martin 2019). Therefore, there is some conceptual overlap between the statistical analysis presented in section 3 and the MaxEnt analysis presented in this section. However, we take these two methods to be achieving something different. On the one hand, a logistic regression model is a statistical means to explore what we can conclude based on

[13] An obvious alternative analytical framework is Stochastic Optimality Theory (Boersma & Hayes 2001), which is also able to capture stochastic linguistic generalizations. For a problem that the sort of the pattern obtained in the current experiment presents to Stochastic OT, see Jäger (2007), Kawahara (2020c) and Zuraw & Hayes (2017). To put it in a nutshell, Stochastic OT is unable to model counting cumulativity effects (Jäger 2007), of which the effects of mora counts are a typical example.

experimental results. On the other hand, the analysis developed here is a generative analysis, which is a model of the knowledge that lies behind the patterns observed in the experiment. See Breiss & Hayes (2020) for further discussion on this difference. See also Kawahara (2021) for specific restrictions that are often imposed upon linguistic analyses but not on statistical analyses; e.g. constraints cannot reward candidates in linguistic analyses, whereas no comparable restrictions hold in statistical modeling.

Here we offer a brief explanation of the MaxEnt math and refer the readers to other published papers for further details (Breiss & Hayes 2020; Hayes 2020; Hayes & Wilson 2008; Kawahara 2020c; McPherson & Hayes 2016; Zuraw & Hayes 2017). In MaxEnt HG, just as in OT, output candidates are evaluated against a set of constraints, each of which bears a numerical weight. Given a set of constraint violation profiles and constraints' weights, each candidate gets a harmony score (H), which is the weighted sum of constraint violations: $H = \sum w_i C_i(x)$, where w_i = the weight of the i-th constraint and $C_i(x)$ = how many times the candidate violates the i-th constraint. The predicted probability of each candidate $x_j, p(x_j)$, is determined by the Softmax Function used in the machine learning literature. We first take e raised to the negative of the harmony score (e^{-H}), the e^{-H} values for all the candidates are summed, and each e^{-H} is relativized with respect to that sum. The Softmax Function assures that all probabilities sum to 1.

To model the current experimental results, we posit the three constraints defined in (1). The first and third constraints are adapted from Kawahara (2020c, 2021).

(1) The list of the constraints

a. *LONGPRE: Assign a violation mark for each mora in a pre-evolution character name.
b. *BACKPRE: Assign a violation mark for each pre-evolution character name consisting of back vowels.
c. *POST: Assign a violation mark for each post-evolution name.

The first constraint prefers long names to be used for post-evolution characters, and corresponds to the numerically-violable constraint S that was used for the illustration of a stripy wug-shaped curve in Figure1(b). The second constraint prefers that names with back vowels are used for post-evolution character names, and this corresponds to the perturber constraint P. The last constraint penalizes post-evolution character names in general, which corresponds to the binary constraint B. This constraint determines the general preference for pre-evolution characters, functioning as the intercept of the linear predictor of the sigmoidal curves.

The MaxEnt tableaux for all the conditions appear in (2). The leftmost column shows the phonological inputs, and the second column shows the two output semantic meanings. The constraint violation profiles are shown in the next three columns. The observed percentages are shown in the rightmost column, which correspond to the grand averages obtained in the experiment.

(2) The MaxEnt Tableaux

		w = 1.52	w = 0.30	w = 6.66				
Input	Output	*LongPre	*BackPre	*Post	Harmony (H)	e^{-H}	Predicted	Observed
2 moras, [a]	Pre	2	1		3.31	0.036	96.59	93.47
	Post			1	6.66	0.001	3.41	6.53
3 moras, [a]	Pre	3	1		4.82	0.008	86.23	87.00
	Post			1	6.66	0.001	13.77	13.00
4 moras, [a]	Pre	4	1		6.33	0.002	58.07	62.60
	Post			1	6.66	0.001	41.93	37.40
5 moras, [a]	Pre	5	1		7.84	0.000	23.45	17.40
	Post			1	6.66	0.001	76.55	82.60
6 moras, [a]	Pre	6	1		9.35	0.0001	6.35	10.70
	Post			1	6.66	0.001	93.65	89.30
2 moras, [i]	Pre	2			3.02	0.049	97.44	96.48
	Post			1	6.66	0.001	2.56	3.52
3 moras, [i]	Pre	3			4.53	0.011	89.38	90.10
	Post			1	6.66	0.001	10.62	9.90
4 moras, [i]	Pre	4			6.03	0.002	65.05	67.20
	Post			1	6.66	0.001	34.95	32.80
5 moras, [i]	Pre	5			7.54	0.001	29.16	24.30
	Post			1	6.66	0.001	70.84	75.70
6 moras, [i]	Pre	6			9.05	0.0001	8.35	11.30
	Post			1	6.66	0.001	91.65	88.70
2 moras, [u]	Pre	2	1		3.31	0.036	96.59	99.40
	Post			1	6.66	0.001	3.41	0.60
3 moras, [u]	Pre	3	1		4.82	0.008	86.23	88.10
	Post			1	6.66	0.001	13.77	11.90
4 moras, [u]	Pre	4	1		6.33	0.002	58.07	57.70
	Post			1	6.66	0.001	41.93	42.30
5 moras, [u]	Pre	5	1		7.84	0.000	23.45	16.70
	Post			1	6.66	0.001	76.55	83.30
6 moras, [u]	Pre	6	1		9.35	0.000	6.35	8.30
	Post			1	6.66	0.001	86.34	84.18

Based on the constraint violation profiles and the observed percentages of each output form, the optimal weights for the three constraints were found using the Solver function of Excel (Fylstra et al. 1998). This was done by maximizing the log-likelihood of the data with respect to the constraint set. In order words, the optimal weights are those that are most likely to generate this dataset. The Excel sheet used for this analysis, as well as screen recording of this calculation process, are available at the osf repository. The obtained optimum weights appear at the top row of the tableaux.

The values predicted by these optimum weights, given the MaxEnt math and the constraint violation profiles, are shown in the penultimate column. Comparing the last two columns of these tableaux, the match between the observed percentages and predicted percentages generally seems to be very good. To visualize the success of this MaxEnt HG analysis, Figure 4 plots the correlation between these two measures.

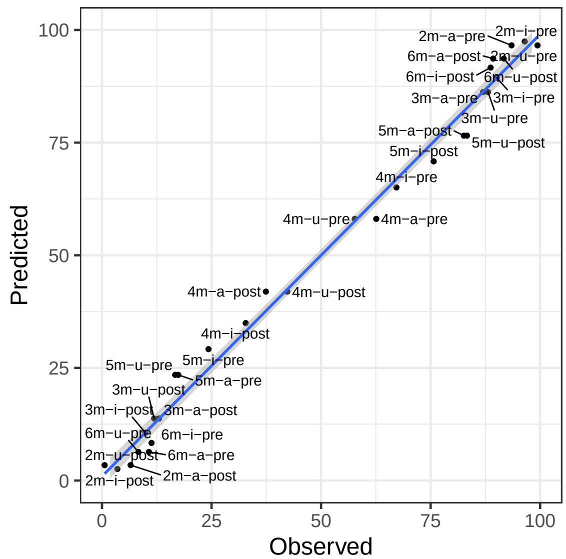


Figure 4: The correlation between the observed percentages in the experiment (the x-axis) and the percentages predicted by the MaxEnt analysis (the y-axis).

6 Conclusion

While MaxEnt HG is in essence a general statistical tool, it has proven to be an extremely useful tool to model various aspects of our linguistic behavior. The current aim of this experiment, building on two previous recent studies (Kawahara 2020c, 2021), was to expand its scope by including sound symbolism as another potential domain for which MaxEnt HG can be a useful analytical tool. The approach we took was heavily inspired by Hayes (2020)—take the mathematical predictions of MaxEnt HG seriously and examine whether they pan out in actual linguistic patterns. We have shown that the quantitative signature of MaxEnt HG, in particular a wug-shaped curve, is observed when Japanese speakers judged the evolution status of non-existing Pokémon names. An increase in mora length results in a sigmoidal curve, and the curves are shifted depending on the vowel quality of the names. An analysis using MaxEnt HG, together with the sorts of constraints that are used in OT research tradition, is shown to be successful in modeling the observed data.

To summarize the key contributions of this paper, first, we replicated the fundamental results of Kawahara (2020c, 2021) that an increase in mora counts results in more post-evolution responses in a sigmoidal fashion, and that this sigmoid curve can be shifted when another factor—e.g. in the current experiment, vowel backness difference—is relevant. This result is a wug-shaped curve, which is a typical probabilistic pattern that MaxEnt HG is predicted to generate, lending further support to the idea that MaxEnt HG is suited to model various aspects of our linguistic behavior (Hayes 2020). Second, as a methodological contribution, we have shown that Bayesian analyses are necessary to access wug-shaped curves, as they allow us to examine the degree of certainty that we can conclude that the different curves have comparable slopes, as predicted by MaxEnt HG. Third, as a case study of sound symbolism, the experiment has shown that it is vowel backness that is relevant in the sound symbolic patterns related to the notion of evolution, which is closely related to size. Finally, by showing that sound symbolic patterns and probabilistic phonological patterns show similar quantitative signatures, we would like to echo the recent claim that exploring sound symbolic patterns can be informative for formal phonological research (Alderete & Kochetov 2017; Kawahara 2020c, 2021; Kawahara & Breiss 2021).

References

Alderete, John & Alexei Kochetov. 2017. Integrating sound symbolism with core grammar: The case of expressive palatalization. *Language* 93. 731–766.

AnderBois, Scott, Adrian Brasoveanu & Robert Henderson. 2012. The pragmatics of quantifier scope: A corpus study. *Proceedings of Sinn und Bedeutung* 16. 15–28.

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

Berlin, Brent. 2006. The first congress of ethonozoological nomenclature. *Journal of Royal Anthropological Institution* 12. 23–44.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32. 45–86.

Boersma, Paul & Joe Pater. 2016. Convergence properties of a Gradual Learning Algorithm for Harmonic Grammar. In John J. McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, 389–434. London: Equinox.

Breiss, Canaan & Bruce Hayes. 2020. Phonological markedness effects in sentential formation. *Language* 96. 338–370.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Royal Netherlands Academy of Science.

Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models using Stan. R package.

Chambers, Chris. 2017. *The 7 deadly sins of psychology*. Princeton: Princeton University Press.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral science*. Lawrence Erlbaum Associates.

Dingemanse, Mark. 2018. Redrawing the margins of language: Lessons from research on ideophones. *Glossa* 3(1). 4, doi:org/10.5334/gjgl.444.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen & Padraic Monaghan. 2015. Arbitrariness, iconicity and systematicity in language. *Trends in Cognitive Sciences* 19(10). 603–615.

Eager, Christopher & Roy Joseph. 2017. Mixed effects models are sometimes terrible. Ms. University of Illinois at Urbana Champaign, `https://arxiv.org/abs/1701.04858`.

Ernestus, Mirjam & Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1). 5–38.

Fylstra, Danial, Leon Lasdon, John Watson & Allan Waren. 1998. Design and use of the Microsoft Excel Solver. *Interfaces* 28(5). 29–55.

Gallistel, Randy C. 2009. The importance of proving the null. *Psychological Review* 116(2). 439–453.

Garellek, Marc, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, Daniel Recasens, Timo B. Roettger, Adrian Simpson & Kristine Yu. 2020. Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Sciences* 9(1).

Gigerenzer, Gerd & Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology* 62. 451–482.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Workshop on Variation within Optimality Theory* 111–120.

Haiman, John (ed.). 1985. *Iconicity in syntax*. Amsterdam: John Benjamins.

Hamano, Shoko. 1998. *The sound-symbolic system of Japanese*. Stanford: CSLI Publications.

Harrison, K. David, Mark Dras & Berk Kapicioglu. 2002. Agent-based modeling of the evolution of vowel harmony. *Proceedings of NELS* 32(2). 217–236.

Hayes, Bruce. 2017. Varieties of noisy harmonic grammar. *Proocedings of Annual Meetings on Phonology* .

Hayes, Bruce. 2020. Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. Ms. UCLA.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39. 379–440.

Hayes, Bruce, Colin Wilson & Anne Shisko. 2012. Maxent grammars for the metrics of Shakespeare and Milton. *Language* 88(4). 691–731.

Hinton, Leane, Johanna Nichols & John Ohala. 1994. *Sound symbolism*. Cambridge: Cambridge University Press.

Hinton, Leane, Johanna Nichols & John Ohala. 2006. *Sound symbolism, 2nd edition*. Cambridge: Cambridge University Press.

Jäger, Gerhad. 2007. Maximum Entropy Models and Stochastic Optimality Theory. In Joan W. Bresnan (ed.), *Architectures, rules, and preferences: Variations on themes*, 467–479. CSLI.

Jespersen, Otto. 1922. Symbolic value of the vowel *i*. In *Linguistica: Selected papers in English, French and German*, vol. 1, 283–303. Copenhagen: Levin and Munksgaard.

Jurafsky, Daniel & James H. Martin. 2019. *Speech and language processing (3rd edition, draft)*. `https://web.stanford.edu/~jurafsky/slp3/`.

Kawahara, Shigeto. 2016. Japanese has syllables: A reply to Labrune (2012). *Phonology* 33. 169–194.

Kawahara, Shigeto. 2020a. Sound symbolism and theoretical phonology. *Language and Linguistic Compass* 14(8). e12372.

Kawahara, Shigeto. 2020b. Teaching and learning guide for "Sound symbolism and theoretical linguistics". *Language and Linguistic Compass* 14(8). e12376.

Kawahara, Shigeto. 2020c. A wug-shaped curve in sound symbolism: The case of Japanese Pokémon names. *Phonology* 37(3). 383–418.

Kawahara, Shigeto. 2021. Testing maxent with sound symbolism: A stripy wug-shaped curve in japanese pokémon names. *Language (Research Report)* .

Kawahara, Shigeto & Canaan Breiss. 2021. Exploring the nature of cumulativity in sound symbolism: Experimental studies of Pokémonastics with English speakers. *Laboratory Phonology* 12(1).

Kawahara, Shigeto, Mahayana C. Godoy & Gakuji Kumagai. 2021. English speakers can infer Pokémon types based on sound symbolism. *Frontiers in Psychology* .

Kawahara, Shigeto, Miwa Isobe, Yukino Kobayashi, Tomoko Monou & Reiko Okabe. 2018a. Acquisition of sound symbolic values of vowels and voiced obstruents by Japanese children: Using a Pokémonastics paradigm. *Journal of the Phonetic Society of Japan* 22(2). 122–130.

Kawahara, Shigeto, Hironori Katsuda & Gakuji Kumagai. 2019. Accounting for the stochastic nature of sound symbolism using Maximum Entropy model. *Open Linguistics* 5. 109–120.

Kawahara, Shigeto & Jeff Moore. 2021. How to express evolution in English Pokémon names. *Linguistics* 59(3). 577–607.

Kawahara, Shigeto, Atsushi Noto & Gakuji Kumagai. 2018b. Sound symbolic patterns in Pokémon names. *Phonetica* 75(3). 219–244.

Kingston, John & Randy Diehl. 1994. Phonetic knowledge. *Language* 70. 419–454.

Kluender, Keith, Randy Diehl & Beverly Wright. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16. 153–169.

Knoeferle, Klemens, Jixing Li, Emanuela Maggioni & Charles Spence. 2017. What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports* 7.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(2). 199–244.

Kruschke, John K. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, John K. & Torrin M. Liddell. 2018. The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychological Bulletin and Review* 25. 178–206.

Kumagai, Gakuji. 2019. A sound-symbolic alternation to express cuteness and the orthographic Lyman's Law in Japanese. *Journal of Japanese Linguistics* 35(1). 39–74.

Kumagai, Gakuji & Shigeto Kawahara. 2019. Effects of vowels and voiced obstruents on Pokémon names: Experimental and theoretical approaches [in Japanese]. *Journal of the Linguistic Society of Japan* 155. 65–99.

Lefkowitz, Lee Michael. 2005. *Maxent Harmonic Grammars and Phonetic Duration*: UCLA Doctoral dissertation.

Linzen, Tal & Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Reviews of Linguistics* .

Makowski, Dominique, Mattan S. Ben-Shachar, Annabel S.H. Chen & Daniel Lüdecke. 2019. Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology* https://doi.org/10.3389/fpsyg.2019.02767.

Makowski, Dominique, Daniel Lüdecke, Mattan S. Ben-Shachar, Michael D. Wilson, Paul-Christian Bürkner, Tristan Mahr, Henrik Singmann, Quentine F. Gronau & Sam Crawley. 2020. `bayestestR`. R package.

Manning, Christopher D., Kevin Clark & John Hewitt. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of National Academy of Sciences* 117(48).

McCarthy, John J. 2002. *A thematic guide to Optimality Theory*. Cambridge: Cambridge Univer-

sity Press.
McCarthy, John J. 2003. OT constraints are categorical. *Phonology* 20(1). 75–138.
McCarthy, John J. 2008. *Doing Optimality Theory*. Oxford: Blackwell-Wiley.
McPherson, Laura & Bruce Hayes. 2016. Relating application frequency to morphological structure: The case of Tommo So vowel harmony. *Phonology* 33. 125–167.
Newman, Stanley. 1933. Further experiments on phonetic symbolism. *American Journal of Psychology* 45. 53–75.
Nielsen, Alan K. S. & Mark Dingemanse. 2020. Iconicity in word learning and beyond: A critical review. *Language and Speech* .
Nielsen, Alan K. S. & Drew Rendall. 2011. The sound of round: Evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology* 65. 115–124.
Nielsen, Alan K. S. & Drew Rendall. 2013. Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology* 67(2). 153–63.
Ohala, John. 1994. The frequency code underlies the sound symbolic use of voice pitch. In Leane Hinton, Johanna Nichols & John Ohala (eds.), *Sound symbolism*, 325–347. Cambridge: Cambridge University Press.
Porte, Graeme (ed.). 2012. *Replication research in applied linguistics*. Cambridge: Cambridge University Press.
Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Malden and Oxford: Blackwell.
R Development Core Team. 1993–. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria.
Roettger, Timo B. & Dinah Baer-Henney. 2019. Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis* 4(1).
Rousseau, Pascale & David Sankoff. 1978. Advances in variable rule methodology. In David Sankoff (ed.), *Linguistic variation: Models and methods*, 57–69. New York: Academic Press.
Sapir, Edward. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology* 12. 225–239.
Shih, Stephanie S. 2017. Constraint conjunction in weighted probabilistic grammar. *Phonology* 34(2). 243–268.
Shih, Stephanie S, Jordan Ackerman, Noah Hermalin, Sharon Inkelas, Hayeun Jang, Jessica Johnson, Darya Kavitskaya, Shigeto Kawahara, Miran Oh, Rebecca L Starr & Alan Yu. 2019. Cross-linguistic and language-specific sound symbolism: Pokémonastics. Ms. University of Southern California, University of California, Merced, University of California, Berkeley, Keio University, National University of Singapore and University of Chicago.
Shinohara, Kazuko & Shigeto Kawahara. 2016. A cross-linguistic study of sound symbolism: The images of size. In *Proceedings of the Thirty Sixth Annual Meeting of the Berkeley Linguistics Society.*, 396–410. Berkeley: Berkeley Linguistics Society.
Stevens, Kenneth & Sheila Blumstein. 1981. The search for invariant acoustic correlates of phonetic features. In Peter Eimas & Joanne D. Miller (eds.), *Perspectives on the study of speech*, 1–38. New Jersey: Earlbaum.
Tesar, Bruce. 2007. A comparison of lexicographic and linear numeric optimization using violation difference ratios. Ms. Rutgers University.

Ultan, Russell. 1978. Size-sound symbolism. In Joseph Greenberg (ed.), *Universals of human language II: Phonology*, 525–568. Stanford: Stanford University Press.

Westbury, Chris, Geoff Hollis, David M. Sidhu & Penny M. Pexman. 2018. Weighting up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language* 99. 122–150.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Winter, Bodo. 2019. *Statistics for linguists*. New York: Taylor & Francis Ltd.

Zimmermann, Richard. 2017. *Formal and quantitative approaches to the study of syntactic change: Three case studies from the history of English*: University of Geneva Doctoral dissertation.

Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: An argument for Harmonic Grammar. *Language* 93. 497–548.