Testing MaxEnt with sound symbolism: A stripy wug-shaped curve in Japanese Pokémon names*

Shigeto Kawahara
Keio University
kawahara@icl.keio.ac.jp

To appear in *Language* (the Research Reports section)

Abstract

One issue that is actively explored in the contemporary linguistics literature is how to account for probabilistic generalizations, for which there are currently various competing theories. To bear on this debate, Hayes (2020) proposes that we examine a grammatical framework in terms of its **quantitative signature**, typical probabilistic patterns that the framework is predicted to generate. In this paper, I zoom in on the quantitative signature of Maximum Entropy Harmonic Grammar (MaxEnt HG), because this framework has proven to be a useful tool to model probabilistic generalizations across different linguistic domains. Given a linear scale of violations of one constraint, MaxEnt yields a sigmoid curve. When another constraint is relevant and can be violated multiple times, this sigmoid curve can be shifted, yielding multiple sigmoid curves, which results in a **stripy wug-shaped curve**. Extending upon a previous study (Kawahara 2020b), the current experiment demonstrates that we observe a stripy wug-shaped curve in a particularly clear manner in patterns of sound symbolism, systematic associations between sounds and meanings. Concretely, the experiment with Japanese speakers shows that the judgment of Pokémons' evolution status is affected by the mora counts of nonce names, resulting in a sigmoid curve, and that this sigmoid curve is shifted according to the number of voiced obstruents contained in the names. The overall results suggest that MaxEnt is a useful tool to model systematic sound-meaning correspondences.

*Many thanks to four engaging anonymous *Language* reviewers, the associate editor, and the editor for comments on previous versions of this paper. Canaan Breiss, Donna Erickson, Bruce Hayes and Gakuji Kumagai also offered useful comments on earlier versions of this paper and/or on this general project on modeling sound symbolic patterns using MaxEnt. This project is supported by JSPS Grants #17K13448 and #18H03579. All remaining errors are mine.

1 Introduction

1.1 General theoretical background

Traditional generative analyses tended to focus on the dichotomous distinction between grammatical and ungrammatical forms. In the syntax research, this thesis was made clear from its outset: "[t]he fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not sentences of L" (Chomsky 1957: p.2, emphasis in the original). One of the aims of the early generative phonology research was also to model the distinction between possible words and impossible words: e.g. in English, *brick* and *blick* are possible words, while *bnick* is an impossible word (Chomsky & Halle 1965). A common assumption had been that grammar—or competence—only makes a dichotomous distinction between grammatical forms and ungrammatical forms (Neeleman 2012; Sprouse 2007). In the words of Pierrehumbert (2001: 195), "[t]he classical generative models are non-probabilistic. Any given sequence is either well-formed under a grammar, or it is completely impossible."

However, an accumulating body of evidence shows that probabilistic generalizations are just as important as categorical generalizations for linguistic inquiry. In sociolinguistics research, for which variation has been the central topic of investigation (Labov 2004), it has been long noted that variation shows systematic probabilistic patterns, which are influenced by both linguistic and non-linguistic considerations. To take a well-studied example of English *t/d*-deletion, the probability of this process has been shown to be systematically affected by various morphological, phonological and socio-linguistic factors (Coetzee & Kawahara 2013; Guy & Boberg 1997). As such, in this research tradition, various grammatical models have been formulated to account for such probabilistic generalizations (Cedergren & Sankoff 1974; Johnson 2009; Rousseau & Sankoff 1978).

Recent generative phonology research has also started paying serious attention to stochastic generalizations, recognizing that phonological knowledge is fundamentally stochastic (Pierrehumbert 2001). One piece of evidence comes from the observation that many sound sequences are neither completely phonotactically legal nor illegal, but instead are intermediate (Daland et al. 2011). Another influential finding is that some "exceptional" phonological alternation patterns can be just as systematic as "regular" phonological patterns, being susceptible to various phonological constraints (Anttila 1997; Ernestus & Baayen 2003; Zuraw 2000). Thus, there is a rise of interest in formal, generative grammatical models which can account for such probabilistic generalizations (Coetzee & Pater 2011). The three most widely discussed frameworks in the contemporary phonology research are (1) Stochastic Optimality Theory (Boersma 1998; Boersma & Hayes 2001), (2) Noisy Harmonic Grammar (Boersma & Pater 2016; Coetzee & Kawahara 2013) and (3) Maximum

Entropy Harmonic Grammar (Goldwater & Johnson 2003; Hayes & Wilson 2008). Which model is best suited to account for the probabilistic aspects of our linguistic knowledge has been one topic that is actively discussed in the contemporary linguistics literature (Anttila & Magri 2018; Breiss 2020; Breiss & Albright 2020; Ernestus & Baayen 2003; Goldrick & Daland 2009; Hayes 2017, 2020; Jäger & Rosenbach 2006; Jäger 2007; Magri et al. 2020; Smith & Pater 2020; Zuraw & Hayes 2017, among many others).

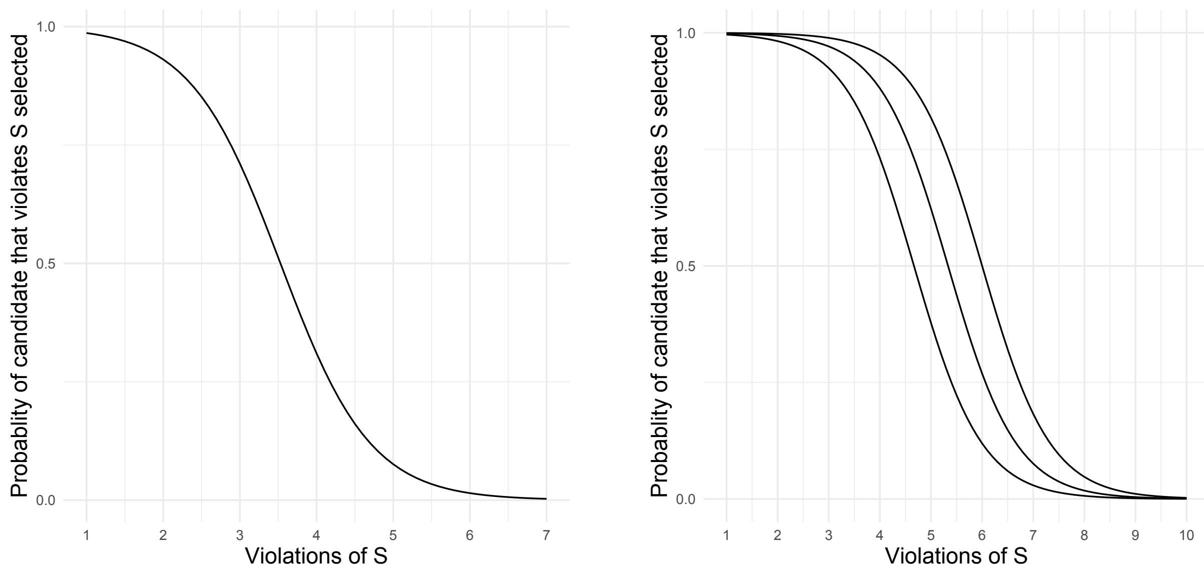
To address this general question, Hayes (2020) proposes that we examine the grammatical model of interest in terms of its **quantitative signature**. The quantitative signature refers to the set of typical probabilistic patterns that a particular framework is predicted to generate. Inspired by this proposal, the current paper zooms in on Maximum Entropy Harmonic Grammar (henceforth, MaxEnt HG) (Goldwater & Johnson 2003; Smolensky 1986), and examines its quantitative signature. This paper focuses on MaxEnt HG, because as extensively reviewed by Hayes (2020), its quantitive signature may be observed across different areas of linguistic patterns, and can thus be considered to be a useful tool to model various aspects of linguistic knowledge. MaxEnt (or something akin to it) has indeed been applied to model data from a wide variety of linguistic domains, including, but not limited to, phonetics (Lefkowitz 2005), phonology (Zuraw & Hayes 2017), syntax (Bresnan et al. 2007), semantics/pragmatics (AnderBois et al. 2012), historical changes (Kroch 1989), and sociolinguistics (Rousseau & Sankoff 1978) (again, see Hayes 2020).

In the context of current linguistic theorization, there are two ways to understand MaxEnt HG. One is to consider it as application of logistic regression modeling for linguistic analyses (Jurafsky & Martin 2019). The other is to consider it as a probabilistic 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. Section 5 presents the mathematical details for how the probability distribution over the candidates is calculated, but following Hayes (2020), here let us focus on the general prediction that the theory makes—i.e. its quantitive signature—without going into the mathematical details.

1.2 The quantitative signature of MaxEnt

To reiterate, the quantitative signature is the set of probabilistic patterns that a particular grammatical framework is predicted to generate. This subsection describes the quantitative signature of MaxEnt HG, as demonstrated by Hayes (2020). One of the motivations for examining the quantitative signature of MaxEnt HG is the recurrent observation that similar quantitative patterns hold across different linguistic domains, and that such probabilistic linguistic generalizations often exhibit a set of multiple sigmoidal curves. These observations are a natural consequence of MaxEnt HG.

To illustrate, suppose that there is a scalar constraint S, whose violation can be assessed on a linear numerical scale (e.g. 1, 2, 3...etc). 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, MaxEnt yields a sigmoid curve, as shown in Figure 1(a). The linear violation scale (i.e. the x-axis) 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.



(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. The weight of the constraint B serves as an intercept term for $-N$. (b) Multiple sigmoid curves, shifted by multiple violations of P, resulting in a stripy wug-shaped curve.

When another constraint—call it P for "Perturber"—is at play, this sigmoid curve can be shifted horizontally 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). Mathematically speaking, a stripy wug-shaped curve has three defining features: (1) it consists of multiple sigmoid curves, (2) they are separated from one another, and (3) the slopes of the sigmoid curves are identical.

As Hayes (2020) demonstrates, a stripy wug-shaped curve is commonly observed in probabilistic phonological alternation patterns (Ernestus & Baayen 2003; McPherson & Hayes 2016; Zuraw & Hayes 2017), categorical perception of speech (Kluender et al. 1988), diachronic changes in syntax (Kroch 1989; Zimmermann 2017), and possibly other linguistic domains.[1] Hayes (2020) further demonstrates that these wug-shaped curves are difficult to derive in another widely-used probabilistic grammatical model, 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 the linguistic knowledge that lies behind these probabilistic patterns.[2]

1.3 The current study

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 in the Pokémon universe. 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 can be considered to be identical and separated from one another (§3 and §4). A MaxEnt-based phonological analysis of the sound symbolic connections found in the experiment shows that the sound symbolic patterns can indeed be nicely modeled by MaxEnt HG, when we posit that MaxEnt HG mediates the mapping from sound to meaning, with the sort of constraints that are used in the OT research

[1] See the website accompanying Hayes (2020) for the actual examples of wug-shaped curves: `https://linguistics.ucla.edu/people/hayes/GalleryOfWugShapedCurves/index.htm` (last access, April 2021). See also Harrison et al. (2002) and references cited therein for additional cases of sigmoidal curves in phonological diachronic changes.

[2] 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.

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).

Before I proceed, I would like to highlight the novel contributions that the current paper makes beyond Kawahara (2020b). First, Kawahara (2020b) only tested a distinction between the presence and absence of a voiced obstruent, and as such, the work was unable to address the question of whether we would observe a stripy wug-shaped curve. The current experiment overcomes that limitation by varying the number of voiced obstruents from 0 to 2. Second, as detailed below (§3 and §4), a stripy wug-shaped curve entails that the slopes of all sigmoidal curves are identical. Whether this holds true or not cannot be addressed by a frequentist statistical analysis, which Kawahara (2020b) adapted, because it cannot be used to "prove the null effect" (Gallistel 2009). The current paper overcomes this limitation by making use of Bayesian analyses, which can quantify the evidence for the null effect. Third, this paper used a better controlled stimulus set, thereby removing some ambiguity in interpretation that Kawahara (2020b) was unable to resolve.

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 Kawahara & Breiss (2021) for a recent overview of the 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 called by a different name. The first Pokémonastics study, which analyzed the existing Pokémon names in Japanese (Kawahara et al. 2018), found 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 sound symbolic 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), which may have a domain-general cognitive basis (Marks 1978). The second observation is perhaps rooted in the well-known observation in Japanese that voiced obstruents denote larger quantities (Hamano 1998). This sound symbolic connection itself may be grounded in the expansion of the oral cavity which occurs during the production of voiced obstruents (Ohala 1983), or in the low frequency energy of voiced obstruents (Kingston & Diehl 1994), which is known to be associated with large images (via a mechanism often referred to as the "Frequency Code": Ohala 1994). Several experimental studies using nonce names have confirmed the productivity of these two sound symbolic patterns in Japanese (Kawahara 2020b; Kawahara & Kumagai 2021), as well as in other languages including Brazilian Portuguese (Godoy et al. 2020) and English (Kawahara & Breiss 2021) . The current experiment makes use of these sound symbolic patterns in order to address the specific prediction that MaxEnt makes, which was illustrated in Figure 1(b).

2.2 Stimuli

The current experiment heavily draws upon Kawahara (2020b), who manipulated the number of moras 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 correspond 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.[3] Since [p] is known to have its own salient sound symbolic values, such as cuteness (Kumagai 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 whole name, resulted in artificial names, especially in long names.

[3] `http://sei-street.sakura.ne.jp/page/doujin/site/doc/tool_genKanaName.html` (last access, April 2021). Using a random generator is important, because it precludes the bias that the experimenter may have prior to the design of the experiment (Westbury 2005). 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]

2.3 Procedure

The experiment was distributed online using SurveyMonkey. The first page of the experiment presented a consent form, which had been approved by the author's institute. The instructions stated that the participation in this experiment is completely voluntary, and that it 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 for each name whether it 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 judgments based on their intuitions rather than thinking about "right" or "wrong" answers. The stimuli were presented in the Japanese *katakana* orthography, which is used for real Pokémon names. They were asked to read each name silently 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 symbolism. Four participants reported that they had participated in another Pokémonastics experiment. The data from these ten 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). For accessible introductory books to Bayesian modeling, see Kruschke (2014) and McElreath (2020), the former of which is rather succinctly summarized in Kruschke & Liddell (2018). Nicemboim & Vasishth (2016) and Franke & Roettger (2019) offer shorter tutorials on Bayesian analyses using linguistic examples. Vasishth et al. (2018) provide a slightly more advanced overview of Bayesian analyses using phonetic data.

Bayesian regression analyses yield, for each estimated parameter, a posterior distribution, given our prior belief about the estimate and the data that are being analyzed. These posterior distributions can be interpreted as directly reflecting our updated (un)certainty about the estimate, after the data are observed. As a useful heuristic, for each estimate that we are interested in, we can examine the middle 95% of the posterior distribution, called the 95% Credible Interval (abbreviated as "95% CI"). One rule of thumb, which is roughly analogous to significance testing in a more traditional frequentist approach, is that if the 95% CI does not include 0, we can be 95% certain that that effect meaningfully impacts the responses. If the 95% CI includes 0, on the other hand, we can examine its posterior distribution in closer detail and test how confident we can be regarding its null effect (see section 3 for further details).

In the current model, the dependent variable was the binary response obtained in the experiment (0 = pre-evolution, 1 = post-evolution). The fixed 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. Four chains of 3,000 iterations were run, and the last 2,000 samples from each chain were analyzed. The default, weakly-informative priors were used. All the $\hat{R}$ values were 1.00, indicating that the chains mixed successfully.

Since the current model is based on logistic regression, a positive slope coefficient (β) indicates that that factor increases post-evolution Pokémon responses. In the spirit of open science initiative in linguistics (see e.g. Winter 2019), the raw data as well as the R code and all Bayesian posterior samples are made available at the osf repository.[4] Interested readers are welcome to examine the posterior samples in further detail.

Taking a Bayesian approach has two 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 (e.g. Gallistel 2009; Kruschke 2014; Kruschke & Liddell 2018). 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 term between the mora count and the voiced obstruent *must* be "null", since the interaction term functions as a slope adjustment term (Winter 2019). In order to be confident that we observe a (stripy) wug-shaped curve, it is necessary to show the null effect of an interaction term, but this is possible only in a Bayesian framework, not in a frequentist framework.

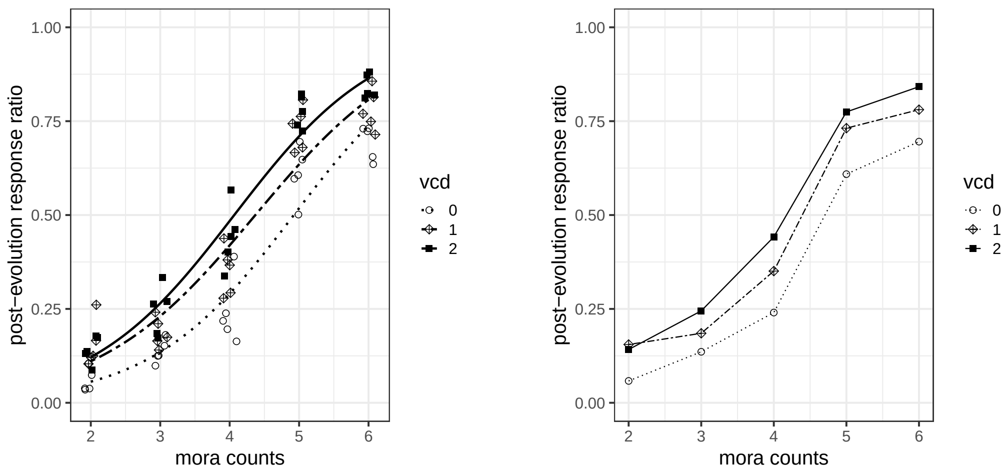
This is one clear limitation of Kawahara (2020b), who was unable to address this question, because the analysis deployed by Kawahara (2020b) was a frequentist analysis—the best that could be concluded based on that analysis was that the interaction term was non-significant. The lack of significance, however, only means that we cannot reject the null hypothesis that the slopes were not identical. The Bayesian analysis presented in the current paper overcomes this limitation.[5]

3 Results

Figure 2(a) plots "post-evolution response ratios" for each item, averaged over all the participants. White circles represent items which contain no voiced obstruents, diamonds show the results of the items which contain one voiced obstruent, and filled 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 curve shown in Figure 1(b).

[4] `https://osf.io/tn6cu/`. They are also available as supplementary materials at the journal website.

[5] Extending on the current results, the patterns analyzed by Hayes (2020) should thus ideally be reanalyzed using a Bayesian framework.



(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, unlike Figure 2(a), 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 does 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.6). It requires very convincing evidence to be certain that a particular name is for a pre-evolution character or for a post-evolution character.

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 indicates 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 is positive and its 95% CI does not include 0, which shows that an increase in mora counts meaningfully increases the probability of the names assigned to a post-evolution character. The slope of voiced obstruents is also positive, and its 95% CI does not include 0. These results suggest that voiced obstruents meaningfully increase post-evolution responses as well.

The 95% CI for the interaction term on the other hand includes 0. Addressing whether the interaction term should be considered to be null is important, because if all the curves involve the identical slope, as predicted by MaxEnt, then the interaction term should play no meaningful

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

	β	error	95% CI
intercept	-0.56	0.08	[-0.72, -0.40]
mora count	1.32	0.09	[1.14, 1.51]
vcd obs	0.49	0.08	[0.34, 0.66]
mora count $\times$ vcd obs	0.03	0.05	[-0.07, 0.13]

role. To address this question, in this paper I make use of a ROPE analysis (Region Of Practical Equivalence) (Kruschke & Liddell 2018; Makowski et al. 2019). The basic idea is that we define a range that is "close enough" to 0. A general rule of thumb is that if the 95% CI is contained in that ROPE, then we can accept the null value for that parameter for practical purposes. Following Makowski et al. (2019), I take the ROPE to range from -0.1 to 0.1 of a standardized parameter (= negligible effect size, as defined by Cohen 1988). In a logistic model, a standardized parameter can be approximated as $\frac{\pi}{\sqrt{3}}$ (Makowski et al. 2019). The ROPE that is of interest therefore ranges from [-0.18, 0.18].

Since the 95% CI of the interaction term of the model (=[-0.07, 0.13]) is fully contained in this ROPE, it is safe to conclude that the slopes between the three curves in Figure 2 are, practically speaking, identical. Moreover, we can calculate how many of the posterior samples are contained in this ROPE, using the `bayestestR` package (Makowski et al. 2020). The result shows that 99.8% of all the posterior samples are contained in the ROPE—we are thus 99.8% certain that the three curves in the current results can be considered to be identical.

4 Discussion

I would like to first discuss in some detail whether the current experimental results shown in Figure 2 should indeed be considered to instantiate a stripy wug-shaped curve. To reiterate, 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) the slopes of the sigmoid curves are identical.[6]

The discussion at the end of section 3 already shows that the current results satisfy the third

[6]Technically speaking, 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 (for which see section 5), the second question can be restated as a question regarding whether we should allow for a locally conjoined constraint in a MaxEnt grammatical model (Shih 2017). See Hayes (2020) for further discussion on this point. For now, I assume, for the sake of theoretical restrictiveness, that different sigmoid curves should have the same slope in stripy wug-shaped curves.

requirement: the ROPE analysis shows that we can be 99.8% certain that the interaction term is null; i.e. the slopes between the three curves can be treated as identical for practical purposes.

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 obstruents meaningfully impacts the post-evolution responses. One may still wonder whether *all* three curves are separated from one another. This question is addressed by an additional analysis presented in the appendix, which shows that all three curves are indeed separated from each other.

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. 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; Breiss & Albright 2020; Hayes 2020; Hayes & Wilson 2008; McPherson & Hayes 2016; Zuraw & Hayes 2017). With this, we turn to a full MaxEnt analysis of the current experimental results in the next section.

Before proceeding, one additional note about the current results is in order. Unlike the idealized mathematical shapes of stripy wug-shaped curves in Figure 1(b), the current wug-shaped curves do not cap out at 1 or 0; i.e. even the most extreme conditions do not show 100% pre-evolution responses or 100% post-evolution responses, although the bimoraic names without a voiced obstruent get rather close to 100% pre-evolution responses. This feature may be related to the nature of sound symbolism, which is inherently stochastic (Kawahara et al. 2019). Given that a language is a system that can associate sounds and meanings in arbitrary ways (Saussure 1916), it may be unlikely that all forms with particular phonological structures behave in the same way so that they are deterministically associated to a particular meaning. Alternatively, it is possible that the current experiment is simply tapping only a subset of the whole sigmoidal curves. This possibility can be addressed by a new experiment using more extreme phonological properties (e.g. longer names). A final possibility is that this feature is merely an artifact of noise that is unavoidable in any behavioral task and/or due to uncontrolled factors, such as vowel quality.

5 A MaxEnt analysis

This section develops a MaxEnt analysis of the sound symbolic connections found in the experiment, using the sort of constraints that have been used in the Optimality Theory research (Prince & Smolensky 1993/2004). Before we proceed, I note at this point that mathematically, MaxEnt is equivalent to multinomial logistic regression (Jurafsky & Martin 2019), and therefore there is some conceptual overlap between the statistical analysis presented in section 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 useful discussion on this difference).[7] I find it particularly important to show that we can build a **generative phonological analysis of sound symbolism**, because sound symbolism had been a topic that was almost never seriously addressed in the generative tradition (Alderete & Kochetov 2017; Kawahara 2020a). To highlight the fact that the analysis in this section builds on the generative phonological research tradition, the set of constraints used below are formulated using a constraint schema proposed by McCarthy (2003).

Since there are already a number of papers that explain how MaxEnt works for linguistic analyses (e.g. Breiss & Hayes 2020; 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

[7] Moreover, there are subtle but important differences between purely statistical logistic regression analyses and linguistic MaxEnt analyses. For example, in logistic regression analyses, there is nothing that prevents slope coefficients to be negative or positive; in MaxEnt analyses, on the other hand, constraints are formulated so that they only *penalize*—but not reward—particular structures or particular mappings between two representations. Whether or not we can reward a constraint violation is at best a contended issue (Kaplan 2018). Perhaps more strikingly, in logistic regression analyses, it is recommended that we center a continuous variable (Winter 2019), as we did in section 3. Directly translating this practice into linguistic MaxEnt analyses amounts to admitting constraints to reward some constraint violations and penalize other violations. To the best of my knowledge, no such proposals have been put forward within a constraint-based analysis of linguistic patterns. Finally, several researchers have proposed to impose particular formal restrictions on the set of constraints admitted by UG (McCarthy 2003; Potts & Pullum 2002); statistical analyses, on the other hand, do not have such limits on the choice of predictor variables. Overall, linguistic MaxEnt analyses are more restrictive than purely logistic regression analyses. Exploring this difference may be a way to address the question of how UG is different from a purely statistical learning device.

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_i 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} or equivalently, $\frac{1}{e^H}$), 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 $\frac{e^{-H(x_j)}}{Z}$.

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 are 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).[8]

(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.

[8] One may object that constraints used in OT/MaxEnt analyses should not refer to arguably Pokémon-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 (Sidhu & Pexman 2018)).

(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

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 (Fylstra et al. 1998) so as to maximize the log-likelihood of the data with respect to the model, i.e. the constraint set. The Excel sheet used to calculate the optimal weights and the predicted values is available at the osf repository mentioned above (also available as supplementary material).[9] 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

[9]The supplementary materials also contains a tutorial video on how to calculate optimal weights using Excel's Solver function.

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. Comparing the last two columns of these tableaux, the match between the observed percentages and predicted percentages generally seems to be very good.

Before closing this section, some remarks on Stochastic Optimality Theory (Stochastic OT) are in order. Stochastic OT is an alternative framework which has been used to model probabilistic phonological patterns (Boersma 1998; Boersma & Hayes 2001; Zuraw 2000). 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 effects 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. Similarly, if *VCDPRE dominates *POST, a post-evolution character wins regardless of how many voiced obstruents that name contains. 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 (i.e. seven free parameters vs. three free parameters). See Hayes (2020), 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) the mora counts affect the judgment of evolvedness in Pokémon names, (2) the number of voiced obstruents also affects the judgment of evolvedness, and (3) these two effects are additive, which requires no meaningful interaction term when modeling the data. The effects of mora count resulted in what looks to be sigmoidal functions. The three sigmoidal curves together result in what Hayes 2020 refers to as a stripy wug-shaped curve. Since a stripy wug-shaped curve is a quantitative signature that MaxEnt is predicted to generate, it provides support for the thesis that MaxEnt is a useful tool to model sound symbolic patterns, and perhaps more broadly, linguistic patterns.

The stripy wug-shaped pattern in sound symbolism found in the current experiment draws an intriguing parallel to the recent observation made in the analyses of probabilistic phonological alternation patterns, including consonant devoicing patterns in Dutch (Ernestus & Baayen 2003), 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)—see Hayes (2020) for other potential cases. 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; Kumagai 2019). In line with these arguments, we can echo Kawahara's (2020b) claim that there may be a meaningful parallel between sound symbolic patterns and other phonological patterns. 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 (section 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. This is reflected in the way the *VCDPRE constraint is formulated in the MaxEnt analysis (section 5) as well—it assigns a violation mark for *every* instance of a voiced obstruent in a pre-evolution character name.

Nevertheless, recall that one major difference between the current experiment and Kawahara (2020b) is the addition of the contrast between 1 voiced obstruent and 2 voiced obstruents. Therefore, it would be informative to more directly show that the difference that arises from this new experimental manipulation is credible. In order to address this issue, 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 3,000 iterations with 1,000 warmups, this new analysis ran 5,000 iterations with 2,000 warmups. Other aspects of the analysis were the same as the one reported in the main body of the paper. All the $\hat{R}$ values were 1.00. See the R markdown file in the supplementary materials for complete details.

The results, summarized in Table 3, 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. 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. In short, the difference between 1 voiced obstruent and 2 voiced obstruents, which Kawahara (2020b) did not test, credibly affected Japanese speakers's responses.

Table 3: 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	-0.41	0.11	[-0.63, -0.19]
mora count	1.25	0.11	[1.05, 1.47]
vcd obs (0 vs. 1)	-0.74	0.15	[-1.03, -0.45]
vcd obs (1 vs. 2)	0.28	0.14	[0.01, 0.55]
vcd obs (0 vs. 1) × mora	0.11	0.10	[-0.09, 0.31]
vcd obs (1 vs. 2) × mora	0.13	0.10	[-0.06, 0.32]

Neither of the interaction terms were at first sight very credible, because their 95% CIs contain 0. However, the 95% CIs for the interaction terms are not fully contained in the ROPE of the hypothesis that they should indeed be treated as null (=[-0.18, 0.18]). Thus I further examined the posterior distributions of these two interaction terms using `bayestestR` (Makowski et al. 2020), and found that 75% and 70% of the whole posterior distributions are contained in this ROPE. This result indicates that we can be at least 70% certain that the slopes between the three slopes are identical in this analysis.[10]

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 3 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 posterior 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 about 2.8% of the time (i.e. 340 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

[10]Following suggestions by Makowski et al. (2019, 2020), I tested several ranges of CIs: 79% and 73% of the 89% CIs are included in the ROPE, and 77% and 71% of the 95% CIs are included in the ROPE.

symbolism (Kawahara & Breiss 2021; see also Breiss & Albright 2020).

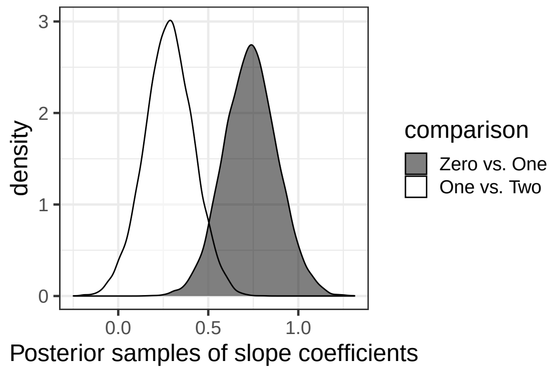


Figure 3: The posterior sample distributions of the slope coefficients for the two voicing differences.

References

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

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

Anttila, Arto. 1997. *Variation in Finnish phonology and morphology*: Stanford University Doctoral dissertation.

Anttila, Arto & Giorgio Magri. 2018. Does MaxEnt overgenerate? Implicational universals in Maximum Entropy Grammar. *Proceeings of the Annual Meeting of Phonology 2017* .

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.

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. 2020. Constraint cumulativity in phonotactics: Evidence from artificial grammar learning studies. *Phonology* 37(4). 551–576.

Breiss, Canaan & Adam Albright. 2020. Cumulative markedness effects and (non-)linearity in phonotactics. Ms. UCLA and MIT.

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.

Cedergren, Henrietta J. & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50. 333–355.

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

Chomsky, Noam & Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1. 97–138.

Coetzee, Andries W. & Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 31(1). 47–89.

Coetzee, Andries W. & Joe Pater. 2011. The place of variation in phonological theory. In John A. Goldsmith, Jason Riggle & Alan Yu (eds.), *The handbook of phonological theory, 2nd edition*, 401–431. Oxford: Blackwell-Wiley.

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

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology* 28(2). 197–234.

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.

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

Franke, Michael & Timo B. Roettger. 2019. Bayesian regression modeling (for factorial designs): A tutorial. Ms. `https://doi.org/10.31234/osf.io/cdxv3`.

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.

Godoy, Mahayana C., Neemias Silva de Souza Filho, Juliana G. Marques de Souza, Hális Alves & Shigeto Kawahara. 2020. Gotta name'em all: An experimental study on the sound symbolism of Pokémon names in Brazilian Portuguese. *Journal of Psycholinguistic Research* 49. 717–740.

Goldrick, Matthew & Robert Daland. 2009. Linking speech errors and phonological grammars: insights from Harmonic Grammar networks. *Phonology* 26(1). 147–185.

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.

Guy, Gregory & C. Boberg. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change* 9. 149–164.

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.

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.

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.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed–effects variable rule analysis. *Language and Linguistic Compass* 3(1). 359–383.

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

Kaplan, Aaron. 2018. Positional licensing, asymmetric trade-offs, and gradient constraints in Harmonic Grammar. *Phonology* 35(2). 247–286.

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). 383–418.

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

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. 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.

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.

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.

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.

Labov, William. 2004. Quantitative analysis of linguistic variation. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics: An international handbook* *of the science of language and society, volume 1: 2nd edition*, 6–21. Berlin: Mouton de Gruyter.

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

Magri, Giorgio, Scott Borgeson & Arto Anttila. 2020. Equiprobable mappings in weighted constraint grammars. *Proceedings of the Society for Computation in Linguistics* 3. 439–440.

Makowski, Dominique, Mattan S. Ben-Shachar & Daniel Lüdecke. 2019. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software* 4(40). 1541.

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.

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.

McElreath, Richard. 2020. *Statistical Rehinking: A Bayesian Course with Examples in R and Stan, 2nd edition*. London: Taylor & Francis Ltd.

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

Neeleman, Ad. 2012. Comments on Pullum. *Mind and Language* 28(4). 522–531.

Nicemboim, Brune & Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational Ideas – Part II. *Language and Linguistic Compass* 10. 591–613.

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.

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.

Pierrehumbert, Janet B. 2001. Stochastic phonology. *GLOT* 5. 1–13.

Potts, Christopher & Geoffrey Pullum. 2002. Model theory and the content of OT constraints. *Phonology* 19. 361–393.

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

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.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Paris: Payot.

Shih, Stephanie S. 2017. Constraint conjunction in weighted probabilistic grammar. *Phonology* 34(2). 243–268.

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

Smith, Brian W. & Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa* 5(1). 24, doi: `http://doi.org/10.5334/gjgl.583`.

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.

Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syn-

tax. *Biolinguistics* 1. 123–134.
Vasishth, Shravan, Bruno Nicenboim, Mary Beckman, Fangfang Li & Eun Jong Kong. 2018. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71. 147–161.
Westbury, Chris. 2005. Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language* 93. 10–19.
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. 2000. *Patterned exceptions in phonology*: University of California, Los Angeles Doctoral dissertation.
Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: An argument for Harmonic Grammar. *Language* 93. 497–548.