A wug-shaped curve in sound symbolism: The case of Japanese Pokémon names

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

Abstract

Whether linguistic patterns show cumulative effects or not is an issue that is currently debated in contemporary phonological studies. This paper attempts to shed new light on this debate from a novel perspective by studying sound symbolism, systematic associations between sounds and meanings. The current experiment shows that when Japanese speakers judge Pokémons' evolution status based on nonce names, the judgments are affected both by their mora count and the presence of a voiced obstruent. The effects of mora count instantiate a case of counting cumulativity, while the interaction between these two factors instantiates ganging-up cumulativity. These two cumulative patterns observed in the current experiment together result in what Hayes (2020) refers to as "wug-shaped curves," a quantitative signature predicted by MaxEnt. The paper shows that the experimental results can indeed be successfully modeled using MaxEnt, equipped with the sort of phonological constraints that have been used in the Optimality Theoretic research. I also examine Stochastic Optimality Theory in light of the current experimental results and show that this theory faces an interesting set of challenges. Finally, in this paper I make a novel methodological proposal for general phonological inquiry and sound symbolism research. The current study was inspired by Hayes (2020), a proposal made within formal phonology. The experiment ended up revealing important, hitherto understudied aspects of sound symbolism, and in turn, it revealed how cumulativity manifests itself in linguistic patterns. The current exploration thus shows by way of a case study that formal phonology and research on sound symbolism can mutually inform one another.

1 Introduction

1.1 A wug-shaped curve

Traditional generative theories of linguistics tended to focus on categorical generalizations, which assumed that grammar makes only dichotomous distinctions between grammatical forms and ungrammatical forms. This assumption was often made clear in the syntactic research in which grammaticality distinction is taken to be binary (e.g. Chomsky 1957; Schütze 1996; Sprouse 2007). The same spirit was shared by the early work in the generative phonological research, in which the crucial distinction had been a distinction between impossible forms (e.g. *bnick*) and possible/existing forms (e.g. *brick* or *blick*) (Chomsky & Halle 1968; Halle 1978). Probabilistic or stochastic generalizations were hardly the focus of formal phonological analyses, although in practice, exceptions to phonological generalizations were usually acknowledged and handled by some means (e.g. Kisseberth 1970).

On the other hand, probabilistic generalizations regarding phonological variations have been a central topic of sociolinguistic research (e.g. Guy 2011; Labov 1966, 1969), in which it is claimed that "variation is the central problem of linguistics" (Labov 2004: 6). It is not uncommon, for example, that the same word can be produced differently in different social or discourse contexts. Some phonological processes can apply with different probabilities in different contexts, and these probabilities can be predicted based on the interaction of various (morpho-)phonological and social factors (e.g. *t/d*-deletion in English: Guy 1991), the observation which has been modeled using various formal frameworks (e.g. Cedergren & Sankoff 1974; Johnson 2009; Guy 1991). Syntactic variations and their historical changes also seem to exhibit systematic quantitive patterns (Kroch 1989; Zimmermann 2017), which have been analyzed from formal perspectives as well (e.g. Bresnan & Hay 2008; Featherston 2005; Kellar 2006).

In harmony with these views, a growing body of recent studies has shown that phonological knowledge is deeply stochastic in nature (e.g. Boersma & Hayes 2001; Coetzee & Pater 2011; Cohn 2006; Daland et al. 2011; Hayes & Londe 2006; Pierrehumbert 2001, 2020; Zuraw 2000). Some phonotactic sequences are neither completely grammatical nor ungrammatical, but can instead be intermediate; indeed, controlled phonotactic judgment experiments typically reveal a continuous, gradient pattern (e.g. Daland et al. 2011).

Therefore, the field of phonology has recently witnessed a rise of interests in formal grammatical models which can account for such stochastic phonological generalizations. Among these, the three most widely employed frameworks are (1) Stochastic Optimality Theory (Boersma 1998; Boersma & Hayes 2001; Zuraw 2000), (2) Noisy Harmonic Grammar (Boersma & Pater 2016; Coetzee 2016; Coetzee & Kawahara 2013; Hayes 2017), and (3) Maximum Entropy Harmonic Grammar (henceforth MaxEnt) (Goldwater & Johnson 2003; Zuraw & Hayes 2017). Teasing apart

these stochastic models of phonology is a topic that is currently debated in contemporary phonological studies (Anttila & Magri 2018; Anttila et al. 2019; Breiss 2020; Breiss & Albright 2020; Hayes 2017, 2020; Jäger & Rosenbach 2006; Jäger 2007; O'Hara 2017; Pizzo 2015; Smith & Pater 2020; Zuraw & Hayes 2017 among many others).

In order to distinguish between these theoretical frameworks, building upon a body of previous studies on probabilistic linguistic patterns (Kroch 1989; McPherson & Hayes 2016; Zimmermann 2017; Zuraw & Hayes 2017), Hayes (2020) proposed to take an abstract, top-down approach, asking the following question: if we take the MaxEnt grammar framework seriously, what predictions does it make in terms of its quantitative signature, i.e., the probabilistic pattern that it typically generates? To be more specific, suppose that there is a scalar constraint, S, that is gradiently violable—i.e., its violations can be assessed on a numerical scale—and a binary constraint, B.[1] Further suppose that these constraints are in direct conflict with each other; i.e. the satisfaction of S entails the violation B, and vice versa. When we simulate the probabilities of the candidate that obeys B and violates S as a function of the number of violations of S, we get a sigmoid (s-shaped) curve, as shown in Figure 1. In reality, the constraint violation profile of S is discrete (ranging from 1 to 7 in Figure 1), but for the sake of illustration, Figure 1 continuously plots for all values, not just the integers. This curve is characterized by the fact that the y-axis values do not change very much when the x-axis values are small (from 1 to 3) or large (from 5 to 7), but there is a radical change in the middle range (from 3 to 5).

[1] Hayes (2020) uses different names (VARIABLE and ONOFF) for these two constraints.

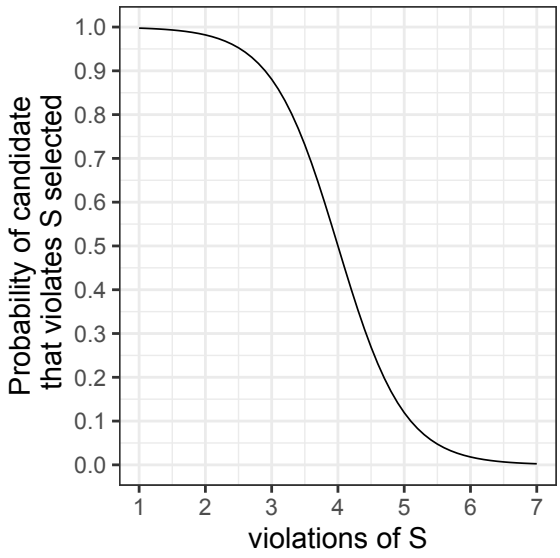
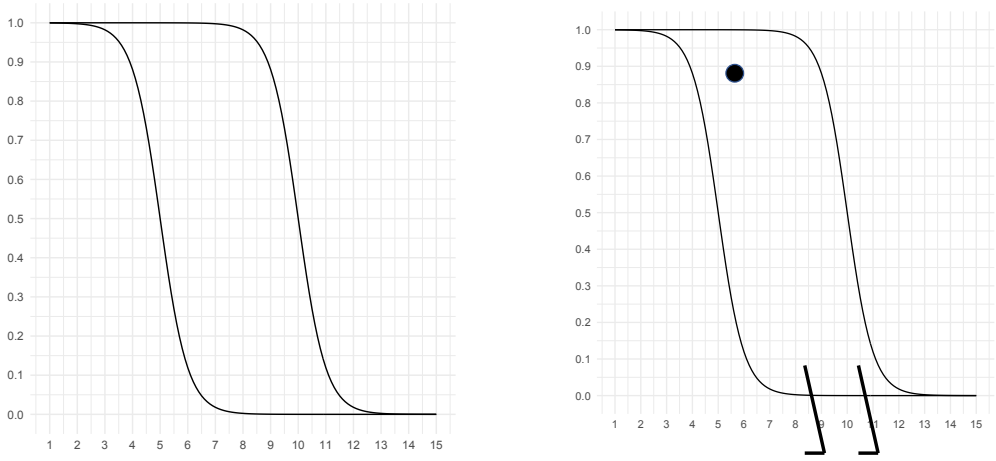


Figure 1: A sigmoid curve predicted by the MaxEnt grammar with a scalar constraint S and a binary constraint B, which are directly opposed to each other. Based on Hayes (2020: 5). The figure is redrawn and the axis labels are edited by the author. The mathematical equation which derives this curve is: $y = \frac{1}{1+e^{-H}}$, where H directly correlates with the number of violations of S. See e.g. Jurafsky & Martin (2019), McPherson & Hayes (2016) as well as §5.1.

Hayes (2020) further considers a case in which two sets of inputs are relevant—each set consists of outputs with the constraint violation profiles that are identical to those in Figure 1, but the two sets differ in terms of whether they violate an additional "perturber" constraint (P) or not. This scenario creates two identical sigmoid curves, shifted from one another on the horizontal axis, as in Figure 2(a). Hayes (2020) calls these "wug-shaped curves," because, as illustrated in Figure 2(b), they resemble the beloved animal well known in the general linguistic community, since the classic work by Berko (1958).

Studying whether a wug-shaped curve is observed in linguistic patterns is important, because the wug-shaped curves are natural outcomes of the MaxEnt grammar, and are also predicted under some versions of Noisy Harmonic Grammar, but not under Stochastic Optimality Theory. Therefore, this top-down approach to examine quantitative signatures of linguistic generalizations offers one strategy to distinguish among three competing stochastic models of grammar. If we are to find wug-shaped curves in linguistic patterns, then it provides support for the MaxEnt grammar or Noisy Harmonic Grammar over Stochastic Optimality Theory. Hayes (2020), building upon McPherson & Hayes (2016) and Zuraw & Hayes (2017), argues that such wug-shaped curves are commonly observed in probabilistic phonology, as well as in other domains of linguistic patterns, such as speech perception (classic categorical perception: Liberman et al. 1957 *et seq.*) and diachronic changes in syntax (Kroch 1989; Zimmermann 2017).



(a) (b)

Figure 2: Wug-shaped curves with two sigmoid functions. Based on Hayes (2020: 7), redrawn by the author.

Inspired by Hayes (2020) and the body of research he builds upon, this paper asks whether we can identify wug-shaped curves in the patterns of sound symbolism, systematic/iconic associations between sounds and meanings (e.g. Hinton et al. 2006). If the answer to this question turns out to be positive, then it supports the idea that MaxEnt is suited to model the knowledge that lies behind sound symbolism (Kawahara et al. 2019; Kawahara 2020a). Moreover, to the extent that MaxEnt is suited as a model of phonological knowledge (e.g. Hayes & Wilson 2008; McPherson & Hayes 2016; Zuraw & Hayes 2017 among many others), it implies that the same mechanism may lie behind phonological patterns and sound symbolic patterns; i.e. that there is a non-trivial parallel between phonological patterns and sound symbolic patterns.

1.2 Cumulativity

Stepping back from Hayes (2020) a bit, one general theoretical issue that lies behind the wug-shaped curves is that of cumulativity. Cumulativity is a topic that is currently addressed in recent linguistic theorization, because addressing this issue potentially helps us tease apart Optimality Theory (henceforth OT) (Prince & Smolensky 1993/2004) with ranked constraints from other constraint-based theories with numerically weighted constraints, such as Harmonic Grammar (Breiss 2020; Breiss & Albright 2020; Farris-Trimble 2008; Hayes et al. 2012; Jäger & Rosenbach 2006; Jäger 2007; McPherson 2016; McPherson & Hayes 2016; Pater 2009; Potts et al. 2010; Zuraw & Hayes 2017).

It is convenient to distinguish two types of cumulativity, counting cumulativity and ganging-up cumulativity (Jäger & Rosenbach 2006; Jäger 2007), because they present different types of

challenges to OT. Counting cumulativity is a case in which multiple violations of the same constraint add up. Suppose, in the context of OT, that Constraint A dominates Constraint B, then OT predicts that a single violation of Constraint A takes precedence over any number of violations of Constraint B—this is a consequence of the strict domination of constraint rankings, one central tenet of OT (Prince & Smolensky 1993/2004). In reality, however, it is not uncommon that a language tolerates one violation of a particular constraint but not two violations. For instance, the native phonology of Japanese tolerates one voiced obstruent within a morpheme, but not two voiced obstruents (a.k.a. Lyman's Law: Ito & Mester 1986, 2003). Such observations are commonly accounted for in OT by positing OCP constraints (Leben 1973; Ito & Mester 1986; Myers 1997) or self-conjoined constraints, which are violated if and only if there are two instances of the same structure (Alderete 1997; Ito & Mester 2003). Grammatical frameworks related to OT, which use numerical weights instead of rankings, can account for counting cumulativity without positing an additional mechanism (e.g. McPherson & Hayes 2016).[2]

Ganging-up cumulativity would be a case in which Constraint A dominates both Constraints B and C, but simultaneous violations of Constraints B and C would "gang-up" to take precedence over the violation of Constraint A. Such a scenario is not predicted under OT, again because of the strict domination of constraint rankings. To analyze a ganging-up cumulativity pattern, OT generally requires local conjunction of Constraints B and C (Crowhurst 2011; Smolensky 1995, 1997). For example, the loanword phonology of Japanese tolerates voiced obstruent geminates in isolation, as well as two voiced obstruent singletons. However, voiced obstruent geminates undergo devoicing when they co-occur with another voiced obstruent. In order to account for this pattern, Nishimura (2006) proposed to locally conjoin *VOICEDOBSGEM and OCP(voice) within the domain of stem. Frameworks with numerically weighted constraints can demonstrably account for this ganging-up cumulativity pattern in Japanese without stipulating a complex locally

[2]There still remains a difference between OT equipped with OCP constraints on the one hand and the related constraint-based theories with weighted constraints. One widely-shared idea in OT (and much of the pre-OT literature) is that there can be a constraint that penalizes two instances of a particular structure, but there are no constraints that penalize exactly three instances. The following quote from Ito & Mester (2003) succinctly represents this view:

> With local conjunction as a recursive operation, ternary (and higher) conjunction...[is] formally derivable... No convincing evidence has been found so far that [a ternary conjoined constraint] is ever linguistically operative separate from [a binary conjoined constraint], which tends to support the old idea in generative linguistics (cf. syntactic movement theory) that the genuine contrast in grammars is not "1 vs. 2 vs. 3 vs. 4 vs...", but "1 vs. greater than 1." (p. 265)

As implied in this quote, studies in OT generally assumed that there are no constraints that are violated if and only if there are exactly three instances of a particular structure. On the other hand, weight-based theories predict no essential differences between one violation mark vs. two violation marks and two violation marks vs. three violation marks, as will be shown in further detail in §5. It used to be believed that phonological systems do not count beyond two (e.g. McCarthy & Prince 1986), although this thesis was recently challenged by Paster (2019). See Kawahara et al. (2020b), McPherson & Hayes (2016), Paster (2019) as well as the experimental results below, for cases which apparently count beyond two.

conjoined constraint (Pater 2009) (see also Potts et al. 2010).

In short, whether phonological patterns show counting/ganging-up cumulativity is an issue that is currently addressed in contemporary phonological theorization, because it bears on the issue of whether the grammatical model should be based on rankings or weights. More broadly speaking, the question is whether the optimization algorithm deployed in the linguistic system is based on lexicographic ordering or numeric ordering (Tesar 2007).

The current paper attempts to shed new light on this debate by examining a pattern that has hitherto been barely analyzed from this perspective; namely, sound symbolism, or systematic/iconic relationships between sounds and meanings (Hinton et al. 2006). The primary question that is addressed in this study is whether sound symbolism shows cumulative effects, both in terms of counting cumulativity and ganging-up cumulativity, and if so, how.

This is an empirical question that is important to address for its own sake, because only a few studies have directly addressed the (non-)cumulative nature of sound symbolism, and therefore this is one aspect of sound symbolism that is only poorly understood. There are some impressionistic reports regarding counting cumulativity in the literature—more segments of the same kind evoke stronger sound symbolic images (Hamano 2013; Martin 1962; McCarthy 1983). Thompson & Estes (2011) addressed whether sound symbolism is categorical or gradient by way of experimentation, and found some evidence for cumulativity in their results. A recent experimental study by Kawahara & Kumagai (2021) found evidence for counting cumulativity, when they studied various sound symbolic values of voiced obstruents in Japanese. D'Onofrio (2014) examined the well-known *bouba-kiki* effect (Ramachandran & Hubbard 2001), in which certain classes of sounds are associated with round figures, whereas other classes of sounds are associated with angular figures. She found that vowel backness, consonant voicing and consonant labiality all contribute to the perception of roundness, instantiating a case of ganging-up cumulativity.[3] No studies, to the best of my knowledge, have addressed the question of whether counting cumulativity and ganging-up cumulativity can co-exist in the same sound symbolic system, as predicted by MaxEnt (though see Kawahara et al. 2020b, which is discussed in some detail in §2).

In a sense, this question—whether the same pattern can show counting cumulativity and ganging-up cumulativity at the same time—is the one addressed by Hayes (2020): each of the two sigmoid curves in a wug-shaped curve can arise when there is counting cumulativity, and the separation between the two curves is a sign of ganging-up cumulativity. It is important to note here, however, that cumulativity is a necessary, but not sufficient, condition for a wug-shaped curve. A sigmoid curve, a crucial component of a wug-shaped curve, entails counting cumulativity, but not vice versa. Counting cumulativity, for example, can manifest itself as a linear function rather than an sigmoid function. See §5.3 for further elaboration on this point.

[3] D'Onofrio (2014) herself does not directly address the issue of cumulativity. This ganging-up cumulative pattern is analyzed using MaxEnt by Kawahara (2020a).

In domains other than sound symbolism, Breiss (2020) shows that we observe both counting cumulativity and ganging-up cumulativity in phonotactic learning patterns in an artificial language learning experiment. Case studies of phonological alternation patterns reported in McPherson & Hayes (2016) and Zuraw & Hayes (2017) can also be understood as simultaneously involving counting cumulativity and ganging-up cumulativity. Apart from these studies, there are not many case studies that have directly addressed this question, especially in the domain of sound symbolism. Since the co-existence of counting cumulativity and ganging-up cumulativity is a natural consequence of MaxEnt, it should be tested with more case studies. One aim of this paper is to address this gap in the literature.

The issue of cumulativity in sound symbolism is interesting to address from a more general theoretical perspective as well. To the extent that cumulativity is a general property of phonological patterns (Breiss 2020; Hayes 2020; McPherson & Hayes 2016; Zuraw & Hayes 2017), and to the extent that sound symbolic effects show similar cumulative properties, then we may conclude that there exists a non-trivial parallel between phonological patterns and sound symbolic patterns (Kawahara 2020a). This parallel would lend some credibility to the hypothesis that sound symbolism is a part of "core" linguistic knowledge, as recently argued by several researchers (Alderete & Kochetov 2017; Jang 2019; Kawahara 2020a,b; Kumagai 2019; Shih 2020). This is a rather radical conclusion, given the fact that sound symbolism has long been considered as residing outside the purview of theoretical linguistics.

1.3 Pokémonastics

In addition to addressing the issue of cumulativity in sound symbolism, the current study can also be considered as a case study of the Pokémonastics research paradigm, within which researchers explore the nature of sound symbolism using Pokémon names (Kawahara et al. 2018; Shih et al. 2019). I refer the readers to Shih et al. (2019) for the discussion of several advantages of this research paradigm, and provide minimal background information necessary for what follows. Pokémon is a game series first released by Nintendo Inc. in 1996 and has become very popular worldwide since then. In this game series, players collect and train fictional creatures called Pokémon, which is a truncation of [**poke**tto **mon**sutaa] 'pocket monster.' One feature that becomes crucial in what follows is that some Pokémon characters undergo evolution, and when they do so, they generally become larger, heavier and stronger. When they evolve, moreover, they are called by a different name; for instance, *Iwaaku* becomes *Haganeeru*.

Kawahara et al. (2018) show that when we systematically examine their names from the perspectives of sound symbolism, post-evolution characters have longer names than pre-evolution characters. They call this "the longer the stronger principle," and attribute this observation to one of the previously known sound symbolic principles, "the iconicity of quantity" (Haiman 1980,

1984), in which larger quantity is expressed by longer phonological material. They also show that post-evolution Pokémon characters are more likely to have names with voiced obstruents than pre-evolution characters. This observation is likely to be related to the observation that in Japanese, voiced obstruents often sound symbolically denote large quantity and/or strength (Hamano 1998; Kawahara 2017). Both of these sound symbolic effects can be seen in the pair *Iwaaku* vs. *Haganeeru*: evolved *Haganeeru* has five moras and contains a voiced obstruent [g], while unevolved *Iwaaku* has only four moras and no voiced obstruents. The experiment below examines these two sound symbolic effects in further detail by way of experimentation.

1.4 The goals of this paper: An interim summary

The take-home messages of the current study are summarized in (1)-(3).

(1) Wug-shaped curve
   a. We observe a wug-shaped curve in sound symbolism.

(2) Cumulativity
   a. Sound symbolism shows counting cumulativity.
   b. Sound symbolism shows ganging-up cumulativity.
   c. These two types of cumulativity can co-exist within a single sound symbolic system.

(3) Theoretical implications
   a. A wug-shaped curve is a natural consequence of the MaxEnt grammar.
   b. Stochastic OT can also model the observed patterns, but it requires additional tweaks.
   c. Sound symbolic mappings can be modeled using the same mechanism as phonological generalizations.
   d. Phonological theories and research on sound symbolism can inform one another.

2 Methods

One precursor of the current experiment is Kawahara et al. (2020b), who report on a judgment experiment on the strengths of Pokémon move names (moves are what Pokémons use when they battle with each other). Kawahara et al. (2020b) manipulated mora length from 2 moras to 7 moras, and showed that the longer the nonce names, the stronger they were judged to be. They also manipulated the presence/absence of a voiced obstruent placed at word-initial position, and found that nonce move names with voiced obstruents were judged to be stronger. Their results are reproduced in Figure 3, which instantiate both counting cumulativity (the effect of mora count) and ganging-up cumulativity (the additive effects of the two factors). However, their experiment is not

suited to address the question of whether we observe wug-shaped curves in sound symbolism, nor were their results amenable to a MaxEnt analysis, because the judged values were continuous—what we need instead is the probability distributions of categorical outcomes.

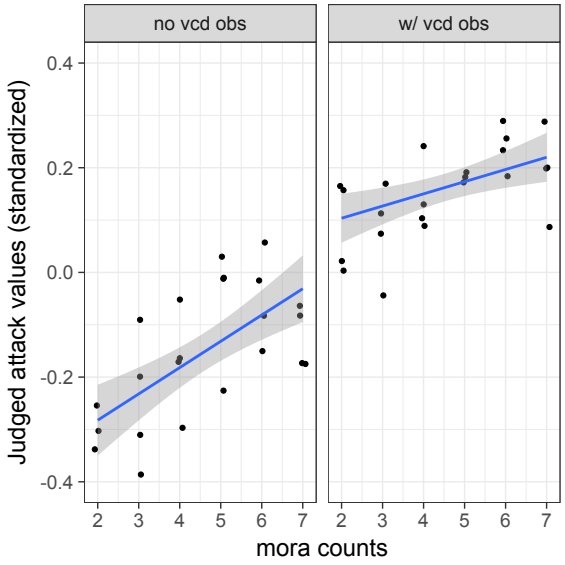


Figure 3: The effects of mora counts and word-initial voiced obstruents on judged attack values in nonce Pokémon move names. The y-axis shows standardized judged attack values, which are continuous. Adapted from Kawahara et al. (2020b), their Figure 4.

The current study builds upon Kawahara et al. (2020b), but in order to obtain a binary categorical response, the experiment asked the participants to judge whether each stimulus name is better suited for a pre-evolution character or post-evolution character. To obtain more reliable estimates of each condition, more items were included for each condition. As we will see below, moreover, the current experiment collected responses from many more participants.

2.1 Stimuli

Table 1 lists the stimuli, in which dots represent mora boundaries. Building on the two previous studies reviewed above (Kawahara et al. 2018, 2020b), the experiment manipulated two variables: mora counts and the presence of a voiced obstruent placed at word-initial position. The mora count was varied in order to examine counting cumulativity, and relatedly, to examine whether it would result in a sigmoid curve. Mora counts varied from 2 to 6, corresponding to minimum and maximum lengths for Pokémon names. This experiment manipulated mora counts rather than segment counts or syllables counts, because mora counts is what was identified as most important in the previous studies (Kawahara et al. 2018, 2020b; Shih et al. 2019), and moreover, the mora is demonstrably the most psycholinguistically salient prosodic counting unit in Japanese (Otake et al.

1993). The perturbing factor (see §1.1) was the presence/absence of a voiced obstruent placed at the name-initial position.

As shown in Table 1, six items were included in each cell. All the names were created using a nonce name generator, which randomly combines Japanese moras to create new names.[4] This random generator was used to preclude the potential bias by the experimenter to select the stimuli that were likely to support their hypothesis prior to the experiment (Westbury 2005). All voiced obstruents appeared word-initially, because a previous study shows that the strength of sound symbolic values of voiced obstruents in Japanese may vary depending on different word positions (Kawahara et al. 2008). No geminates, long vowels, or coda nasals appeared anywhere in the stimuli; i.e. all syllables were open syllables. Moreover, because of its potentially salient sound symbolic values, such as cuteness (Kumagai 2019), [p] was excluded in the current stimulus items.

[4] `http://sei-street.sakura.ne.jp/page/doujin/site/doc/tool_genKanaName.html` (last access, August 2020).

Table 1: The list of stimuli. Dots represent mora boundaries.

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

2.2 Procedure

The experiment was distributed as an online experiment using SurveyMonkey (`https://www.surveymonkey.com`: last access, August 2020). Within each trial, participants were given one nonce name at a time and asked to judge whether that name is better for a pre-evolution character or a post-evolution character, i.e., the task was to make a binary decision. The stimuli were presented in the Japanese *katakana* orthography, which is used to represent real Pokémon

names. The participants were asked to base their decision on their intuition, without thinking too much about "right" or "wrong" answers. The order of the stimuli was randomized for each participant.

2.3 Participants

The experiment was advertised on a Pokémon fan website.[5] A total of 857 participants completed the experiment over a single night. Since some previous Pokémonastics experiments were advertised on the same website (e.g. Kawahara et al. 2020a), 124 of them reported that they either had participated in another Pokémonastics experiment or studied sound symbolism before. Three participants were non-native speakers of Japanese. After excluding the data from these speakers, the data from the remaining 730 participants entered into the subsequent analysis.

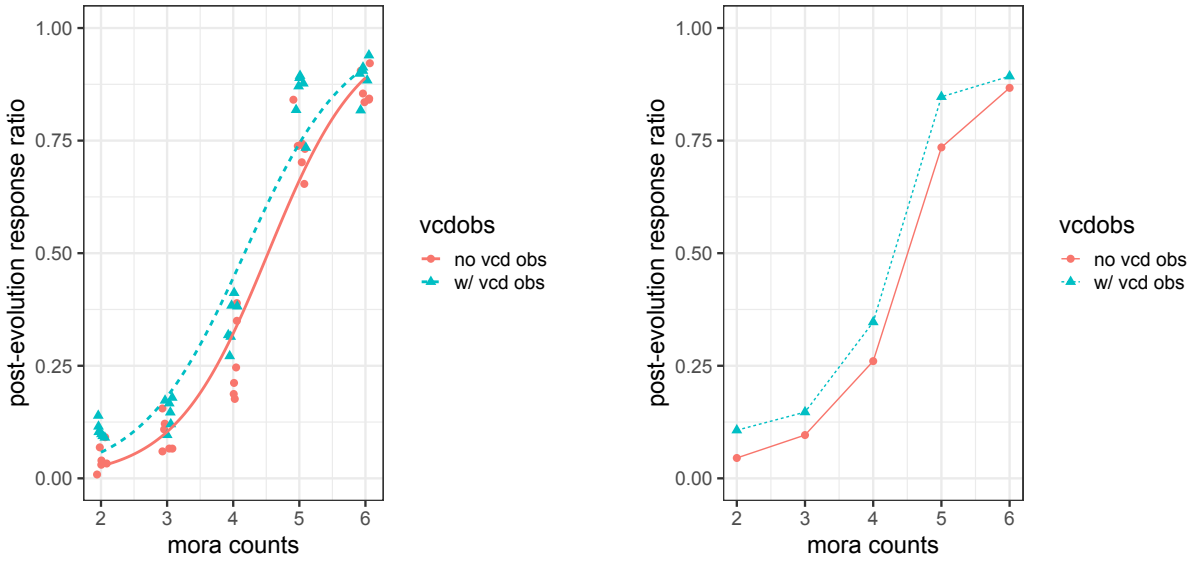
2.4 Analysis

For statistical analysis, a logistic linear mixed effects model was fit with the response (pre-evolution vs. post-evolution) as the dependent variable (Jaeger 2008). The fixed independent variables include the mora count and the presence of a voiced obstruent as well as its interaction. The mora count was centered, because it is a continuous variable (Winter 2019). Random factors were participants and items. The model with maximum random structure with both slopes and intercepts (Barr et al. 2013) did not converge; hence a simpler model with only random intercepts was interpreted.

3 Results

Figure 4 shows the results. Figure 4(a) plots "post-evolution response ratios" for each item, averaged over all the participants. The items for the condition with voiced obstruents are shown with green triangles and the items for the condition without voiced obstruents are shown with red circles. A logistic curve is superimposed for each voicing condition (green dotted line = the condition with a voiced obstruent; red solid line = the condition without a voiced obstruent).

[5] `http://pokemon-matome.net` (last access, August 2020)



(a) (b)

Figure 4: (a) The by-participant averages for each item. The items with a voiced obstruent are shown with green triangles and those without a voiced obstruent are shown with red circles. To avoid overlap, the points were horizontally jittered by 0.1. Logistic curves are superimposed—the green dotted line represents the condition with a voiced obstruent, whereas the red solid line represents the condition without a voiced obstruent. (b) The line-plots with grand averages for each voicing condition.

These results look like wug-shaped curves, schematically illustrated in Figure 2, consisting of two sigmoid curves separated from each other on the horizontal axis. The relationships between the x-axis and y-axis appear to be closer to sigmoid curves than to a linear function, in that the slope is evidently steepest in the middle range. This observation is also clear in Figure 4(b), which illustrates the overall pattern by presenting grand averages for each condition—this analysis does not presuppose that sigmoid curves would fit the data points well. The slopes are rather steep between the 3-mora condition and the 5-mora condition. On the other hand, the slopes are not very steep between the 2-mora condition and the 3-mora condition or between the 5-mora condition and 6-mora condition. As Hayes (2020: 3) puts it, "certainty is evidentially expensive"—it requires very strong evidence to be certain that a particular name is for a pre-evolution character or for a post-evolution character. A more elaborate defense of using a wug-shaped curve to fit the data is provided in §5.3, once we have developed a full analysis of the current data using MaxEnt.

The model summary of the linear mixed effects model appears in Table 2. It shows that the two main factors are statistically significant: the longer the names, the more likely that they are judged to be better names for post-evolution characters, and the names with a voiced obstruents are judged to be better suited for post-evolution characters. The interaction between the two factors was not

significant.

Table 2: The summary of the logistic linear mixed effects model.

	β	$s.e.$	z	p
intercept	-0.76	0.11	-7.14	$< .001$
mora count	1.43	0.08	19.04	$< .001$
vcd obs	0.63	0.15	4.26	$< .001$
mora count $\times$ vcd obs	-0.135	0.11	-1.42	$n.s.$

4 Discussion

The effect of mora count instantiates a case of counting cumulativity (Jäger & Rosenbach 2006), in that each mora count additively contributes to the post-evolution judgment. This effect of mora count is evident both when there is a voiced obstruent name-initially and when there is not. The effect of a voiced obstruent in name-initial position manifests itself as a shift between the two sigmoid curves. The effects of mora counts and a voiced obstruent together instantiate a case of a ganging-up cumulativity (Jäger & Rosenbach 2006)—both factors contribute to the judgment of evolvedness. Overall, the current results show that counting cumulativity and ganging-up cumulativity can coexist within a single sound symbolic system. This conclusion is compatible with the results of an artificial language learning experiment on phonotactic learning reported by Breiss (2020), as well as some phonological observations made by Hayes (2020), McPherson & Hayes (2016), and Zuraw & Hayes (2017). See also Breiss (2020) and Kawahara (2020a) for summaries of cumulative effects in phonological alternations as well as wellformedness judgment patterns of surface phonotactics.

While the results in Figure 4 seem to instantiate a clear case of "wug-shaped" curves, one may wonder if the results could have been otherwise. The answer is positive, as there were multiple alternative patterns that could have arisen from the current experimental design. For example, the effect of mora counts could have been cumulative but linear, instead of being sigmoidal. Indeed, the effect of mora counts in the existing Pokémon names actually looks to be more linear than sigmoidal (see the Appendix).

Or, the results could have been non-cumulative. It could have been the case, for example, that there was a "length threshold" such that any names shorter than that threshold were judged to be pre-evolution names, but the actual results did not follow such a pattern. Neither was it the case that the presence of a voiced obstruent made the names post-evolution names 100% of the time. Instead, both mora counts and voiced obstruents gradiently increased the probabilities of each

name being judged as a post-evolution name.[6] This point is related to another important aspect of sound symbolism; namely, its stochastic nature (Dingemanse 2018; Kawahara et al. 2019). At a more general level, Gigerenzer & Gaissmaier (2011) discuss several cases in which when people make decisions, they take "a fast and frugal decision heuristics"—they take into account the most important information only and disregard other information (just as OT with strict domination would do). If people had applied such a fast-and-frugal heuristics decision making approach in the current experiment, the results would have been neither stochastic nor cumulative.

Finally, the stochastic nature of sound symbolism draws a parallel to a growing body of evidence that many if not all phonological generalizations have to be stated in a stochastic or probabilistic way; e.g., some structures tend to be preferred over others, and some alternations occur with different probabilities in different environments (see §1.1). The current results thus reveal an intriguing parallel between phonological patterns and sound symbolic patterns.

5 A MaxEnt analysis

The experimental results reported in §3 seem to instantiate a wug-shaped curve, a quantitative signature of the MaxEnt grammar model; the results thus appear to lend support for this grammatical model from the perspective of sound symbolism. To provide more concrete support for the MaxEnt grammar model, this section develops an analysis of the experimental results using MaxEnt, equipped with the sorts of constraints that have been used in the Optimality Theoretic tradition (Prince & Smolensky 1993/2004).[7] One fundamental idea behind this analysis is that sound symbolic connections—mapping between sounds and meanings—can be understood as involving essentially the same mechanism as the phonological input-output mappings (Kawahara et al. 2019; Kawahara 2020a). The model deploys the sort of constraints that are familiar from the OT tradition

[6] A very small subset of the participants (17 out of 730, ca. 2%) showed categorical patterns with respect to the effects of mora length, in that they assigned names of all mora lengths to either pre-evolution characters or post-evolution characters 100% of the time and showed no intermediate responses. A MaxEnt grammar, which is developed below in §5, can yield (near-)categorical results when the weight of the relevant constraint is very high. No participants judged names with voiced obstruents to be post-evolution character names 100% of the time.

[7] Wug-shaped curves are also predicted under some versions of Noisy Harmonic Grammar, depending upon precisely how noise is added to the calculation of harmony scores. To be specific, Noisy Harmonic Grammar fails to generate sigmoid curves if noise can be added to the weight of each constraint, as noise is multiplied by the number of violations. It is able to generate sigmoid curves, as long as noise is added after harmony values are calculated (see Hayes 2017, 2020; McPherson & Hayes 2016 for detailed discussion on this point). Since wug-shaped curves do not themselves distinguish MaxEnt from the latter implementation of Noisy Harmonic Grammar (both are versions of Harmonic Grammar), I focus on the former framework.

Another quantitative framework that can model stochastic generalizations in phonology is the inverted-exponential model proposed by Guy (1991), which derives different probabilities by positing that an optional phonological rule can apply different numbers of times in different morphological conditions. I set this analysis aside in the paper for three reasons: (i) it is not clear how a rule-based approach can be used to model sound symbolic connections (Kawahara 2020a), (ii) the current probabilistic patterns have nothing to do with morphological differences, and (iii) this exponential model does not derive sigmoid curves (McPherson & Hayes 2016).

(Prince & Smolensky 1993/2004). To underscore the parallel between phonological analyses and the analysis of sound symbolism developed in this paper, I adapt a particular formalism that has been used to define constraints in the OT research tradition (McCarthy 2003).

5.1 A brief review of MaxEnt

This section briefly reviews how MaxEnt works in the context of linguistic analyses. This section repeats section 3 of Kawahara 2020a; readers who are familiar with MaxEnt can safely skip this section. The MaxEnt grammar is similar to OT (Prince & Smolensky 1993/2004) in that a set of candidates is evaluated against a set of constraints. Unlike OT, however, constraints are weighted rather than ranked. Consider a toy example in (4). The set of candidates that are evaluated are listed in the leftmost column. The top row lists the set of constraints that are relevant, and each constraint is assigned a particular weight. The tableau shows the violation profiles of each constraint—which candidate violates which constraints how many times.

(4) A toy example tableau of a MaxEnt analysis

	Constraint A Weight = 3	Constraint B Weight = 2	Constraint C Weight = 1	H-score	eHarmony	*Z*	predicted percentages
Candidate 1	1			1*3=3	e^{-3}= 0.0498	0.0565	88
Candidate 2		2	1	2*2+1*1=5	e^{-5}= 0.0067	0.0565	12

Based on the constraint violation profiles, for each candidate x, its Harmony score (H-score(x)) is calculated using the formula in (1):

$$\text{H-score}(x) = \sum_{i}^{N} w_i C_i(x) \;\; (N \text{ is the number of the constraints}) \tag{1}$$

where w_i is the weight of the i-th constraint, and $C_i(x)$ is the number of times candidate x violates the i-th constraint. For example, Candidate 2 in the tableau (4) violates Constraint B twice and Constraint C once; its H-score is therefore $2 * 2 + 1 * 1 = 5$.

The H-scores are negatively exponentiated (eHarmony, e^{-H} or $\frac{1}{e^H}$: Wilson 2014), which is proportional to the probability of each candidate. Intuitively, the more constraint violations a candidate incurs, the higher the H-score, and hence the lower the eHarmony (e^{-H}). Therefore, more violations of constraints lead to lower probability of that candidate. The eHarmony values are relativized against the sum of the eHarmony values of all the candidates, which is referred to

as Z:

$$Z = \sum_{j}^{M} (e^{-H})_j \text{ (M is the number of the candidates)} \quad (2)$$

In the example in (4), Z is $0.0498 + 0.0067 = 0.0565$. The predicted probability of each candidate $x_j, p(x_j)$, is $\frac{eHarmony(x_j)}{Z}$.

5.2 The MaxEnt analysis of the current results

Like most phonological analyses in OT and other related frameworks, the current MaxEnt analysis of sound symbolism consists of inputs, outputs and constraints that evaluate the mapping between these two levels of representations. The inputs are phonological forms, and the outputs are their sound symbolic meanings, here either pre-evolution character names or post-evolution character names. The set of constraints deployed in the current analysis is given in (5).[8] These constraints essentially correspond to the markedness constraints in OT in that they evaluate the wellformedness of output structures. The definition of the constraints follows the format proposed by McCarthy (2003).

(5) Constraints deployed in the current analysis

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 prevents long names from being used for pre-evolution characters. This constraint is a formal expression of "the longer the stronger principle" (Kawahara et al. 2018) or "the iconicity of quantity" (Haiman 1980, 1984). The constraint is a single gradient/scalar constraint (Hsu & Jesney 2017; McPherson & Hayes 2016) in that it is a reflection of one principle whose violations can be assessed on a numerical scale.[9] This constraint corresponds to the scalar con-

[8] If one is concerned that notions like "pre-evolution" and "post-evolution" are too language/culture-specific to be mentioned in the OT-style constraints, which are generally taken to be universal, they can each be replaced with "small entity" and "large entity," since Pokémon characters generally become larger after evolution. Size, together with shape, is one semantic dimension that is most clearly signaled by sound symbolism across many languages (Sidhu & Pexman 2018).

[9] The use of a scalar constraint is not new, even in the OT research tradition. The HNUC constraint proposed by Prince & Smolensky (1993/2004) can be understood as this type of constraint, although Prince & Smolensky did not use actual numbers. See, for example, de Lacy (2006) and Gouskova (2004) who offer extensive discussion on how various phonological scales should be formally captured in OT, although they deploy a family of constraints instead of a single scalar constraint (see §6). See McCarthy (2003) for a review of gradient constraints in OT and criticisms against them. See also McPherson & Hayes (2016: 149) for other examples of scalar constraints that have been used

straint S that was used to schematically illustrate the wug-shaped curves in §1.1. The second constraint is a formal expression of the preference that names with voiced obstruents should be used for post-evolution character names, and this corresponds to the perturber constraint P that was used in §1.1. The last constraint is a *STRUC constraint (Prince & Smolensky 1993/2004) that penalizes post-evolution character names in general, which corresponds to the binary constraint B discussed in §1.1. For the current analysis we need this *STRUC constraint, because there has to be some constraint that favors pre-evolution character names. All of these constraints are statistically motivated by a log-likelihood ratio test, to be presented below in Table 3.

Hayes (2020) recommends that we conceive of constraints as evidence to make a decision about which candidate to choose. The constraints posited in (5) do precisely this: the first two constraints offer sound symbolic evidence to decide on post-evolution names when the candidates are long (*LONGPRE) or when they contain a voiced obstruent (*VCDPRE). The last constraint helps us to decide on a pre-evolution name in general. The weights associated with each constraint reflect the strengths, or cogency, of each evidence.

The MaxEnt tableaux for all types of inputs are shown in (6). The leftmost column shows each phonological form, and the second column shows how each phonological form is mapped onto two meanings: pre-evolution character names vs. post-evolution character names. The constraint violation profiles are shown in the third-fifth columns. The observed percentages of each condition, shown in the rightmost column, 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 these constraints were calculated using the Solver function of Excel (see Supplementary Material 1). The weights that were obtained by this analysis are shown at the top row of the tableaux. These weights, together with the constraint profiles, allow us to calculate H-scores, eHarmony scores, and predicted percentages, using the procedure reviewed in §5.1.

in linguistic theories.

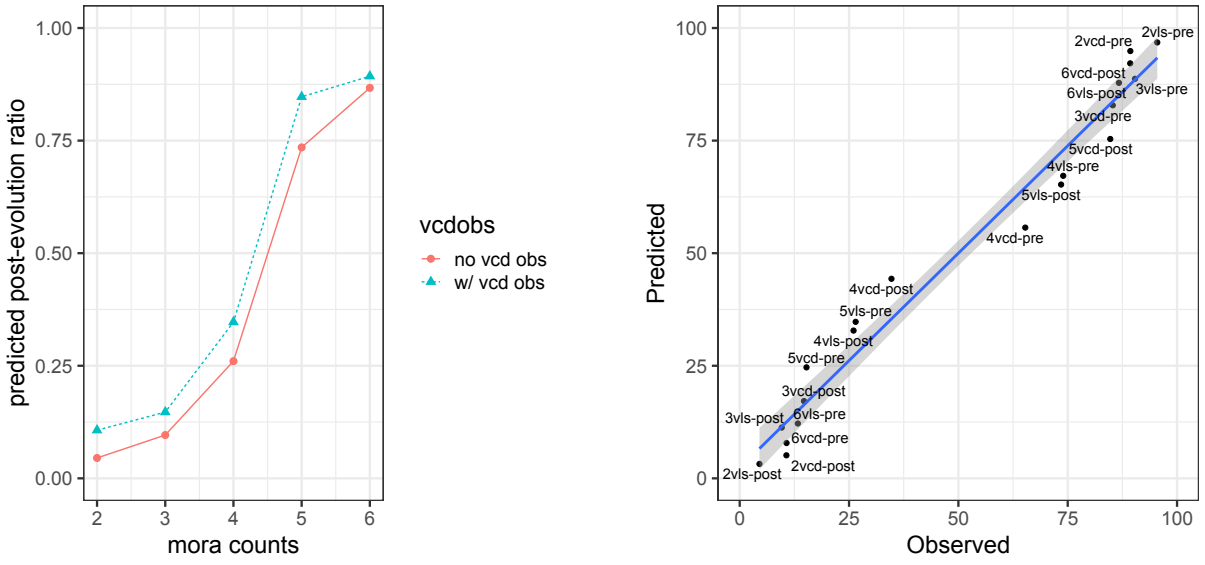
(6) The MaxEnt Tableaux

		w = 1.35	w = 0.49	w = 6.10				
Input	Output	*LongPre	*VcdPre	*Post	H-score	eHarmony	Predicted	Observed
2 moras, vls	Pre	2			2.69	0.068	96.80	95.48
	Post			1	6.10	0.002	3.20	4.52
3 moras, vls	Pre	3			4.04	0.018	88.72	90.39
	Post			1	6.10	0.002	11.28	9.61
4 moras, vls	Pre	4			5.38	0.005	67.18	73.97
	Post			1	6.10	0.002	32.82	26.03
5 moras, vls	Pre	5			6.73	0.001	34.76	26.51
	Post			1	6.10	0.002	65.24	73.49
6 moras, vls	Pre	6			8.08	0.0003	12.18	13.29
	Post			1	6.10	0.002	87.82	86.71
2 moras, vcd	Pre	2	1		3.18	0.042	94.89	89.32
	Post			1	6.10	0.002	5.11	10.68
3 moras, vcd	Pre	3	1		4.53	0.011	82.84	85.30
	Post			1	6.10	0.002	17.16	14.70
4 moras, vcd	Pre	4	1		5.87	0.003	55.68	65.30
	Post			1	6.10	0.002	44.32	34.70
5 moras, vcd	Pre	5	1		7.22	0.001	24.64	15.27
	Post			1	6.10	0.002	75.36	84.73
6 moras, vcd	Pre	6	1		8.57	0.0002	7.84	10.71
	Post			1	6.10	0.002	92.16	89.29

The observed and the predicted values are very close to each other. To visualize the success of this MaxEnt analysis, Figure 5(a) shows the probability curves that are predicted by this MaxEnt model (cf. the actual results of the experiment in Figure 4). Figure 5(b) plots the correlation between the observed and the predicted values, which shows a good fit between the two measures.

One general advantage of MaxEnt is that it allows us to assess the necessity of each constraint using a well-established statistical method, i.e., a log-likelihood ratio test (see e.g. Wasserman 2004 and Winter 2019; see also Breiss & Hayes 2020, Hayes et al. 2012 and Hayes & Jo 2020 for applications of this test in linguistic analyses). We can do so by comparing two grammatical models—for the current analysis, 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 two-constraint models. We then compare their log-likelihood values by taking their ratios, which would tell us 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 shown in Table 3, which demonstrates that



(a) (b)

Figure 5: (a) The probability curves predicted by the MaxEnt analysis. (b) The correlation between the observed and the predicted percentages obtained from the MaxEnt analysis.

all three constraints are statistically motivated to explain the data—in other words, each of the constraints plays a role in the explanation of the data, in addition to what is explained by the other two constraints (see the Appendix of Breiss & Hayes 2020).

Table 3: The results of the log-likelihood ratio tests. The log-likelihood of the best fitting model with the three constraints was -432.3. See Supplementary Material 1.

	Δlikelihood	$\chi^2(1)$	p
*LONGPRE	249.88	499.77	$< .001$
*VCDPRE	4.10	8.20	$< .01$
*POST	256.65	513.30	$< .001$

Next, a more complex model was tested with a fourth constraint that represents the interaction term between *LONGPRE and *VCDPRE, which is equivalent to the locally conjoined version of these two constraints (cf. Shih 2017). The result shows that addition of this constraint did not improve the model fit at all. The Solver actually assigned 0 weight to the conjoined constraint, even when we allow its weight to be negative. 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 the MaxEnt model itself, obviating the need to posit a specific constraint to capture the interaction between the two factors (see Zuraw & Hayes 2017).

5.3 MaxEnt and wug-shaped curves revisited

Having fully developed the MaxEnt analysis, we can now address one general question regarding wug-shaped curves: whether it is possible to objectively assess if a given data is best fit with a wug-shaped curve. To reiterate its definition, a wug-shaped curve, generated by MaxEnt, is a mathematical object consisting of two identical sigmoid curves separated on the x-axis. It thus has three essential features: (1) it consists of two sigmoid curves, (2) the two curves are identical, and (3) they are separate curves. No real data would perfectly fit this mathematical definition, because they involve some natural variability. Therefore, the question boils down to the issue of how well wug-shaped curves fit the observed data.

Testing whether the two curves are separated on the x-axis or not is relatively straightforward: it can be assessed by examining the effect of the perturber. In the current analysis, the perturber corresponds to the constraint *VCDPRE, which was significant in the MaxEnt analysis developed in §5.2. Whether the two curves are identical or not can be addressed by examining the interaction term, because the interaction term represents whether—and how much—the slope should be adjusted from one curve to the other (Winter 2019: 138). If the interaction term between *LONGPRE and *VCDPRE were significant, we could have reasonably concluded that the two curves were not identical to each other. Since the inclusion of the interaction term did not improve the fit of the model, we cannot reject the null hypothesis that the two curves are identical.

In reality, however, it is improbable to obtain two curves that are literally identical, because again, the data in the real world is subject to natural variability. To what extent we allow the two sigmoid curves to be different is a matter that should be examined by way of empirical investigation rather than something that should be determined *a priori*. Two similar yet non-identical sigmoid curves would result in a slightly "distorted" wug-shaped curve. This issue, however, is not just about two lines on a graph; it must instead be understood as an issue of whether we should allow interaction terms—or conjoined constraints—to play a substantial role in the MaxEnt grammar. McPherson & Hayes (2016) and Zuraw & Hayes (2017) posit no interaction terms for their analyses, while Shih (2017), on the other hand, argues that constraint conjunction is required even in the MaxEnt grammar. More quantitative studies are necessary to settle this issue.

Finally, most challenging is how to conclude whether the pattern is best modeled using a sigmoid curve or not, which concerns the general issue of which mathematical function to use to fit the data. One useful heuristic is to make use of log-likelihood, the log probability of the observed data being generated by the model (see Zuraw & Hayes 2017 who use this measure to compare different linguistic models). For example, fitting linear functions to the current data yields $p(evolved) = -0.51 + 0.228 \times \text{mora} + 0.067 \times \text{vcd obs}$. The log-likelihood of this linear model is -501.0,[10] which is worse than the sigmoidal MaxEnt model, which has log-likelihood of -432.3.

[10]See Supplementary Material 3. This model predicted that bimoraic forms without voiced obstruents should be

Log-likelihood represents summed log probabilities, so they are always negative. The higher the log-likelihood (i.e. the closer it is to 0), the more likely that the data is generated by the model (i.e. the data is better fit by the model).

However, relying on log-likelihood alone does not allow us to conclude that the sigmoid function is *the* function that underlines the actual data. In principle, we can posit a mathematical function with high complexity to achieve the perfect fit to the data; in fact, a function that fits the data perfectly would intersect every data point. However, such functions would be non-restrictive, non-predictive, and non-generalizable; i.e., they suffer from the general problem of overfitting (Good & Hardin 2006). In order to balance between the goodness of the fit to the data and model complexity, additional statistical measures, such as AIC (Akaike Information Criterion: Akaike 1973), which take into account the number of free parameters, may prove to be useful (see Shih 2017 as well as §6).

Comparing the different sorts of mathematical functions, of which there are many, is beyond the scope of the present paper; in general, however, the choice of mathematical functions to fit linguistic data should be guided by cross-linguistic quantitive observations. For now, we are reasonably confident that mathematical functions generated by MaxEnt are suited to model cross-linguistic quantitative patterns, as reviewed in §1.1.

Finally, to conclude this discussion, the current MaxEnt analysis makes specific predictions for forms that contain two voiced obstruents. One of the experiments reported by Kawahara & Kumagai (2021) shows that nonce names with two voiced obstruents are more likely to be judged as post-evolution characters names than nonce names with one voiced obstruent. This result suggests that the effects of voiced obstruents are cumulative, just like the effects of mora count. The definition of *VCDPRE formulated in (5) actually predicts this cumulative behavior, since forms with two voiced obstruents are assigned two violation marks when they are mapped onto the pre-evolution character. Since the weights of the constraints are already calculated and since the constraint violation profiles are known, the current MaxEnt model makes specific quantitative predictions.[11] These predictions of the MaxEnt grammar are illustrated in Figure 6, which instantiates a "stripey wug" consisting of three sigmoid curves (Bailey 1973; Hayes 2020; McPherson & Hayes 2016; Zimmermann 2017; Zuraw & Hayes 2017). While the current experiment was limited to those items which contained only one voiced obstruent, these predicted values can be tested in future experimentation.

post-evolution characters "-5.4% of the time," which is impossible, instantiating a general problem of fitting a linear function to probability distributions (Jaeger 2008). I simply replaced this value with 1×10^{-6}. This points to one strength of MaxEnt—since harmony is negatively exponentiated, it never yields probabilities below zero.

[11] This analysis assumes that the sound symbolic values of voiced obstruents are of equal strength between word-initial positions and word-medial positions, which may be an oversimplification, as Kawahara et al. (2008) show that voiced obstruents in initial positions may evoke stronger images. Word-internal voiced obstruents may not thus increase post-evolution responses as much as word-initial voiced obstruents.

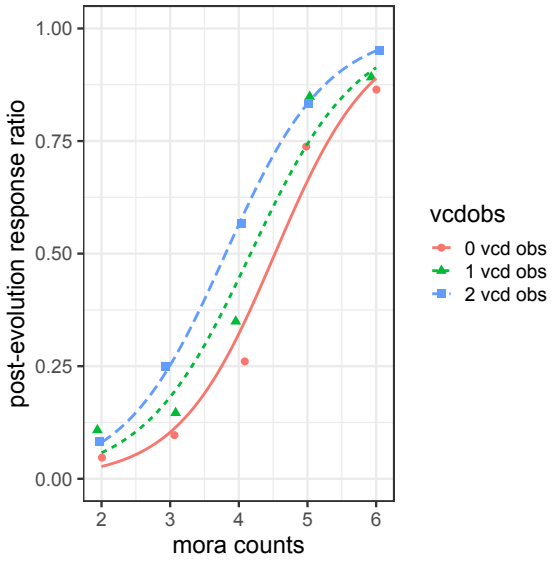


Figure 6: Predictions of the current MaxEnt model regarding forms with two voiced obstruents, instantiating a "stripey wug." See Supplementary Material 2.

This analysis serves to illustrate one strength of the MaxEnt grammar with explicit constraint formulation: it makes specific quantitative predictions about forms that we have not seen yet. As discussed above, choosing a relatively simple model avoids over-fitting and is more likely to generate good predictions for new data.

5.4 Some notes on MaxEnt and logistic regression

Finally, I note at this point that MaxEnt is mathematically equivalent to a (multinomial) logistic regression (see in particular Jurafsky & Martin 2019: chapter 5, as well as Breiss & Hayes 2020 and Shih 2017). A mixed effects logistic regression analysis was reported in §3 as a means to test the experimental results without any particular linguistic theories/analyses in mind. On the other hand, a MaxEnt analysis was developed in this section as an explicit, formal analysis within generative grammar to model the knowledge that may underlie the patterns that were identified in the experiment. In order to emphasize that this MaxEnt analysis is a generative phonological analysis, I deployed a particular constraint schemata that is proposed in the OT research tradition (McCarthy 2003).

The fact that logistic regression, a general statistical tool, is so well-suited to model linguistic patterns is an interesting and thought-provoking observation. As the Associate Editor notes, one way to understand this convergence is that since MaxEnt (or logistic regression) is demonstrably a "useful method for discerning...causes in data in general, it makes sense that something akin to logistic regression might be used by children to discern causes (embodied as highly weighted

constraints) in the grammar of the language they learn." In this view, UG deploys some form of logistic regression to learn patterns in the ambient data (see in particular Hayes & Wilson 2008, as well as Smolensky 1986).

Another way to understand MaxEnt within the current phonological research is to consider it as a stochastic extension of Optimality Theory (Prince & Smolensky 1993/2004; see also Breiss & Hayes 2020), which invites an interesting question of whether UG can be reduced to a domain-general statistical tool. Providing a full fledged answer to this question is beyond the scope of this paper. However, even if the mapping between two linguistic representations is mediated by a general statistical device, there can be other aspects of UG that remain domain-specific; these include, but are most likely not limited to, (i) the content of the constraints (i.e. CON), (ii) what the set of the vocabulary that this constraint set refers to (e.g. distinctive features such as [+son] and [+voiced] as well as the levels in prosodic hierarchy such as moras and syllables), (iii) how constraint violations can and cannot be assessed (e.g. whether constraints can reward a candidate), and (iv) whether constraints can be conjoined or not, and if yes, to what extent (e.g. Coetzee 2016; Coetzee & Kawahara 2013; Crowhurst 2011; de Lacy 2006; McCarthy 2003; Potts & Pullum 2002 among many others). Restricting CON may be necessary to explain cases in which speakers behaviors substantially diverge from what is predicted by the statistical patterns in the lexicon (e.g. Becker et al. 2011; Garcia 2019; Jarosz 2017). Additionally, UG may also function as particular biases toward, for example, phonetically natural patterns, which can be formalized in the MaxEnt framework in terms of biases on constraint weights (Hayes et al. 2009; Hayes & White 2013; Wilson 2006). In short, UG can be a meta-theory of constraints. Since MaxEnt allows us to statistically access the necessity of each constraint by way of log-likelihood tests, it may prove to be a useful tool to explore what CON consists of in a quantitively rigorous manner (Shih 2017).

6 Analyses with Stochastic Optimality Theory

While Hayes (2020) as well as Zuraw & Hayes (2017) have shown that patterns with wug-shaped curves cannot be modeled well with Stochastic OT (Boersma 1998; Boersma & Hayes 2001), this section reports several attempts to fit a Stochastic OT model to the current data. In Stochastic OT, each constraint is assigned a particular ranking value, which is perturbed by a Gaussian noise at each time of evaluation. Each evaluation is computed just as in Classic OT with strict domination, predicting a single winner per each evaluation trial. The probability distributions of variable outputs are calculated over multiple evaluation cycles.

To analyze the current experimental results using Stochastic OT, first, the same data structure that was used for the MaxEnt analysis in (6) was fed to OTSoftware (Hayes et al. 2014) with the Gradual Learning Algorithm as the learning algorithm. The initial ranking values of all constraints

were set to be 100 (the default value). The initial plasticity and the final plasticity were set to be 0.01 and 0.001, respectively. There were 1,000,000 learning trials, and the grammar was tested for 1,000,000 cycles in order to get the predicted probability distribution. The results of all the learning simulations presented in this section are available as Supplementary Material 4.

This learning simulation resulted in the following ranking values: *LONGPRE = 99.6, *VCDPRE = 98.1, *POST = 100.4. All the constraints were active in at least one of the evaluation trials. The problem with this analysis using Stochastic OT is that it was not able to model the effects of mora counts at all; indeed, Stochastic OT does not handle counting cumulativity effects well in general (Hayes 2020; Jäger 2007). For all the conditions without voiced obstruents, regardless of the mora counts, post-evolution candidates were predicted to win 40.0% of the time and pre-evolution candidates were predicted to win 60.0% of the time. For all the conditions with voiced obstruents, post-evolution characters were predicted to win 46.6% of the time, whereas the pre-evolution characters were predicted to win 53.4% of the time. Stochastic OT was thus able to model the effect of voiced obstruents (40% vs. 46.6%), which seems to reflect the actual observed post-evolution response values averaged across all the mora length conditions (40.1% vs. 46.8%). However, it was unable to learn the effects of mora counts.

The failure to model the counting cumulativity effects of mora counts is due to the fact that Stochastic OT is no different from Classic OT (Prince & Smolensky 1993/2004) at each time of evaluation. OT does not distinguish between, for example, one violation mark vs. two violation marks (* vs. **) and one violation mark vs. four violation marks (* vs. ****). Therefore, if *POST dominates *LONGPRE at a particular time of evaluation, then the pre-evolution candidate is predicted to win at that particular time of evaluation, no matter how many violations of *LONGPRE the pre-evolution candidate incurs. Similarly, if *LONGPRE dominates *POST, the post-evolution candidate wins no matter how long the pre-evolution candidate is. The number of violations simply does not add up in Classic OT or Stochastic OT, because of strict domination. For these reasons, it was not able to account for the counting cumulativity effects of mora counts.

This problem can be (partially) remedied by splitting up *LONGPRE into a set of separate constraints which each penalizes a pre-evolution name with a particular mora length; i.e. *LONGPRE3MORA, *LONGPRE4MORA, *LONGPRE5MORA, and *LONGPRE6MORA (see footnote 21 of McPherson & Hayes 2016 as well as Boersma 1998, de Lacy 2006, and Gouskova 2004). A new learning simulation was run with the same parameter settings. With the expanded set of constraints, it learned the following values: *LONGPRE3MORA = 97.2, *LONGPRE4MORA = 99.7, *LONGPRE5MORA = 103.7, *LONGPRE6MORA = 103.2, *VCDPRE = 98.7, *POST =101.6. When we plot the predicted values based on these ranking values, the Stochastic OT analysis created two separate curves for the two voicing conditions, as shown in Figure 7. However, these curves formed the "open jaw" pattern, in which we observe the convergence of the two curves at

one end and divergence between the two curves at the other end, and the difference between the two curves grow monotonically toward the left end (compare this pattern with the result of the MaxEnt analysis, Figure 5(a)).

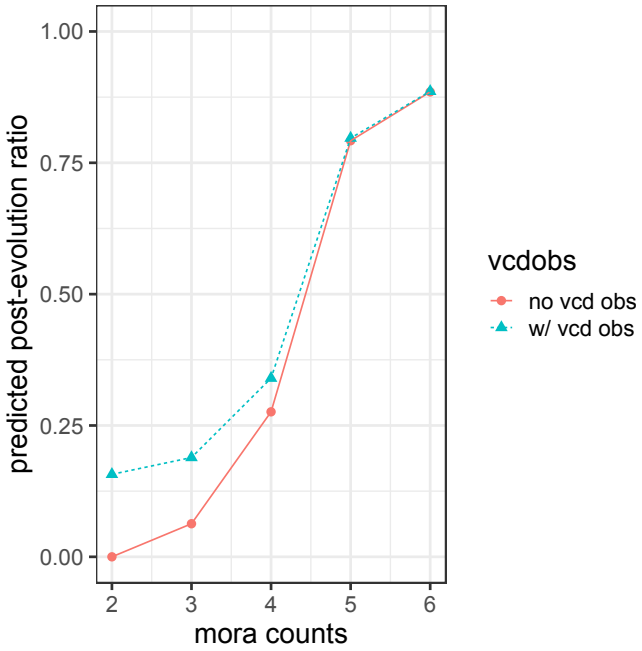


Figure 7: The probability patterns predicted by the GLA, when *LONGPRE is split into a family of different constraints. The "open-jaw" pattern.

The problem comes from the fact that the ranking value of the perturber constraint—*VCDPRE—is too far away from the ranking values of *LONGPRE5MORA, *LONGPRE6MORA, and *POST, essentially resulting in "near strict domination." As a result, *VCDPRE does not have a visible influence on 5-mora long names and 6-mora long names. This problem is a general one (Hayes 2020): the perturber constraint can have one ranking value, and hence has a hard time exerting its influence across the whole x-axis range, when it is placed near one end of the constraint value continuum.

This aspect of Stochastic OT was identified by Zuraw & Hayes (2017) in their quantitive analysis of French liaison. Indeed the general constraint profiles for the current analysis are similar to those for their analysis of French. The set of *LONGPREXMORA constraints and *VCDPRE are synergistic in that they both favor post-evolution names, and the other constraint, *POST, favors pre-evolution names. Zuraw & Hayes (2017: 530) offer an intuitive explanation of how this type of constraint violation profile results in a pattern like the one in Figure 7. Citing an unpublished work by Giorgio Magri, they characterize this pattern as "[two curves] will be uniformly converging in one direction and diverging in the other...where [the] differences...grow monotonically toward

the right of the plot" (p. 530). The pattern in Figure 7 looks precisely like what Zuraw & Hayes describe, with a very minor difference that the divergence is largest on the left (rather than on the right) of the plot in Figure 7.

Bruce Hayes pointed out (p.c.) that Stochastic OT may perform better if the perturber P constraint (*VCDPRE) is reformulated in such a way that it agrees with the binary constraint B (*POST). Following this suggestion, I reformulated *VCDPRE as a constraint that penalizes a post-evolution name which does not start with a voiced obstruent, as in (7).

(7) The new perturber constraint

a. STARTWITHVCDOBSPOST: Assign a violation mark for each post-evolution name which does not start with a voiced obstruent.

This new constraint reminds us of a positional markedness constraint, which for example, requires a low-sonority segment in onset positions (Smith 2002). Unlike *VCDPRE, this constraint penalizes a post-evolution name (rather than a pre-evolution name), when it does not have a particular property. The ranking values that the GLA learned with this new perturber constraint are: *LONGPRE3MORA = 64.6, *LONGPRE4MORA = 65.9, *LONGPRE5MORA = 69.9, *LONGPRE6MORA = 70.4, STARTWITHVCDOBSPOST = 66.6, *POST = 67.5, which yields two curves shown in Figure 8.

The two curves look better separated in Figure 8 than in Figure 7, because the ranking value of the perturber constraint, STARTWITHVCDOBSPOST, is placed in the middle of the constraint ranking continuum in this analysis. We can see that the difference between the two curves is largest for 4-mora long names, and the difference becomes smaller as the name gets shorter or longer. If we had a larger range of x-axis values, the separation of the two curves should eventually disappear at both ends, predicting a "cucumber curve," in which the difference between the two curves monotonically become larger as we move toward the middle range in the horizontal axis.

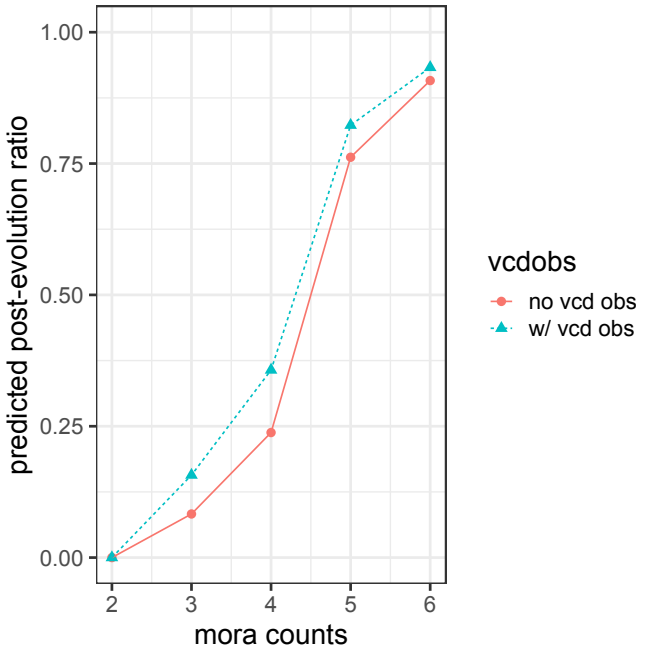


Figure 8: The probability patterns predicted by the GLA with the new perturber constraint in (7).

As demonstrated in this section, Stochastic OT requires that we split the scalar constraint (*LONGPRE) into a set of multiple constraints (Boersma 1998; McPherson & Hayes 2016) to account for the counting cumulativity effect, thus requiring the greater number of free parameters. In addition, the problem identified by Hayes (2020), also observed in the current analyses, is a general one: the perturber constraint can have only one ranking value, so its influence is localized. When it is placed in the middle of the ranking value continuum, as in Figure 8, we may observe a global separation of the two curves, as long as the x-axis range is sufficiently limited. If the x-axis has a wider range, however, it is predicted that the perturber cannot influence the whole x-axis range.

Finally, to conclude this section on Stochastic OT, the log-likelihood—a measure of deviation between the observed data and the model predictions—of the two Stochastic OT analyses are calculated, which were -459.6 and -546.8.[12] These values are lower than that of the MaxEnt model (-432.3) (recall that log-likelihood values that are closer to 0 are better). Moreover, the Stochastic OT models and the MaxEnt model differ in terms of the number of free parameters (i.e. the number of constraints): 6 vs. 3. Therefore, the AIC (Akaike Information Criterion) is calculated for each model, yielding 931.2, 1105.5 (the two Stochastic OT models) and 870.7 (the MaxEnt model)—a model with lower AIC makes a better prediction about the data (Akaike 1973).[13]

[12]Since we cannot take log of 0, it was replaced with $\frac{1}{10^6}$.

[13]There are two caveats. First, *LONGPRE in the MaxEnt model can assign a wider range of constraint violation marks than the set of *LONGPREXMORA constraints in the Stochastic OT model does, because the former is a scalar

7 Concluding remarks

7.1 Summary

The current project was largely inspired by the research program proposed by Hayes (2020). In order to compare various stochastic linguistic models, it is useful to think abstractly about what quantitative predictions the competing theories make. Taking MaxEnt as an example, Hayes (2020) shows that we should be able to identify wug-shaped curves under certain circumstances. The current experiment addressed this prediction in the domain of sound symbolism, and has shown that we can indeed identify wug-shaped curves when certain variables are systematically manipulated for the judgment of evolvedness in Pokémon names. To the extent that wug-shaped curves are typical quantitative signatures of MaxEnt, it shows that MaxEnt is a grammatical framework that is suited to model sound symbolic patterns in natural languages (Kawahara et al. 2019; Kawahara 2020a). To put the results in a more theory-neutral fashion, Japanese speakers take into account different sources of information (mora counts and a voiced obstruent) in a cumulative way, more specifically, in the way that is naturally predicted by MaxEnt.

Viewed from a slightly different—albeit related—perspective, the current experiment addressed the general issue of cumulativity in sound symbolism. The effects of mora counts instantiated a case of counting cumulativity, in that each mora count contributes to the judgment of evolvedness in a sigmoidal fashion. The overall patterns also instantiate a ganging-up cumulativity in that the effects of voiced obstruents and those of mora counts additively contributed to the judgment of evolvedness. Such cumulative patterns are natural consequences of MaxEnt.

7.2 Phonological patterns and sound symbolic patterns

To the extent that MaxEnt is a useful tool to model phonological patterns including both input-output mappings and surface phonotactics judgment patterns, as many previous studies have already shown (e.g. Hayes & Wilson 2008; McPherson & Hayes 2016; Zuraw & Hayes 2017), the overall results point to an intriguing parallel between phonological patterns and sound symbolic patterns. Traditionally, sound symbolism barely received serious attention from formal phonologists (Alderete & Kochetov 2017; Kawahara 2020b). However, the current results suggest that there may be non-negligible similarities between sound-meaning mappings and phonological input-output mappings (as well as wellformedness judgments of surface phonotactic patterns). Phonological patterns and sound symbolic patterns share two important properties—stochasticity and cumulativity—both of which naturally follow from the MaxEnt grammar. This conclusion

constraint and the latter is a binary constraint. Second, since this comparison between MaxEnt and Stochastic OT is based on a single case study, I do not consider the arguments presented here to be final. See Breiss (2020) and Zuraw & Hayes (2017) for other recent case studies which offer quantitative comparisons of MaxEnt and Stochastic OT.

in turn implies that sound symbolism may not be as irrelevant to formal phonological theory as has been assumed in the past, echoing the claim recently made by several researchers (Alderete & Kochetov 2017; Jang 2019; Kawahara 2020b; Kumagai 2019; Shih 2020).[14]

If this hypothesis is on the right track, one question that arises is how closely these two systems are related to one another. I am unable to offer a full fledged answer to this general question here, but can partially address it by asking a more concrete question: whether sound symbolic constraints of the sort that are used in the current paper can coerce phonological changes. Alderete & Kochetov (2017) argue that such patterns do exist. Patterns of expressive palatalization, often found in baby-talk registers, exhibit properties that are different from "regular" phonological palatalization processes; e.g. the former can target all the coronal segments in a word without a clear trigger like a high front vowel (e.g. /o**s**akana-**s**aɴ/ → [o**ɕ**akana-**ɕ**aɴ] 'fish-y' in Japanese). They thus argue that expressive palatalization patterns are caused by sound symbolic requirements, instead of constraints that are purely phonological, and propose a family of EXPRESS(X) constraints, which demands that a particular meaning is expressed by a particular sound. Expressive palatalization may thus instantiate a case in which sound-symbolic constraints coerce phonological changes. See Jang (2019) and Kumagai (2019) for other possible examples of this sort.

7.3 The final conclusion

Finally, I would like to close this paper by putting forward the following methodological thesis: phonological theory can inform research on sound symbolism. While sound symbolism is currently studied very actively, most of such research is being conducted by psychologists, cognitive scientists and cognitive linguists, and few formal phonologists pay serious attention to sound symbolism. However, the current research, inspired by Hayes (2020), has revealed important aspects of sound symbolism—their cumulative nature and how they can be modeled using MaxEnt. Hayes (2020) offers an abstract "top-down" approach, which takes one theory seriously and considers its consequences. If it hadn't been for this approach, I would not have conducted the current experiment. More generally speaking, then, phonological theory can inform research on sound symbolism in important ways. In turn, I hope to have shown that sound symbolism may offer a new testing ground to examine how a cumulative nature of linguistic patterns manifests itself, and in this sense, studies of sound symbolism can inform phonological theories as well. All in all, I hope to have shown with the current case study that phonological theories and research on sound symbolism can and should mutually inform one another.

[14]To this, we can add studies of metrics conducted by phonologists. To the extent that metrics can be a topic of phonological inquiry, which in fact they have been, I do not see any fundamental reasons to exclude sound symbolism from phonological inquiry either. See e.g. Hayes et al. (2012) for a MaxEnt analysis of metrics.

Appendix: Patterns in the existing names

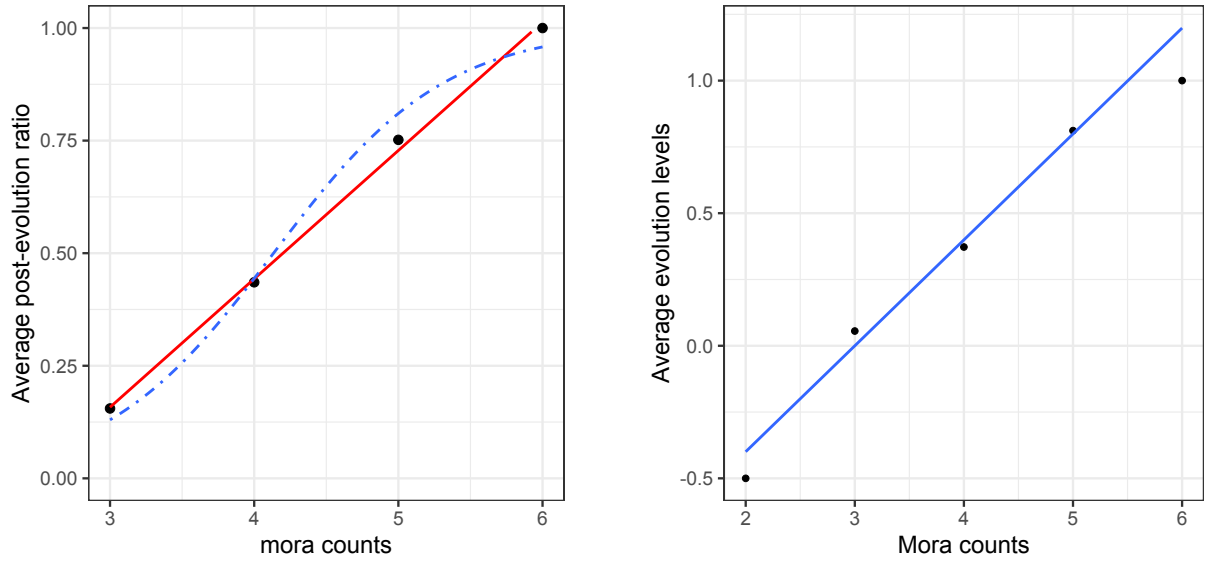
One may wonder how the existing patterns of Pokémon names behave with respect to the issues discussed in the main text. To address this question, I have used the dataset compiled by Kawahara et al. (2018), which includes all the data up to those characters included in the 6th generation, for which there are about 700 characters. Some Pokémon characters do not undergo evolution at all, and those were removed from the analysis. Some other Pokémons were "baby" Pokémons, which were introduced as a pre-evolution version of an already existing character at a later series. While there are not so many of them (N=16), they were also excluded. Pokémon can undergo evolution twice; in the current analysis, as long as they are evolved once, they were counted as post-evolution. There was only one name that is 6-moras long, so this data point has to be interpreted with caution. The total N was 585 in this analysis.

In order to examine whether we observe a sigmoid curve, Figure 9(a) plots the relationship between the mora length and the averaged probabilities of post-evolution. Both a linear function (red, solid) and a sigmoid curve (blue, dotted) are superimposed. There does not seem to be a good reason to believe that the sigmoid curve fits the data better than the linear function. The analysis reported by Kawahara et al. (2018), which makes use of a four-way distinction in terms of evolution—baby Pokémon, no-evolution, evolved once, and evolved twice (coded as -1, 0, 1, 2, respectively)—likewise shows a similar linear trend, as shown here as Figure 9(b) (edited by the author based on their Figure 7).

Based on the inspection of Figure 9, we may tentatively conclude that sigmoid curves (and hence wug-shaped curves) emerged in the current experimental settings, despite the absence of such patterns in the existing names.

An anonymous reviewer has raised the question of where this difference between the real names and experimental results comes from, asking if "there [is] anything inherent to MaxEnt ... that would force a linear input distribution to be converted into a sigmoidal distribution." The answer is positive. Because of the mathematics that lies behind MaxEnt, given a scaler constraint, it has to result in a sigmoid curve, not in a linear curve (Jurafsky & Martin 2019; McPherson & Hayes 2016; Zuraw & Hayes 2017; Winter 2019).

A question that arises is why we observe a linear pattern in the existing names instead of a sigmoid curve. My tentative hypothesis is that since the experiment focused on sound symbolism using nonce names, it was able to tap how sound symbolic knowledge reveals itself in a more pure and direct form than looking at the set of existing names. In the existing names, sound symbolism is not the only factor that determines Pokemon names; other factors are also taken into consideration, such as occasionally using real words to describe the character; e.g. *hitokage* 'fire lizard' is a kind of a lizard (=*tokage*) which spits out fire (=*hi*). Another complication is that the Pokémon lexicon has evolved over a number of generations, with new Pokémon characters



(a) (b)

Figure 9: (a) The relationship between the mora counts and the averaged probabilities of post-evolution in the existing names, in which evolution is coded as a binary variable. (b) The correlation between the number of moras and the average evolution levels, in which evolution is coded as a four-way variable (see text).

added in each generation. The question of why the existing names show a linear pattern requires further scrutiny, but the current experimental results nevertheless remain encouraging, because, to reiterate, MaxEnt can take a linear input but it has to return a sigmoidal output, which is what the current experiment has found.

Acknowledgements

The paper would not have materialized into its current form without the help of many people, who commented on various incarnations of this project, asked insightful questions, and/or offered analytical help and suggestions. They include Arto Anttila, Andries Coetzee, Christian DiCarnio, Donna Erickson, Hironori Katsuda, Bruce Hayes, Laura McPherson, Jason Shaw, three anonymous *Phonology* reviewers, the Associate Editor, the Editors, as well as the participants at Berkeley Phonology Discussion Group, the NINJAL prosody study group and the IERS workshop on the phonetics-phonology interface hosted by International Christian University. Canaan Breiss deserves a special thanks for many intellectual conversations that we have had about almost all the issues discussed in this paper. I would like to thank Kero for putting the current online experiment on his blog, which helped me to gather the data in a very efficient manner. This project is supported by the JSPS grants #17K13448 and #18H03579 as well as the NINJAL collabora-

tive research project 'Cross-linguistic Studies of Japanese Prosody and Grammar.' Files used for analyses in the current paper are available as supplementary materials. All remaining errors are mine.

References

Akaike, Hirotugu. 1973. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory* 267–281.

Alderete, John. 1997. Dissimilation as local conjunction. In Kiyomi Kusumoto (ed.), *Proceedings of the North East Linguistics Society 27*, 17–31. Amherst: GLSA.

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

Anttila, Arto, Scott Borgeson & Giorgio Magri. 2019. Equiprobable mappings in weighted constraint grammars. *SIGMORPHON 2019* .

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

Bailey, Charles-James N. 1973. *Variation and linguistic theory*. Arlington: Center for Applied Linguistics.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68. 255–278.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 88(2). 231–268.

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

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. Ms. UCLA.

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 & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245–259.

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

Coetzee, Andries W. 2016. A comprehensive model of phonological variation: Grammatical and non-grammatical factors in variable nasal place assimilation. *Phonology* 33. 211–246.

Coetzee, Andries W. & Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 30(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.

Cohn, Abigail. 2006. Is there gradient phonology? In Gisbert Fanselow, Caroline Fery, Matthias Schlesewsky & Ralf Vogel (eds.), *Gradience in grammar: Generative perspectives*, 25–44. Oxford: Oxford University Press.

Crowhurst, Megan. 2011. Constraint conjunction. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology*, 1461–1490. Oxford: Blackwell-Wiley.

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

de Lacy, Paul. 2006. *Markedness: Reduction and preservation in phonology*. Cambridge: Cambridge University Press.

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

D'Onofrio, Annette. 2014. Phonetic detail and dimensionality in sound-shape correspondences: Refining the *bouba-kiki* paradigm. *Language and Speech* 57(3). 367–393.

Farris-Trimble, Ashley. 2008. *Cumulative faithfulness effects in phonology*: Indiana University Doctoral dissertation.

Featherston, Sam. 2005. The decathlon model of empirical syntax. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 187–208.

Garcia, Guilherme Duarte. 2019. When lexical statistics and the grammar conflict: Learning and repairing weight effects on stress. *Language* 95(4). 612–641.

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

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

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

Gouskova, Maria. 2004. Relational markedness in OT: The case of syllable contact. *Phonology* 21(2). 201–250.

Guy, Gregory. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3. 1–22.

Guy, Gregory. 2011. Variability. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology*, 2190–2213. Oxford: Blackwell-Wiley.

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

Haiman, John. 1984. *Natural syntax: Iconicity and erosion*. Cambridge: Cambrdige University Press.

Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the sounds

of their language. In Morris Halle, Joan Bresnan & George A. Miller (eds.), *Linguistic theory and psychological reality*, 294–303. Cambridge: MIT Press.
Hamano, Shoko. 1998. *The sound-symbolic system of Japanese*. Stanford: CSLI Publications.
Hamano, Shoko. 2013. Hoogen-ni okeru giongo-gitaigo-no taikeiteki-kenkyuu-no igi. In Kazuko Shinohara & Ryoko Uno (eds.), *Chikazuku oto-to imi: Onomatope kenkyuu-no shatei*, Tokyo: Hitsuzi Syobo.
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 & Jinyoung Jo. 2020. Balinese stem phonotactics and the subregularity hypothesis. Ms. UCLA.
Hayes, Bruce & Zsuzsa Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23. 59–104.
Hayes, Bruce, Bruce Tesar & Kie Zuraw. 2014. Otsoft 2.5. software package, `http://www.linguistics.ucla.edu/people/hayes/otsoft/`.
Hayes, Bruce & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44. 45–75.
Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39. 379–440.
Hayes, Bruce, Colin Wilson & Anne Shisko. 2012. Maxent grammars for the metrics of Shakespeare and Milton. *Language* 88(4). 691–731.
Hayes, Bruce, Kie Zuraw, Péter Siptár & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4). 822–863.
Hinton, Leane, Johanna Nichols & John Ohala. 2006. *Sound symbolism, 2nd edition*. Cambridge: Cambridge University Press.
Hsu, Brian & Karen Jesney. 2017. Loanword adaptation in Québec French: Evidence for weighted scalar constraints. *Proceedings of the 34th West Coast Conference on Formal Linguistics* 249–258.
Ito, Junko & Armin Mester. 1986. The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry* 17. 49–73.
Ito, Junko & Armin Mester. 2003. *Japanese morphophonemics*. Cambridge: MIT 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. 2019. How cute do I sound?: The iconic function of segmental alternation in Korean baby-talk register, *aegyo*. Ms. University of Southern California.
Jarosz, Gaja. 2017. Defying the stimulus: Acquisition of complex onsets in Polish. *Phonology* 34(2). 269–298.
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/.
Kawahara, Shigeto. 2017. *Introducing Phonetics through Sound Symbolism*. Tokyo: Hitsuzi Syobo.
Kawahara, Shigeto. 2020a. Cumulative effects in sound symbolism. Ms. Keio University.
Kawahara, Shigeto. 2020b. Sound symbolism and theoretical phonology. *Language and Linguistic Compass* 14(8). e12372.
Kawahara, Shigeto, Mahayana C. Godoy & Gakuji Kumagai. 2020a. Do sibilants fly? Evidence from a sound symbolic pattern in Pokémon names. *Open Linguistics* 6. 1–15.
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.
Kawahara, Shigeto, Kazuko Shinohara & Yumi Uchimoto. 2008. A positional effect in sound symbolism: An experimental study. In *Proceedings of the Japan Cognitive Linguistics Association 8*, 417–427. Tokyo: JCLA.
Kawahara, Shigeto, Michinori Suzuki & Gakuji Kumagai. 2020b. Sound symbolic patterns in Pokémon move names in Japanese. *ICU Working Papers in Linguistics 10. Festschrift for Prof. Junko Hibiya in the occasion of her retirement from ICU* 17–30.
Kellar, Frank. 2006. Linear Optimality Theory as a model of gradience in grammar. In Gisbert Fanselow, Caroline Féry, Ralf Vogel & Matthias Schlesewsky (eds.), *Gradience in grammar: Generative perspectives*, 270–287. Oxford: Oxford University Press.
Kisseberth, Charles. 1970. The treatment of exceptions. *Papers in Linguistics* 2. 44–58.
Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(2). 199–244.
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. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45. 715–762.
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.
Leben, Will. 1973. *Suprasegmental phonology*: MIT Doctoral dissertation.
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.
Martin, Samuel. 1962. Phonetic symbolism in Korean. In N. Poppe (ed.), *American studies in Uralic and Altaic linguistics*, Indiana University Press.
McCarthy, John J. 1983. Phonological features and morphological structure. In J. Richardson, M. Marks & A. Chukerman (eds.), *Proceedings from the parasesion on the interplay of phonology, morphology and syntax*, 135–161. Chicago: CLS.
McCarthy, John J. 2003. OT constraints are categorical. *Phonology* 20(1). 75–138.

McCarthy, John J. & Alan Prince. 1986. Prosodic morphology. Ms., University of Massachusetts and Rutgers University.
McPherson, Laura. 2016. Cumulativity and ganging in the tonology of Awa suffixes. *Language: Phonological Analysis* 92(1). e38–e66.
McPherson, Laura & Bruce Hayes. 2016. Relating application frequency to morphological structure: The case of Tommo So vowel harmony. *Phonology* 33. 125–167.
Myers, Scott. 1997. OCP effects in Optimality Theory. *Natural Language and Linguistic Theory* 15(4). 847–892.
Nishimura, Kohei. 2006. Lyman's Law in loanwords. *Phonological Studies* 9. 83–90.
O'Hara, Charlie. 2017. How abstract is more abstract? Learning abstract underlying representations. *Phonology* 34(2). 324–345.
Otake, Takashi, Giyoo Hatano, Anne Cutler & Jacques Mehler. 1993. Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language* 32. 258–278.
Paster, Mary. 2019. Phonology counts. *Radical* 1. 1–61.
Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33. 999–1035.
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.
Pierrehumbert, Janet B. 2001. Stochastic phonology. *GLOT* 5. 1–13.
Pizzo, Presley. 2015. *Investigating properties of phonotactic knowledge through web-based experimentation*: University of Massachusetts, Amherst Doctoral dissertation.
Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker. 2010. Harmonic grammar with linear programming: From linear systems to linguistic typology. *Phonology* 27(1). 1–41.
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.
Ramachandran, Vilayanur S. & Edward M. Hubbard. 2001. Synesthesia–a window into perception, thought, and language. *Journal of Consciousness Studies* 8(12). 3–34.
Schütze, Carson. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
Shih, Stephanie S. 2017. Constraint conjunction in weighted probabilistic grammar. *Phonology* 34(2). 243–268.
Shih, Stephanie S. 2020. Gradient categories in lexically-conditioned phonology: An example from sound symbolism. *Proceedings of the 2019 Annual Meeting on Phonology* .
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.
Smith, Brian W. & Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa* 5(1). 24, doi: `http://doi.org/10.5334/gjgl.583`.
Smith, Jennifer. 2002. *Phonological augmentation in prominent positions*: University of Mas-

sachusetts, Amherst Doctoral dissertation.
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.
Smolensky, Paul. 1995. On the internal structure of the constraint component CON of UG. Talk presented at the University of California, Los Angeles (ROA-86).
Smolensky, Paul. 1997. Constraint interaction in generative grammar II: Local conjunction, or random rules in universal grammar. Handout of talk given at Hopkins Optimality Theory Workshop/Maryland Mayfest, Baltimore.
Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1. 123–134.
Tesar, Bruce. 2007. A comparison of lexicographic and linear numeric optimization using violation difference ratios. Ms. Rutgers University.
Thompson, Patrick D. & Zachary Estes. 2011. Sound symbolic naming of novel objects is a graded function. *Quarterly Journal of Experimental Psychology* 64(12). 2392–2404.
Wasserman, Larry. 2004. *All of statistics: A concise course in statistical inference*. New York: Springer.
Westbury, Chris. 2005. Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language* 93. 10–19.
Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30(5). 945–982.
Wilson, Colin. 2014. Maximum entropy models. Tutorial presentation, MIT.
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.