A stripy wug-shaped curve in sound symbolism

Abstract

In recent years, Maximum Entropy Harmonic Grammar (MaxEnt HG) has been successfully applied to model various linguistic patterns. Building on these studies, the current experiment examines its quantitative signatures, typical probabilistic patterns that this framework is predicted to generate (Hayes 2020). Given a linear scale of violations of one constraint, MaxEnt yields a sigmoid curve. When another constraint is relevant, this sigmoid curve can be shifted, yielding two sigmoid curves, which together look like a wug, our beloved animal in the linguistics community. When the second constraint can be violated more than once, MaxEnt predicts that it yields a set of multiple sigmoid curves separated from one another, resulting in a "stripy wug-shaped curve." The current experiment demonstrates that we observe a stripy wug-shaped curve in patterns of sound symbolism, systematic associations between sounds and meanings. Concretely, the experiment shows that the judgment of Pokémons' evolution status is cumulatively affected by the mora counts of nonce names, resulting in a sigmoid curve, and that this sigmoid curve can be shifted according to the number of voiced obstruents contained in the names. This paper concludes that MaxEnt HG is a useful framework to model sound-meaning mappings, and suggests that there may be a meaningful parallel between sound symbolic patterns and probabilistic phonological patterns.

1 Introduction

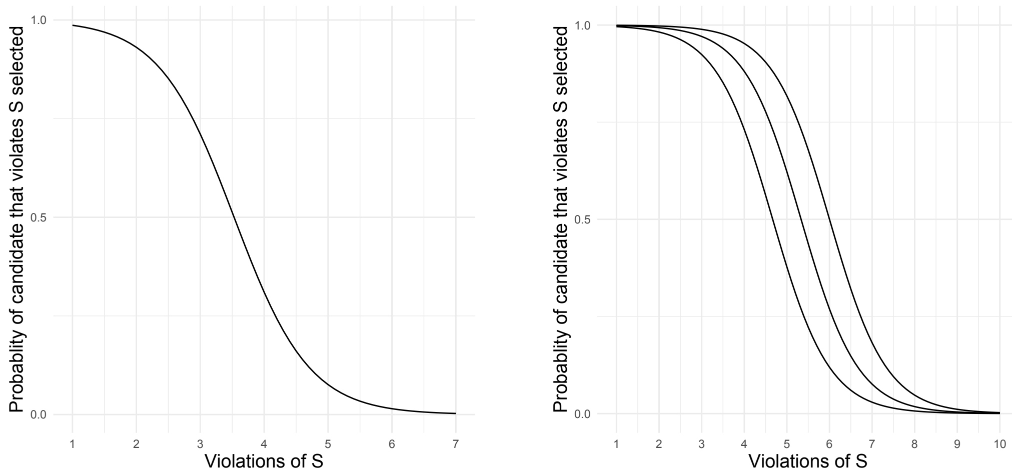
Traditional generative analyses tended to focus on the dichotomous distinction between grammatical and ungrammatical forms (Chomsky 1957; Chomsky & Halle 1968). However, the field of modern linguistic theories has witnessed a rise of interests in probabilistic generalizations (e.g. Pierrehumbert 2020 for a recent review). One framework that has been used to model a wide range of probabilistic generalizations in phonology and other linguistic patterns is Maximum Entropy Harmonic Grammar (MaxEnt HG) (e.g. Goldwater & Johnson 2003; Smolensky 1986). In the context of current phonological theorization, there are two ways to understand MaxEnt HG. One is to consider it as application of logisitc regression models for linguistic analyses (Jurafsky & Martin 2019). The other is to consider it as a stochastic version of Optimality Theory (OT) (Prince

& Smolensky 1993/2004). MaxEnt HG, just like OT, consists of inputs and outputs as well as CON, the set of constraints that regulate the mapping between these two levels of representations. Unlike OT, however, in MaxEnt HG, the constraints are weighted rather than ranked, and MaxEnt assigns a probability distribution over a set of output candidates, rather than deterministically choosing one output candidate as a winner, as OT does.

In order to evaluate the validity of MaxEnt as a grammatical model, Hayes (2020) highlights very specific predictions that MaxEnt makes in terms of the typical probabilistic patterns that it generates, which he refers to as "its quantitive signatures." To illustrate, suppose that there is a scaler constraint S, whose violation can be assessed on a linear numerical scale. Further suppose that there is a binary constraint, B, whose constraint violation directly conflicts with that of S. When we plot the number of violations of the constraint S on the x-axis and the probability of the candidate that violates S being selected as a winner on the y-axis, it results in a sigmoid curve, as shown in Figure 1(a). The linear violation scale is converted to a sigmoidal curve in MaxEnt, because it involves a logistic transformation ($\frac{1}{1+e^{-N}}$) in calculating the probability distribution of output candidates.

When another constraint—call it P for "Perturber"—is at play, this sigmoid curve can be shifted on the x-axis, yielding another sigmoid curve, which results in what Hayes (2020) refers to as a "wug-shaped curve." When this third constraint is violated twice, it yields yet another sigmoid curve. Together, this whole scenario results in a "stripy wug-shaped curve," which is schematically shown in Figure 1(b). Hayes (2020) argues that a stripy wug-shaped curve is commonly observed in probabilistic phonological alternation patterns (McPherson & Hayes 2016; Zuraw & Hayes 2017), as well as in categorical perception of speech (Liberman et al. 1957) and diachronic changes in syntax (Kroch 1989; Zimmermann 2017). He further demonstrates that these wug-shaped curves are difficult to derive in another widely-used stochastic model of phonology, namely, Stochastic OT (Boersma 1998; Boersma & Hayes 2001). Hayes (2020) thus concludes that to the degree that wug-shaped curves are omnipresent in various linguistic patterns, support is provided for MaxEnt being a useful tool to model linguistic knowledge that lies behind these probabilistic patterns.[1]

[1]Noisy Harmonic Grammar (Noisy HG) can derive (stripy) wug-shaped curves as well, given certain assumptions about how noise is added to the calculation of overall harmony (Hayes 2017). Noisy HG, therefore, can be a viable alternative to MaxEnt. Following Hayes (2017, 2020), I will not attempt to tease apart these two theories, because the difference in quantitative predictions that these two theories make can be very subtle. The crucial distinction that the presence of wug-shaped curves bears upon is the one between those grammatical models with weighted constraints (Noisy HG and MaxEnt HG) and those grammatical models with ranked constraints (Stochastic OT).



(a)

(b)

Figure 1: (a) A sigmoid curve generated by MaxEnt. The function is defined as $f(x) = \frac{1}{1+e^{-N}}$, where N is linearly correlated with x, the number of violation marks assigned by S. (b) Multiple sigmoid curves, resulting in a stripy wug-shaped curve.

The current paper demonstrates that this quantitive prediction of MaxEnt holds in a particularly clear manner in the domain of sound symbolism, systematic associations between sounds and meanings (Hinton et al. 1994). The current experiment builds upon an earlier experiment reported in Kawahara (2020b), who manipulated the mora counts and the presence of a voiced obstruent in nonce names. Kawahara (2020b) asked native speakers of Japanese whether each name is better suited for a pre-evolution Pokémon character or a post-evolution character, the latter of which is generally heavier and larger. The results of this experiment showed that varying the mora count of nonce names increases the post-evolution responses in a sigmoidal manner. The experiment also showed that the presence of a name-initial voiced obstruent horizontally shifts the entire sigmoid curve, resulting in a wug-shaped curve. Expanding upon Kawahara (2020b), the current experiment varies the number of voiced obstruents from 0 to 2, which is fully crossed with mora count differences, in order to examine whether this new manipulation results in a stripy wug-shaped curve, illustrated in Figure 1(b).

To preview the results, this manipulation indeed results in what looks to be three separate sigmoid curves, yielding a stripy wug-shaped curve. The Bayesian modeling analysis shows that we can conclude, with a reasonable amount of confidence, that the three sigmoid curves are identical and separated from one another (§3 and §4). A MaxEnt analysis shows that the experimental results are indeed nicely modeled by MaxEnt HG, when we posit that MaxEnt mediates the mapping from sound to meaning, with the sort of constraints that are used in the OT research tradition (§5). Overall, this paper provides further support to the recent proposal that MaxEnt HG is a useful tool to model sound symbolic patterns (Kawahara et al. 2019; Kawahara 2020b). Some intriguing parallels between sound symbolic patterns and probabilistic phonological patterns are discussed at the end of the paper (§6).

2 Methods

2.1 Background

The current experiment is a case study of Pokémonastics, a general research paradigm in which researchers explore the nature of sound symbolic patterns in natural languages using Pokémon names. I refer the readers to Shih et al. (2019) for several research advantages of this research program; here it suffices to note that Pokémon characters can undergo evolution, and when they do so, they are assigned a different name. The first Pokémonastics study, which analyzed the existing Pokémon names in Japanese (Kawahara et al. 2018), pointed out that the names of evolved characters tend to be longer, and are more likely to contain voiced obstruents; for example, *Anopusu* evolves into *Aamarudo*, the latter of which is longer (4 moras vs. 5 moras) and contains a voiced obstruent, [d]. The first principle is arguably an instance of what has been known as "the iconicity of quantity" in the literature on sound symbolism, in which larger words tend to denote larger quantity (Dingemanse et al. 2015; Haiman 1980; see Marks 1978 for its general cognitive basis). The second observation is perhaps rooted in the well-known observation in Japanese that voiced obstruents denote larger quantities (Hamano 1998), which itself may be grounded in the expansion of the oral cavity which occurs during the production of voiced obstruents (Ohala 1983). Several experimental studies using nonce names have confirmed the productivity of these two sound symbolic patterns in the Japanese Pokémon universe (Kawahara 2020b; Kawahara & Kumagai 2019, 2021). The current experiment makes use of these sound symbolic patterns in order to address the specific prediction MaxEnt makes which was illustrated in Figure 1(b).

2.2 Stimuli

The current experiment heavily draws upon Kawahara (2020b), who manipulated the mora counts and the presence of a word-initial voiced obstruent in nonce names. Table 1 lists the stimuli of the current experiment, in which dots represent mora boundaries. The experiment manipulated both the number of moras and the number of voiced obstruents. Mora counts were varied from 2 to 6, which each corresponds to the minimum and maximum lengths that are allowed in the real Japanese Pokémon names. Since the stimuli contained no heavy syllables, the mora boundaries and syllable boundaries always coincided with each other in the current stimulus set. This is one aspect in which the current study improves upon Kawahara (2020b), whose stimuli included diphthongs (e.g. [d**oi**wanu]), failing to control for the syllable counts.

In the current stimulus set, the mora count manipulation was fully crossed with another factor, the number of voiced obstruents, varying from 0 to 2. Voiced obstruents, when they were present, were placed either word-initially or in the first two syllables. The positions of voiced obstruents were therefore consistent across all the mora count conditions.

Five items were included in each cell, resulting in a total of 75 items (5 mora conditions $\times$ 3 voicing conditions $\times$ 5 items). All the stimulus names were created using an online nonce name generator, which combines Japanese moras randomly to create new names.[2] Since [p] is known to have its own salient sound symbolic values, such as cuteness (Kumagai 2019) and smallness (Shih et al. 2019), this segment was not used in the current stimuli. Another potential factor is vowel quality, which was not controlled in the experiment for two reasons. One is that Kawahara et al. (2018) did not find a substantial impact on vowel quality in the existing Pokémon names in Japanese; the other is that an attempt to control for vowel quality, for example by using the same vowel in the whole names, resulted in very artificial names, especially in long names.

2.3 Procedure

The experiment was distributed online using SurveyMonkey. The first page of the experiment presented a consent form, which has been approved by the author's institute. The instruction of the experiment stated that the participation in this experiment is voluntary, but requires some basic familiarity with Pokémon. The participants were reminded that in the Pokémon universe, there are pre-evolution characters and post-evolution characters, and that post-evolution characters tend to be larger and stronger.

Within each trial, the participants were provided with one nonce name and were asked to judge whether each name is better suited for a pre-evolution character or a post-evolution character. The order of the stimuli was uniquely randomized for each participant. They were asked to make their judgment based on their intuitions rather than thinking about "right" or "wrong" answers. The stimuli were presented in the Japanese *katakana* orthography, although they were asked to read each name in their head before answering the questions.

2.4 Participants

The call for participation was primarily advertised on Twitter. A total of 144 native speakers of Japanese completed the online experiment. Six participants reported that they had studied sound

[2] `http://sei-street.sakura.ne.jp/page/doujin/site/doc/tool_genKanaName.html` (last access, September 2020). The items in the first two columns in Table 1 were largely adapted from Kawahara (2020b), who used the same online name generator. Some items, which involved diphthongs (i.e. vowel sequences with falling sonority: Kubozono 2015), were "hand-corrected" by inserting a consonant, in order to control for the number of syllables in addition to the number of moras.

Table 1: The list of the stimuli used in the experiment. Dots represent mora boundaries, which coincide with syllable boundaries.

	0 vcd obs	1 vcd obs	2 vcd obs
2 moras	[su.tsu]	[za.mu]	[bu.zu]
	[no.çi]	[gu.ka]	[ʑi.da]
	[jo.ni]	[gi.ke]	[da.za]
	[ho.mu]	[ba.ɾu]	[ge.bu]
	[ni.mi]	[go.ɸu]	[go.de]
3 moras	[ku.ɕi.me]	[bu.ɾo.se]	[da.bu.so]
	[jo.ɾu.so]	[go.se.he]	[do.da.no]
	[se.sa.ɾi]	[bo.ma.sa]	[ga.da.to]
	[mu.su.ha]	[gu.ne.ju]	[bu.ge.ɾu]
	[ɾi.to.no]	[da.su.ɾo]	[ʑi.de.mi]
4 moras	[ku.ki.me.se]	[bi.to.ɾe.ni]	[ba.de.ju.ɸu]
	[so.ha.ko.ni]	[za.ni.te.ja]	[bu.ga.so.ja]
	[ɾa.ɕi.no.ɾo]	[ga.çi.ke.ɾo]	[ze.ga.ki.ɸu]
	[ko.te.nu.ne]	[da.ka.ni.ɾi]	[ga.de.ha.wa]
	[a.mo.çi.ni]	[do.ki.ɾa.nu]	[gi.do.ke.he]
5 moras	[ha.ku.te.çi.no]	[bi.so.ɸu.sa.ta]	[gi.ze.mi.ke.me]
	[ɾo.ta.ɾa.na.to]	[da.ɾa.su.to.ki]	[ba.go.ki.ɾu.ke]
	[so.ka.ne.ni.ɾe]	[de.mu.sa.te.he]	[de.gu.mu.ɾa.tsu]
	[ɾu.ɾi.ha.me.ke]	[gi.a.so.ta.e]	[do.gu.ha.ɾa.mu]
	[sa.na.çi.ta.ni]	[de.nu.ɾa.so.me]	[bo.ga.to.he.ɾa]
6 moras	[ju.ɾo.ka.mu.mo.ja]	[gu.se.ɸu.çi.ɾa.mo]	[bo.da.ɾo.ɸu.so.ɸu]
	[mu.ku.ho.ɾo.ho.te]	[go.na.ɸu.to.ko.so]	[zu.ga.çi.ne.te.so]
	[ɾa.ha.ɾi.tɕi.ɾu.tsu]	[do.ja.to.sa.mi.ta]	[da.ga.su.me.ta.ɾa]
	[ne.nu.he.mo.sa.nu]	[da.na.ɾi.no.mi.ki]	[be.ga.he.ɾa.ka.ɾo]
	[ɾu.no.nu.ɾo.te.tɕi]	[zo.te.he.so.ju.ɾa]	[gi.go.na.ke.to.sa]

symbolism. Four participants reported that they had participated in another Pokémonastics experiment. The data from these participants were excluded. The data from the remaining 134 participants entered into the subsequent statistical analysis.

2.5 Statistical analyses

The result of the experiment was statistically assessed with a Bayesian mixed effects logistic regression model, using the `brms` R package (Bürkner 2017). The dependent variable was the binary response obtained in the experiment (0 = pre-evolution, 1 = post-evolution). The predictor variables were mora counts and the number of voiced obstruents, both of which were numerical, and hence centered (Winter 2019). The interaction between the two fixed factors was included, for reasons detailed below. Random factors included free-varying random intercepts for participant and item, as well as random slopes for both of the fixed effects as well as their interaction by participants. The random slopes and intercepts were uncorrelated, because there were no theoretical reasons to expect that there should be a correlation. Four chains of 2,000 iterations were run, and the last 1,000 samples from each chain were analyzed. The default, weakly-informative priors were used. The highest $\hat{R}$ value was 1.01, indicating that the chains mixed successfully. The experimental data file and the R markdown syntax are available as Supplementary Material A.

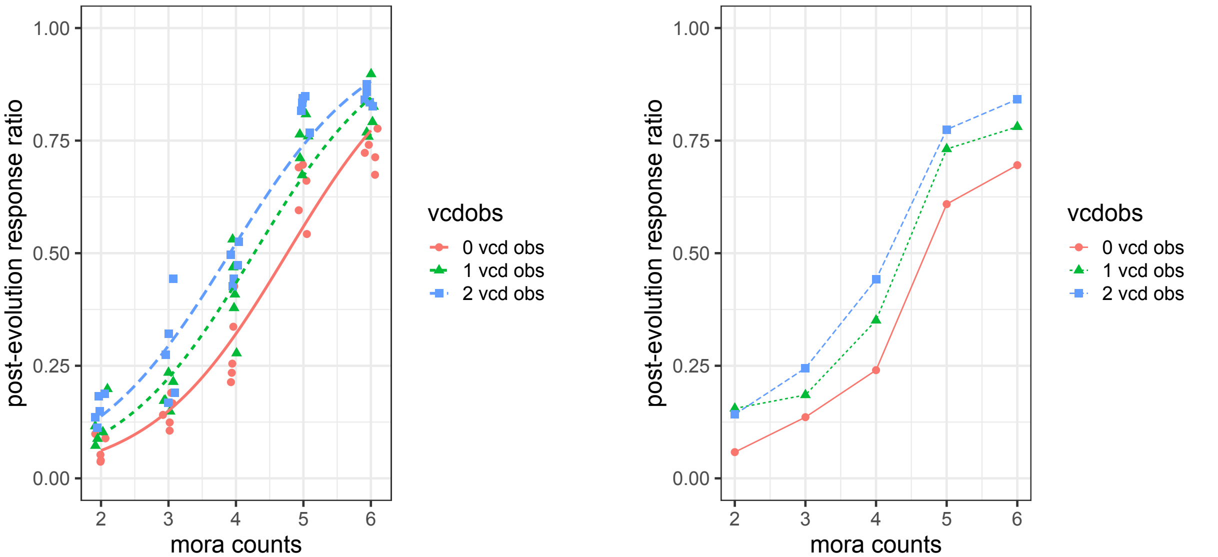
Bayesian models yield a posterior distribution of possible values for each parameter, which can be interpreted by examining the middle 95% of these values, called the 95% Credible Interval (abbreviated as "95% CI"). Since the current model uses logistic regression, a positive slope coefficient (β) indicates that that factor increases post-evolution Pokémon responses.

Taking a Bayesian approach has two notable advantages for the current experiment. First, this method makes it possible to fit the complex model with an interaction term together with a complex random effect structure without convergence issues. Second, perhaps more importantly, the Bayesian approach allows us to gather evidence for the null results, rather than merely failing to reject the null hypothesis (Gallistel 2009). Examining whether the interaction term plays a meaningful role or not is particularly important in the current experiment for the following reason. A stripy wug-shaped curve consists of three *identical* wug-shaped curves, which means that the interaction between the mora count and the voiced obstruent *should not* play a substantial role in the model, since the interaction term functions as a slope adjustment term (Winter 2019). If the interaction term turns out to be a meaningful predictor in the model, it means that the slopes differ between the multiple curves.[3]

3 Results

Figure 2(a) plots "post-evolution response ratios" for each item, averaged over all the participants. Red circles represent items which contain no voiced obstruents, green triangles show the results of the items which contain one voiced obstruent, and blue rectangles show the results for the condition which contains two voiced obstruents. The `ggplot2` package (Wickham 2016) was used to superimpose a logistic curve for each voicing condition. The three logistic curves appear to be well separated from one another, just like the schematic stripy wug-shaped pattern shown in Figure 1(b).

[3]This was one clear limitation of Kawahara (2020b), who was unable to address this question, because the analysis deployed by Kawahara (2020b) was a frequentist analysis—the best that could be concluded based on that analysis was that the interaction term was non-significant. The lack of significance, however, only means that we cannot reject the null hypothesis that the slopes were not identical; i.e. it does not provide evidence that the slopes between the curves are likely to be identical. The Bayesian analysis presented in the current paper overcomes this limitation.



(a) (b)

Figure 2: (a) The by-participant averages for each item. To avoid overlap, the points were horizontally jittered by 0.1. Logistic curves are superimposed using `ggplot2` for each voicing condition. (b) The line-plots with grand averages for each voicing condition.

Figure 2(b) illustrates the overall results by presenting grand averages for each condition. This analysis does not presuppose that sigmoid curves would fit the observed data points well. Despite this, the general pattern looks like a stripy wug-shaped curve, consisting of three separate sigmoid curves. For each curve, the slope is evidently steepest between the 3-mora condition and 5-mora condition; on the other hand, the change from 2-mora to 3-mora or the change from the 5-mora to 6-mora do not seem to impact the judgment as much. These are characteristics of sigmoid curves, as observed in Figure 1(a). The slogan that Hayes (2020) uses to describe this observation is “certainty is evidentially expensive” (p.3). It requires very convincing evidence to be certain that a particular name is for a pre-evolution character or for a post-evolution character. A logistic transformation, which underlies MaxEnt, serves to model this observation.

The model summary of the Bayesian mixed effects logistic regression analysis appears in Table 2. The intercept is negative; since both mora counts and the effects of voiced obstruents were centered, this negative intercept means that for names that are 4-mora long containing one voiced obstruent, the post-evolution responses are lower than 50%. The slope of mora count was positive and its 95% CI does not include 0, which means that increase in mora counts meaningfully increased the probability of the names assigned to a post-evolution character. The slope of voiced obstruents was also positive, and its 95% CI does not include 0. These results suggest that voiced obstruents meaningfully increased post-evolution responses as well. The 95% CI for the interaction term on the other hand includes 0, which means that the effects of mora counts and those of voiced obstruents were independent of one another.

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

	β	error	95% CI
intercept	-0.42	0.07	[-0.56, -0.28]
mora count	1.10	0.09	[0.92, 1.29]
vcd obs	0.45	0.07	[0.30, 0.59]
mora count $\times$ vcd obs	0.03	0.04	[-0.05, 0.12]

4 Discussion

4.1 A stripy wug-shaped curve

I would like to first ponder upon whether the current experimental results shown in Figure 2 should indeed be considered to instantiate a stripy wug-shaped curve. The stripy wug-shaped curve—as discussed by Hayes (2020) and also as predicted by MaxEnt—has three defining features: (1) it consists of multiple sigmoid curves, (2) they are separated from one another, and (3) slopes of the sigmoid curves are identical.[4]

The third requirement is satisfied by the current result, because the interaction term did not meaningfully impact the post-evolution responses. Recall that the Bayesian analysis allows us to gather evidence for the null effects, and therefore we can be reasonably confident that the three sigmoidal curves have the same slope.

The second requirement is satisfied by the current data, since the 95% CI for the effect of voicing does not contain 0; changes in the number of voiced obstruent seem to have meaningfully impacted the post-evolution responses. One may be concerned that the Bayesian analysis reported in §3 alone may not show that *all* three curves are separated from one another, because the voicing condition was coded as a numerical variable. An additional analysis addressing this concern is offered in the Appendix.

The first feature—whether the data would be best fit with a sigmoidal curve—is the most challenging aspect to defend. Linguistic data obtained in an experiment always involve some natural variability, and they therefore never perfectly fit the mathematical definition of sigmoids.

[4] MaxEnt can generate a set of sigmoid curves whose slopes differ from each other, as long as it admits a meaningful interaction term. Whether MaxEnt offers a suitable grammatical framework is one question, but whether it should allow for meaningful interaction terms maybe a different, albeit related, question. In the context of linguistic analyses, the second question can be restated as a question regarding whether we should allow for a locally conjoined constraint in a grammatical model with weighted constraints (Shih 2017). See §5 for further discussion on this point. For now, I assume that different sigmoid curves should have the same (or comparable) slope in (stripy) wug-shaped curves.

Moreover, there are many mathematical functions that can be potentially fit to the data. We can go so far as to fit a mathematical function which intersects every observed data point, but such a function with high mathematical complexity would suffer from the general problem of overfitting (Good & Hardin 2006). For the current experiment, I 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.

An obvious alternative candidate is a linear function, which is not suitable to model the current results for two reasons. First, since we are dealing with the probability space, the y-axis needs to be bound between 0 and 1, but there is nothing in a linear function that guarantees that this restriction is met (Jaeger 2008). Second, a linear function does not capture the observation that the slope in the middle range is steeper than the slopes in the low and high ends of the x-axis continuum.

It is not possible to examine every possible mathematical function here. However, generally speaking, which mathematical function can and should be used to model linguistic data is a topic that should be explored by cross-linguistic considerations. The current state of the field is that we can be reasonably confident that sigmoids, generated by MaxEnt, are suited to model various linguistic patterns (Breiss 2020; Hayes 2020; Hayes & Wilson 2008; McPherson & Hayes 2016; Zuraw & Hayes 2017). See §5 for a MaxEnt analysis of the current experimental results.

4.2 Cumulativity

In addition to showing that we observe a stripy wug-shaped pattern in sound symbolic patterns, the current experiment also bears on a different—albeit related—issue that is of theoretical interest for some contemporary phonological/linguistic theories; namely, cumulativity. Whether linguistic patterns show cumulative properties or not is actively debated in the literature, because it may bear on the question of whether a linguistic optimization system should be based on lexicographic ranking (as in OT) or numeric weightings (as in HG) (e.g. Breiss 2020; McPherson & Hayes 2016; Prince & Smolensky 1993/2004; Zuraw & Hayes 2017; Tesar 2007). The latter framework predicts that cumulativity is the norm, whereas the former framework explicitly prohibits it. The current experiment shows that sound symbolic effects are generally cumulative, just like previous experimental studies on Pokémon names (Kawahara 2020b; Kawahara & Kumagai 2021; Kawahara & Breiss 2020).

More specifically, Japanese speakers are sensitive to each step in the mora count scale. This effect of mora count instantiates a case of counting cumulativity (Jäger 2007; Jäger & Rosenbach 2006) in sound symbolism, in that the effects of multiple tokens of the same structure (=mora) additively contributes to the judgment of evolvedness. The number of voiced obstruents increased the post-evolution responses as well, which is also a case of counting cumulativity. Moreover, the effects of mora counts and those of voiced obstruents added up, which instantiates ganging-up cumulativity, in which the effects of two different factors add up (Jäger 2007; Jäger & Rosenbach 2006). Overall, the current experiment shows that both counting cumulativity and ganging-up cumulativity can coexist within a single sound symbolic system (Kawahara 2020b; Kawahara & Breiss 2020), drawing a parallel to a recent observation in probabilistic phonological patterns which suggest the co-existence of counting and ganging-up cumulativity (Breiss 2020; McPherson & Hayes 2016; Zuraw & Hayes 2017).

5 A MaxEnt analysis

This section develops a MaxEnt analysis of the experimental results, using the sort of constraints that have been used in the Optimality Theory research (Prince & Smolensky 1993/2004). Since there are already a number of papers that explain how MaxEnt works for linguistic analyses (e.g. Breiss & Hayes 2020; Hayes & Wilson 2008; Kawahara 2020b; McPherson & Hayes 2016; Zuraw & Hayes 2017), I only provide a brief explanation in this paper. Just like OT, output candidates are evaluated against the set of constraints, each of which is assigned a particular weight. Each candidate receives a harmony score (H), which is the weighted sum of constraint violations: $H = \sum w_i C_i(x)$, where w represents the weight of the i-th constraint and $C_i(x)$ represents how many times a candidate violates the i-th constraint. The harmony score is negatively exponentiated (e^{-H}), which is proportional to the probability of each candidate. The more constraint violations a particular candidate incurs, the higher the harmony score H, the lower e^{-H}, and hence the lower the probability of that candidate. The e^{-H} values for all the candidates are summed into Z; i.e. $Z = \sum (e^{-H})_j$. The predicted probability of each candidate $x_j, p(x_j)$, is $\frac{e^{-H(x_j)}}{Z}$.

Before we proceed, I note at this point that MaxEnt is mathematically equivalent to (multinomial) logistic regression (Jurafsky & Martin 2019), and therefore there is some conceptual overlap between the statistical analysis presented in §3 and the MaxEnt analysis presented in this section. However, the former was used to explore what we can conclude based on the experimental results; on the other hand, the MaxEnt analysis presented in this section is a generative phonological analysis that is meant to model the knowledge that lies behind the patterns observed in the experiment (see Breiss & Hayes 2020 for relevant discussion). To highlight the fact that the analysis in this section builds on the generative phonological research tradition, the set of constraints used below are those that are formulated with a constraint schema proposed by McCarthy (2003).

The general idea that lies behind the analysis developed below is that we can understand sound symbolism—mappings from sounds to meanings—just like phonological input-output mappings, which is evaluated by a set of constraints that are familiar from traditional phonological research (Kawahara et al. 2019; Kawahara 2020b). The set of constraints that were proposed by Kawahara

(2020b) can actually be directly applied to model the current results, which are shown in (1).[5]

(1) Constraints deployed in the current analysis, adapted from Kawahara (2020b)

a. *LONGPRE: Assign a violation mark for each mora in a pre-evolution character name.

b. *VCDPRE: Assign a violation mark for each voiced obstruent in a pre-evolution character name.

c. *POST: Assign a violation mark for each post-evolution name.

The first constraint is a formal expression of "the iconicity of quantity" (Haiman 1980), which prefers long names to be used for post-evolution characters. This constraint corresponds to the scalar constraint S that was used to schematically illustrate a stripy wug-shaped curve in Figure1(b). The second constraint prefers that names with voiced obstruents be used for post-evolution character names, and this corresponds to the perturber constraint P that was used in Figure1(b). The last constraint penalizes post-evolution character names in general, which corresponds to the binary constraint B. This constraint serves as a "baseline" constraint, determining the general preference for pre-evolution characters.

The MaxEnt tableaux for all the conditions are shown in (2). The leftmost column shows the input forms (i.e. the phonological forms), and the second column shows the outputs (i.e. the two semantic meanings, pre-evolution character names vs. post-evolution character names). The constraint violation profiles are shown in the next three columns. The observed percentages of each condition are shown in the rightmost column, which were taken from the grand averages obtained in the experiment.

Based on the constraint profiles and the observed percentages of each output form, the optimal weights of the three constraints were calculated using the Solver function of Excel. The weights were not allowed to be negative or higher than 50. The weights that were obtained by this analysis are shown at the top row of the tableaux, from which we can calculate the predicted values using the procedure that is outlined at the beginning of this section. The values that are predicted by the MaxEnt analysis are shown in the penultimate column. The Excel sheet used to calculate the optimal weights and the predicted values is available as Supplementary Material B.

[5] One may object that constraints used in OT/HG-analyses should not refer to arguably culture-specific notions such as "evolution." An alternative formulation is to replace the notion of evolution with size, because size is a semantic dimension that is signaled by sound symbolism in various languages (see e.g. Sidhu & Pexman 2018). For the sake of expositional clarity, however, I continue using constraints which refer to evolution status.

(2) The MaxEnt Tableaux

		w = 0.93	w = 0.40	w = 4.55				
Input	Output	*LongPre	*VcdPre	*Post	Harmony (H)	e^{-H}	Predicted	Observed
2 moras, vls	Pre	2			1.86	0.155	93.65	94.18
	Post			1	4.55	0.011	6.35	5.82
3 moras, vls	Pre	3			2.79	0.061	85.32	86.42
	Post			1	4.55	0.011	14.68	13.58
4 moras, vls	Pre	4			3.72	0.024	69.60	75.97
	Post			1	4.55	0.011	30.40	24.03
5 moras, vls	Pre	5			4.66	0.010	47.44	39.10
	Post			1	4.55	0.011	52.56	60.90
6 moras, vls	Pre	6			5.59	0.0037	26.24	30.45
	Post			1	4.55	0.011	73.76	69.55
2 moras, 1 vcd	Pre	2	1		2.27	0.104	90.77	84.48
	Post			1	4.55	0.011	9.23	15.52
3 moras, 1 vcd	Pre	3	1		3.20	0.041	79.49	81.49
	Post			1	4.55	0.011	20.51	18.51
4 moras, 1 vcd	Pre	4	1		4.13	0.016	60.43	64.93
	Post			1	4.55	0.011	39.57	35.07
5 moras, 1 vcd	Pre	5	1		5.06	0.006	37.58	26.87
	Post			1	4.55	0.011	62.42	73.13
6 moras, 1 vcd	Pre	6	1		5.99	0.0025	19.18	21.94
	Post			1	4.55	0.011	80.82	78.06
2 moras, 2 vcd	Pre	2	2		2.67	0.069	86.77	85.82
	Post			1	4.55	0.011	13.23	14.18
3 moras, 2 vcd	Pre	3	2		3.60	0.027	72.10	75.52
	Post			1	4.55	0.011	27.90	24.48
4 moras, 2 vcd	Pre	4	2		4.53	0.011	50.46	55.82
	Post			1	4.55	0.011	49.54	44.18
5 moras, 2 vcd	Pre	5	2		5.47	0.004	28.65	22.54
	Post			1	4.55	0.011	71.35	77.46
6 moras, 2 vcd	Pre	6	2		6.40	0.002	13.66	15.82
	Post			1	4.55	0.011	86.34	84.18

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 this success, Figure 3 plots the correlation between the observed percentages in the experiment on the x-axis and the percentagse predicted by the MaxEnt analysis on the y-axis, which shows a good fit between the two measures.

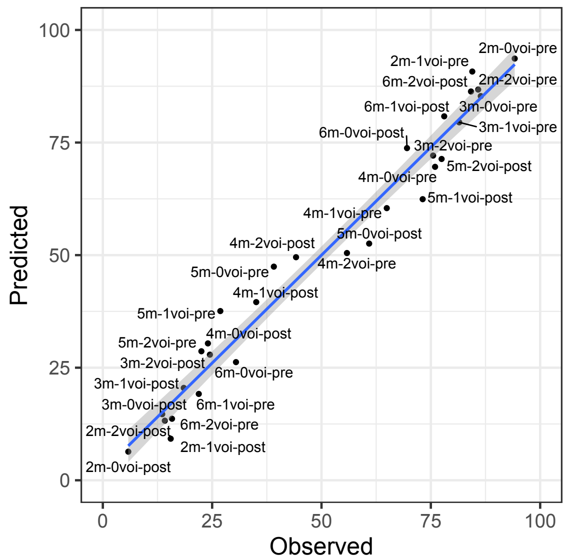


Figure 3: The correlation between the observed percentages in the experiment (the x-axis) and the percentages predicted by the MaxEnt analysis (the y-axis).

Since a MaxEnt analysis is a statistical model (Jurafsky & Martin 2019), we can assess the necessity of each constraint (i.e. factor) using a log-likelihood ratio test (Winter 2019). To do so, we compare the full model with the three constraints and smaller models with two of the three constraints. By removing one of the three constraints, we obtain three simpler models with two remaining constraints. We then compare their log-likelihood values by taking their ratios, which addresses the question of whether the full model fits the data better than the simpler models to a statistically significant degree. The results of these log-likelihood ratio tests are summarized in Table 3, which demonstrates that all three constraints are statistically justified to be included in the MaxEnt analysis. In other words, each of the three constraints plays a role in the explanation of the data, in addition to what is explained by the other two constraints. See the Appendix in Breiss & Hayes (2020) for related discussion.

Table 3: The results of the log-likelihood ratio tests. The log-likelihood of the best fitting model with the three constraints is -784.6. See Supplementary File B.

	Δlikelihood	$\chi^2(1)$	p
*LONGPRE	228.3	456.6	$< .001$
*VCDPRE	14.2	28.4	$< .001$
*POST	247.4	494.8	$< .001$

In addition, a more complex model was tested by adding the fourth constraint which represents the interaction term between *LONGPRE and *VCDPRE, which is equivalent to the locally conjoined version of these two constraints (Shih 2017). The Solver assigned a weight that is close to 0 (=0.005) to the conjoined constraint, and addition of this constraint did not improved the model fit at all. This is a welcome result, since the interaction of the effects of voiced obstruents and those of mora counts followed directly from the architecture of MaxEnt itself, obviating the need to posit a specific constraint to capture the interaction between the two factors (see Zuraw & Hayes 2017 who reached the same conclusion in their phonological analyses; though see Shih 2017 who argued that a locally-conjoined constraint may be necessary even in a MaxEnt grammar). The result here nicely aligns with the Bayesian logistic regression modeling presented in §3.

Before closing this section, some remarks on Stochastic Optimality Theory (Stochastic OT) (Boersma 1998; Boersma & Hayes 2001) are in order. Stochastic OT is an alternative framework which has been used to model probabilistic phonological patterns. It is no different from Classical OT at each time of evaluation. However, constraints are assigned ranking values, and these values can be perturbed by a Gaussian noise, which derives probabilistic variations. One challenge that this framework faces in modeling the current dataset is that it cannot handle counting cumulativity in general (Hayes 2020; Jäger 2007; McPherson & Hayes 2016; Zuraw & Hayes 2017). The problem is that since each evaluation trial proceeds as in Classical OT with strict domination (Prince & Smolensky 1993/2004), if *POST dominates *LONGPRE at a particular time of evaluation, a pre-evolution character wins no matter how long the name under evaluation is. Similarly, if *VCDPRE dominates *POST, a post-evolution character wins regardless of whether a name contains one voiced obstruent or two voiced obstruents. Stochastic OT therefore does not handle the effects of different mora counts or voiced obstruents. In Stochastic OT, therefore, it is necessary to expand *LONGPRE and *VCDPRE into sets of multiple constraints (Boersma 1998; McPherson & Hayes 2016): i.e. *LONGPRE3MORA, *LONGPRE4MORA, *LONGPRE5MORA , *LONGPRE6MORA, *ONEVCDPRE and *TWOVCDPRE (see Kawahara 2020b for a full implementation of an analysis of his data using a set of *LONGPRE*X*MORA constraints). All in all, then, Stochastic OT requires four additional free parameters than MaxEnt. See Kawahara (2020b) and Zuraw & Hayes (2017) for further discussion on some challenges that Stochastic OT faces in modeling wug-shaped curves.

6 Conclusion

The descriptive findings of the current experiments can be summarized as follows: (1) mora count cumulatively affects the judgment of evolvedness in Pokémon names, (2) voiced obstruents also cumulatively affect the judgment of evolvedness, and (3) these two effects are additive, which requires no complex interaction term when modeling the data. The effects of mora count and voiced obstruents instantiate counting cumulativity in sound symbolism. The fact that two separate effects interacted cumulatively suggests that we observe ganging-up cumulativity in sound symbolism as well.

The effects of mora count resulted in what looks to be sigmoidal functions. The three sigmoidal curves, which arose due to the ganging-up cumulativity, together result in what Hayes (2020) refers to as a "stripy wug-shaped curve." Since a stripy wug-shaped curve is a typical quantitative signature that MaxEnt is predicted to generate, it provides support for the thesis that MaxEnt is a useful tool to model linguistic patterns.

The coexistence of counting cumulativity and ganging-up cumulativity, which manifests itself as a stripy wug-shaped pattern in sound symbolism, draws an intriguing parallel to the recent observation made in the analyses of probabilistic phonological alternation patterns, including vowel harmony in Tommo So (McPherson & Hayes 2016) as well as liaison in French, nasal substitution in Tagalog, and vowel harmony in Hungarian (Zuraw & Hayes 2017). The coexistence of counting and ganging-up cumulativity has also recently been shown to hold in an artificial language learning experiment of phonotactic learning with native speakers of English (Breiss 2020).

Traditionally, sound symbolism has been viewed to reside outside the purview of phonological analyses, although some recent proposals argue that sound symbolic principles should be integrated with the "core" phonological grammar (Alderete & Kochetov 2017; Jang 2020; Kumagai 2019). In line with these arguments, the current results may suggest that there is a meaningful parallel between sound symbolic patterns and other phonological patterns (Kawahara 2020a). Suppose that the current proposal—that MaxEnt HG equipped with OT-style constraints is a useful tool to model sound symbolism—is on the right track, and also suppose that MaxEnt is suited to model phonological, and perhaps other linguistic patterns as well, as many previous studies have shown. Taken together, then, it points to a conclusion that the same mechanism may be regulating sound symbolic mappings and phonological patterns. I submit that this is an interesting hypothesis that can and should be explored more extensively in future research, especially given that sound symbolism did not receive much attention from theoretical phonologists in the past.

Appendix

One question that may be raised regarding the conclusion that the three sigmoid curves are all separated from one another in the current experiment (§4) is that the current Bayesian regression analysis coded the effects of voiced obstruent as a numerical variable, rather than a three-level categorical variable. This numerical coding was theoretically motivated, because we are interested in how the number of voiced obstruents affected the post-evolution responses, and it was the count of voiced obstruents that was of interest. This is reflected in the way the *VCDPRE constraint is formulated in §5 as well—it assigns a violation mark for every instance of a voiced obstruent in a pre-evolution character name.

With this said, in order to more directly address this concern, the Bayesian logistic regression model was rerun with the effect of voicing treated as an unordered categorical factor with "1 voiced obstruent" as the reference level (coded as 0 because this variable was centered). Since Bulk ESS and Tail ESS were too high with 2,000 iterations with 1,000 warmups, this new analysis ran 5,000 iterations with 2,000 warmups. All the $\hat{R}$ values were 1.00. See Supplementary Material A.

The results, summarized in Table 4, show that the difference between 0 vs. 1 voiced obstruent and the difference between 1 vs. 2 voiced obstruents both meaningfully impacted the post-evolution responses. Neither of the interaction terms were meaningful, because their 95% CIs contain 0. These results show that the three sigmoidal curves are separated from one another, even if we consider the three different voicing conditions as three manifestations of an unordered categorical factor, and under that assumption too, we can conclude that the three slopes are likely to be identical.

Table 4: Summary of the Bayesian mixed effects logistic regression model, in which the voicing effect is coded as an unordered categorical variable.

	β	error	95% CI
intercept	-0.31	0.10	[-0.51, -0.12]
mora count	1.03	0.10	[0.83, 1.23]
vcd obs (0 vs. 1)	-0.65	0.13	[-0.90, -0.39]
vcd obs (1 vs. 2)	0.27	0.12	[0.02, 0.51]
vcd obs (0 vs. 1) $\times$ mora	0.10	0.09	[-0.07, 0.28]
vcd obs (1 vs. 2) $\times$ mora	0.16	0.09	[-0.01, 0.33]

Additionally, this new analysis shows that the difference between the 0 voiced obstruent condition and the 1 voiced obstruent condition seems to be larger than the difference between the 1 voiced obstruent condition and the 2 voiced obstruent condition. To visualize, Figure 4 shows the distributions of posterior samples of the two relevant slope coefficients, which play a role in characterizing the log-odds of post-evolution responses. Since all the poster samples for the first difference were negative, I took the absolute values to compare the differences in magnitude. This analysis shows that the slope for the shift from 0 voiced obstruent to 1 voiced obstruent generally shows posterior samples that are larger in magnitude than the slope for the shift from 1 voiced obstruent to 2 voiced obstruents. Only less than 3.7% of the time (i.e. 448 out of 12,000) was the second slope larger in magnitude than the first slope. One way to understand this observation is to consider this as a case of non-linear (more specifically, sub-linear) counting cumulativity in sound symbolism (Kawahara & Breiss 2020).

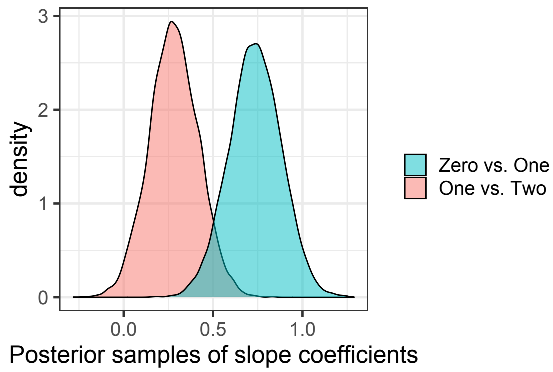


Figure 4: The distributions of posterior samples of the two voicing differences.

References

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

Boersma, Paul. 1998. *Functional phonology: Formalizing the interaction between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

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

Breiss, Canaan. 2020. Constraint cumulativity in phonotactics: Evidence from artificial grammar learning studies. Ms. UCLA.

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

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

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam & Moris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

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.

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

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.

Good, Phillip I & James W. Hardin. 2006. *Common errors in statistics: (and how to avoid them)*. Wiley.

Haiman, John. 1980. The iconicity of grammar: Isomorphism and motivation. *Language* 56(3). 515–540.

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

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

Hayes, Bruce. 2020. Assessing grammatical architectures through their quantitative signatures. Talk presenetd at BLS.

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

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

Jaeger, Florian T. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59. 434–446.

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.

Jäger, Gerhad & Anette Rosenbach. 2006. The winner takes it all—almost: Cumulativity in grammatical variation. *Linguistics* 44(5). 937–971.

Jang, Hayeun. 2020. How cute do I sound?: The iconic function of segmental alternation in Korean baby-talk register, *aegyo*. *Language Sciences* .

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

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

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

Kawahara, Shigeto & Canaan Breiss. 2020. Exploring the nature of cumulativity in sound symbolism: Experimental studies of Pokémonastics with English speakers. Ms. Keio University and UCLA.

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 & Gakuji Kumagai. 2019. Expressing evolution in Pokémon names: Experimental explorations. *Journal of Japanese Linguistics* 35(1). 3–38.

Kawahara, Shigeto & Gakuji Kumagai. 2021. What voiced obstruents symbolically represent in Japanese: Evidence from the Pokémon universe. *Journal of Japanese Linguistics* 37(1).

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

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

Kubozono, Haruo. 2015. Diphthongs and vowel coalescence. In Haruo Kubozono (ed.), *The handbook of Japanese language and linguistics: Phonetics and phonology*, 215–249. Mouton.

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.

Liberman, A. M., K. S. Harris, H. S. Hoffman & B. C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54(5). 358–368.

Marks, Lawrence. 1978. *The unity of the senses: Interrelations among the modalities*. New York: Academic Press.

McCarthy, John J. 2003. OT constraints are categorical. *Phonology* 20(1). 75–138.

McPherson, Laura & Bruce Hayes. 2016. Relating application frequency to morphological structure: The case of Tommo So vowel harmony. *Phonology* 33. 125–167.

Ohala, John. 1983. The origin of sound patterns in vocal tract constraints. In Peter MacNeilage (ed.), *The production of speech*, 189–216. New York: Springer-Verlag.

Pierrehumbert, Janet. 2020. 70+ years of probabilistic phonology. In B. Elan Dresher & Harry van der Hulst (eds.), *Oxford handbook on the history of phonology*, Oxford University Press.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Malden and Oxford: Blackwell.

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.

Sidhu, David & Penny M. Pexman. 2018. Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review* 25(5). 1619–1643.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. In D. Rumelhart, J. McClelland & PDPR Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1: Foundations, 194–281. Cambridge, MA: Bradford Boooks/MIT Press.

Tesar, Bruce. 2007. A comparison of lexicographic and linear numeric optimization using violation difference ratios. Ms. Rutgers University.

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.