Phonology **37** (2020) 383–418.
doi:10.1017/S0952675720000202

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

Shigeto Kawahara
Keio University

An experiment showed that Japanese speakers' judgement of Pokémons' evolution status on the basis of nonce names is affected both by mora count and by the presence of a voiced obstruent. The effects of mora count are a case of counting cumulativity, and the interaction between the two factors a case of ganging-up cumulativity. Together, the patterns result in what Hayes (2020) calls 'wug-shaped curves', a quantitative signature predicted by MaxEnt. I show in this paper that the experimental results can indeed be successfully modelled with MaxEnt, and also that Stochastic Optimality Theory faces an interesting set of challenges. The study was inspired by a proposal made within formal phonology, and reveals important previously understudied aspects of sound symbolism. In addition, it demonstrates how cumulativity is manifested in linguistic patterns. The work here shows that formal phonology and research on sound symbolism can be mutually beneficial.

* E-mail: KAWAHARA@ICL.KEIO.AC.JP.
The paper would not have reached its current form without the help of the 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 DiCanio, Donna Erickson, Bruce Hayes, Hironori Katsuda, Laura McPherson, Jason Shaw, three anonymous *Phonology* reviewers, the associate editor and 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 the International Christian University. Canaan Breiss deserves a special thanks for many conversations that we have had about almost all the issues discussed in this paper. I would like to thank Kero for hosting 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 collaborative research project 'Cross-linguistic studies of Japanese prosody and grammar'. Files used for analyses in the current paper are available as supplementary materials at https://doi.org/10.1017/S0952675720000202. All remaining errors are mine.

1 Introduction

1.1 A wug-shaped curve

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

On the other hand, probabilistic generalisations regarding phonological variations have been a central topic of sociolinguistic research (e.g. Labov 1966, Guy 2011), in which it has been claimed that variation is 'the central problem of linguistics' (Labov 2004: 6). For example, it is not uncommon for the same word to 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 on the basis of the interaction of various (morpho-)phonological and social factors (e.g. *t*/*d*-deletion in English: Guy 1991), an observation which has been modelled in various formal frameworks (e.g. Cedergren & Sankoff 1974, Guy 1991, Johnson 2009). Syntactic variations and their historical changes also seem to exhibit systematic quantitative patterns (Kroch 1989, Zimmermann 2017); these have also been analysed from formal perspectives (e.g. Featherston 2005, Bresnan & Hay 2008).

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, Pierrehumbert 2001, Cohn 2006, Hayes & Londe 2006, Coetzee & Pater 2011, Daland *et al*. 2011). Some phonotactic sequences are neither completely grammatical nor ungrammatical, but intermediate; indeed, controlled phonotactic judgement experiments typically reveal a continuous gradient pattern (e.g. Daland *et al*. 2011).

The field of phonology has recently witnessed a rise of interest in formal grammatical models which can account for such stochastic phonological generalisations. Among these, the three most widely employed frameworks are (i) Stochastic Optimality Theory (Boersma 1998, Boersma & Hayes 2001), (ii) Noisy Harmonic Grammar (Coetzee & Kawahara 2013, Boersma & Pater 2016, Hayes 2017) and (iii) Maximum Entropy Harmonic Grammar (henceforth MaxEnt) (Goldwater & Johnson 2003, Zuraw & Hayes 2017). How these stochastic models of phonology should be teased apart is currently a topic of debate in phonological

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

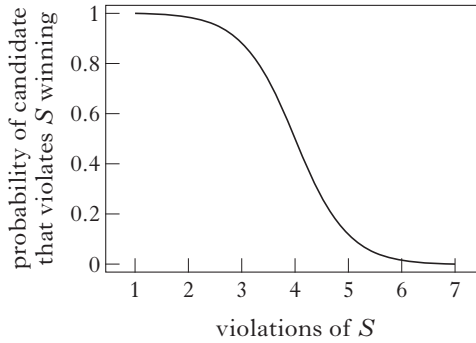

Figure 1

A sigmoid curve predicted by the MaxEnt grammar with a scalar constraint *S* and a binary constraint *B*, whose violations conflict. Based on Hayes (2020: 5). The mathematical equation which derives this curve is $y = 1 / (1 + e^{-H})$, where H directly correlates with the number of violations of *S*. See Jurafsky & Martin (2019) and McPherson & Hayes (2016), as well as §5.1.

studies (Jäger & Rosenbach 2006, Jäger 2007, Hayes 2017, 2020, Zuraw & Hayes 2017, Anttila & Magri 2018, Breiss 2020, among many others).

In order to distinguish between these theoretical frameworks, Hayes (2020), building upon a body of previous studies on probabilistic linguistic patterns (Kroch 1989, McPherson & Hayes 2016, Zimmermann 2017, Zuraw & Hayes 2017), proposes an abstract, top-down approach, asking the following question: if we take the MaxEnt grammar framework seriously, what predictions does it make for its quantitative signature, i.e. the probabilistic pattern that it typically generates? More specifically, 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 Fig. 1. In reality, the constraint-violation profile of *S* is discrete (ranging from 1 to 7 in Fig. 1), but for the sake of illustration, Fig. 1 continuously plots for all values, not just the integers. This curve is characterised 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 display radical change in the middle range (from 3 to 5).

Hayes also 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 Fig. 1, but the two sets differ in terms of whether they violate an additional 'perturber' constraint (*P*) or not. This scenario

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

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

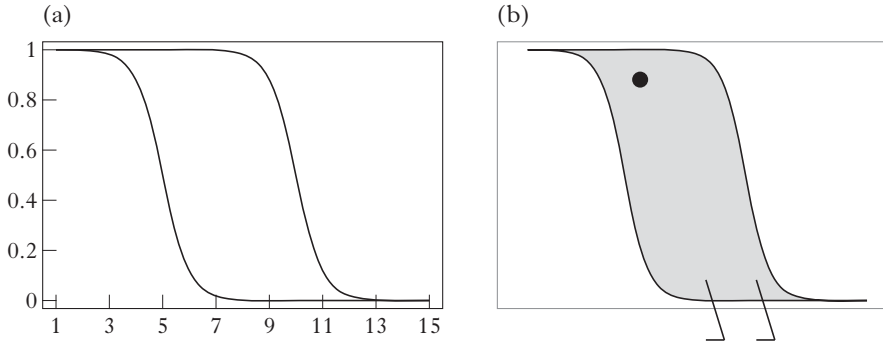

Figure 2
Wug-shaped curves with two sigmoid functions. Adapted from Hayes (2020: 7).

creates two identical sigmoid curves, shifted from one another on the horizontal axis, as in Fig. 2a. Hayes (2020) calls these 'wug-shaped curves', because, as illustrated in Fig. 2b, they are reminiscent of the beloved animal familiar to the general linguistic community since the classic work of Berko (1958).

Studying whether wug-shaped curves are observed in linguistic patterns is important, because they are natural outcomes of MaxEnt grammars, and are also predicted under some versions of Noisy Harmonic Grammar, but not in Stochastic Optimality Theory. This top-down approach to examining quantitative signatures of linguistic generalisations therefore offers a possible strategy for distinguishing among three competing stochastic models of grammar. If we find wug-shaped curves in linguistic patterns, this provides support for MaxEnt 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 categorical perception of speech sounds (Liberman *et al*. 1957) and diachronic changes in syntax (Kroch 1989, Zimmermann 2017).

Building on these studies, this paper asks whether we can identify wug-shaped curves in the patterns of sound symbolism, i.e. systematic/iconic associations between sounds and meanings (Hinton *et al*. 2006). If the answer to this question turns out to be positive, this provides support for 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 appropriate as a model of phonological knowledge (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.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

1.2 Cumulativity

One general theoretical issue that lies behind wug-shaped curves is that of cumulativity. This is a topic that has been addressed in recent linguistic theorisation, because it potentially helps us to distinguish Optimality Theory with ranked constraints (Prince & Smolensky 1993) from constraint-based theories with numerically weighted constraints, such as Harmonic Grammar (Jäger & Rosenbach 2006, Jäger 2007, Pater 2009, Potts *et al*. 2010, Hayes *et al*. 2012, McPherson & Hayes 2016, Zuraw & Hayes 2017, Breiss 2020).

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 Optimality Theory. In the context of OT, we find counting cumulativity when two or more violations of a lower-ranked constraint take precedence over the violation of a higher-ranked constraint. Consider the schematic case of counting cumulativity in (1a). As (1a.i) shows, Constraint A dominates Constraint B. However, as in (1a.ii), two violations of Constraint B are considered to be more important than a single violation of Constraint A.

If Constraint A dominates Constraint B in an OT analysis, then a single violation of Constraint A should take precedence over any number of violations of Constraint B – this is a consequence of the strict domination of constraint rankings, a central tenet of OT (Prince & Smolensky 1993). In reality, however, it is not uncommon for a language to tolerate one violation of a particular constraint, but not two violations, instantiating a case of counting cumulativity. For instance, the native phonology of Japanese allows one voiced obstruent within a morpheme, but not two (Lyman's Law; Itô & Mester 1986, 2003). Such observations are commonly accounted for in OT by positing OCP constraints (Leben 1973, Itô & 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 any additional mechanism (e.g. McPherson & Hayes 2016).[2]

[2] There remains a difference between OT equipped with OCP constraints and related constraint-based theories with weighted constraints. One widely shared idea in OT is that there can be a constraint that penalises two instances of a particular structure, but there are no constraints that penalise exactly three instances. The following quote from Ito & Mester (2003: 265) 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''. As this implies, studies in OT have generally assumed that there are no constraints that are violated if and only

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

(1) a.

		Constraint A	Constraint B	Constraint C
i.	☞ Candidate 1		*	
	Candidate 2	*		
ii.	Candidate 3		**	
	☞ Candidate 4	*		

b.

		Constraint A	Constraint B	Constraint C
i.	☞ Candidate 1		*	
	Candidate 2	*		
ii.	☞ Candidate 3			*
	Candidate 4	*		
iii.	Candidate 5		*	*
	☞ Candidate 6	*		

Ganging-up cumulativity is illustrated by the set of tableaux in (1b). In (1b.i) and (1b.ii) Constraint A dominates Constraint B and Constraint C respectively. Ganging-up cumulativity is said to hold when the simultaneous violation of Constraint B and Constraint C takes precedence over a single violation of Constraint A, as in (1b.iii); i.e. violations of Constraint B and Constraint C 'gang up' to take precedence over a violation of Constraint A. To analyse a ganging-up cumulativity pattern, OT generally requires local conjunction of Constraints B and C (Smolensky 1995, Crowhurst 2011). 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) proposes the local conjunction of *VoicedObsGem and OCP[voice] within the stem domain. Frameworks with numerically weighted constraints can account for this ganging-up cumulativity pattern in Japanese without stipulating a complex locally conjoined constraint (Pater 2009; see also Potts *et al.* 2010).

In short, whether phonological patterns show counting or ganging-up cumulativity bears on the issue of whether the grammatical model

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 has recently been challenged by Paster (2019). See McPherson & Hayes (2016), Paster (2019) and Kawahara, Suzuki & Kumagai (2020), as well as the experimental results below, for cases which apparently count beyond two.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

should be based on rankings or weights. More generally, the question is whether the optimisation algorithm deployed in the linguistic system is based on lexicographic ordering or numeric ordering (Tesar 2007).

In this paper I attempt to shed new light on this debate by examining a pattern that has hitherto hardly been analysed from this perspective, namely, sound symbolism. The primary question that is addressed in this study is whether sound symbolism shows counting cumulativity effects and/or ganging-up cumulativity effects, 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 considered the (non-)cumulative nature of sound symbolism, and this is one aspect of sound symbolism that is only poorly understood. There are some impressionistic reports regarding counting cumulativity in the literature which suggest that more segments of the same kind evoke stronger sound-symbolic images (e.g. Martin 1962, McCarthy 1983, Hamano 2013). Thompson & Estes (2011) carried out experiments to establish whether sound symbolism is categorical or gradient, and found some evidence for cumulativity in their results. A recent experimental study by Kawahara & Kumagai (to appear) found evidence for counting cumulativity in various sound-symbolic values of voiced obstruents in Japanese. D'Onofrio (2014) examined the *bouba-kiki* effect (Ramachandran & Hubbard 2001), in which certain classes of sounds are associated with round figures and other classes 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] To the best of my knowledge, there have been no studies that have addressed the question of whether counting cumulativity and ganging-up cumulativity can coexist in the same sound-symbolic system, as predicted by MaxEnt (though see Kawahara, Suzuki & Kumagai 2020, 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, 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 be manifested as a linear function, rather than a sigmoid function. See §5.3 for further elaboration on this point.

In domains other than sound symbolism, Breiss (2020) shows that we observe both counting and ganging-up cumulativity in phonotactic learning patterns in an artificial language learning experiment. Case studies of phonological alternation patterns reported in McPherson & Hayes

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

(2016) and Zuraw & Hayes (2017) can also be understood as simultaneously involving counting and ganging-up cumulativity. There have not been many other case studies that have directly addressed this question, especially in the domain of sound symbolism. Since the coexistence of counting cumulativity and ganging-up cumulativity is a natural consequence of MaxEnt, 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 (McPherson & Hayes 2016, Zuraw & Hayes 2017, Breiss 2020, Hayes 2020), and if 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 has recently been argued (Alderete & Kochetov 2017, Kumagai 2019, Jang 2020, Kawahara 2020a, b, Shih 2020). This is a rather radical conclusion, given the fact that sound symbolism has long been considered as being outside the purview of theoretical linguistics.

1.3 Pokémonastics

In addition to addressing the issue of cumulativity in sound symbolism, this 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 discussion of this research paradigm, and provide minimal background information necessary for what follows.

Pokémon is a game series which was first released by Nintendo Inc. in 1996, and has become very popular worldwide. In this game series, players collect and train fictional creatures called Pokémons (Pokémon is a truncation of *poketto monsutaa* 'pocket monster'). One feature that will be 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 acquire 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 attribute this observation to a previously formulated sound-symbolic principle, '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 than pre-evolution characters to have names with voiced obstruents. This is likely to be related to the observation that 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.

The rest of this paper proceeds as follows. §2 reports the methods of the experiment, which was designed to address the question of whether we observe a wug-shaped curve in sound symbolism. The results of the experiment demonstrate that sound symbolism shows both counting and ganging-up cumulativity, and that these two types of cumulativity can coexist within a single sound-symbolic system (§3 and §4). These cumulative patterns result in a wug-shaped curve, which can naturally be modelled using MaxEnt (§5). §6 discusses several attempts to use Stochastic OT to model the current results, which shows that this framework requires additional tweaks to fit the wug-shaped pattern observed in the experiment. §7 offers concluding remarks, arguing that formal phonology and research on sound symbolism can inform one another.

2 Methods

One precursor of the current experiment is Kawahara, Suzuki & Kumagai (2020), who carried out a judgement experiment on the strengths of Pokémon move names (moves are what Pokémons use when they battle with each other). Kawahara, Suzuki & Kumagai manipulated mora count from two to seven 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 in word-initial position, and found that nonce move names with voiced obstruents were judged to be stronger. Their results are reproduced in Fig. 3, which instantiates both counting cumulativity (the effect of mora count) and ganging-up cumulativity (the additive effects of the two factors). However, their experiment is not suitable for addressing 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.

The current study builds upon Kawahara, Suzuki & Kumagai (2020), but in order to obtain a binary categorical response, participants in the experiment were asked to judge whether each stimulus name was better suited for a pre-evolution or post-evolution character. To obtain more reliable estimates of each condition, more items were included for each condition. Moreover, in this study responses were collected from many more participants.

2.1 Stimuli

The stimuli used in the experiment are listed in Table I. Building on the two studies reviewed above (Kawahara *et al*. 2018, Kawahara, Suzuki &

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

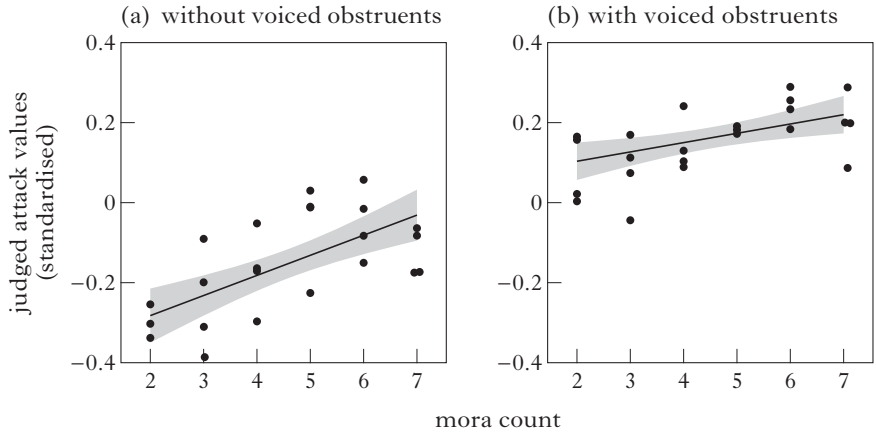

Figure 3
The effects of mora count and word-initial voiced obstruents on judged attack values in nonce Pokémon move names. The *y*-axis shows standardised judged attack values, which are continuous. Adapted from Kawahara, Suzuki & Kumagai (2020: Fig. 4).

Kumagai 2020), two variables were manipulated: mora count and the presence of a voiced obstruent in word-initial position. The mora count was varied, in order to investigate counting cumulativity and, relatedly, to examine whether varying the mora count would result in a sigmoid curve. Mora counts varied from two to six, corresponding to minimum and maximum lengths for Pokémon names. The experiment manipulated mora counts rather than segment or syllable counts, because mora counts were identified as most important in the previous studies (Kawahara *et al*. 2018, Shih *et al*. 2019, Kawahara, Suzuki & Kumagai 2020); moreover, the mora is the most psycholinguistically salient prosodic counting unit in Japanese (Otake *et al*. 1993). The perturbing factor (see §1.1) was the presence or absence of a voiced obstruent in name-initial position.

As shown in Table I, there were six items 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 potential bias by the experimenter to select the stimuli that were likely to support their hypothesis (Westbury 2005). All voiced obstruents appeared word-initially, because a previous study had shown that the strength of sound-symbolic values of voiced obstruents in Japanese may vary depending on word position (Kawahara *et al*. 2008). No geminates, long vowels or coda nasals appeared anywhere in the stimuli; i.e. all syllables were open. Moreover, because of its potentially

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

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

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

Table I

The stimuli used in the experiment. Mora boundaries are represented as dots.

salient sound-symbolic values, such as cuteness (Kumagai 2019), [p] was excluded from the stimuli.

2.2 Procedure

The experiment was distributed as an online experiment using SurveyMonkey.[5] Within each trial, participants were given one nonce

[5] https://www.surveymonkey.com (accessed August 2020).

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

name at a time, and asked to judge whether that name was 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 randomised for each participant.

2.3 Participants

The experiment was advertised on a Pokémon fan website.[6] A total of 857 participants completed the experiment over a single night. Some previous Pokémonastics experiments had been advertised on the same website (e.g. Kawahara, Godoy & Kumagai 2020), and 124 participants reported that they had either taken part in another Pokémonastics experiment or had studied sound symbolism before. Three participants were non-native speakers of Japanese. The data from these speakers was excluded, and 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 fitted, with response (pre-evolution *vs*. post-evolution) as the dependent variable (Jaeger 2008). The fixed independent variables included mora count and the presence of a voiced obstruent as well as its interaction. Mora count was centred, because it is a continuous variable (Winter 2019). Participants and items were random factors. 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 4a plots 'post-evolution response ratios' for each item, averaged over all the participants. The items for the condition with a voiced obstruent are shown with black squares and the items for the condition without a voiced obstruent are shown with grey circles. A logistic curve is superimposed for each voicing condition.

These results look like the wug-shaped curves illustrated schematically in Fig. 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 clearly steepest in the middle range. This is also clear in Fig. 4b, which illustrates the overall pattern by presenting grand averages

[6] http://pokemon-matome.net (accessed August 2020).

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

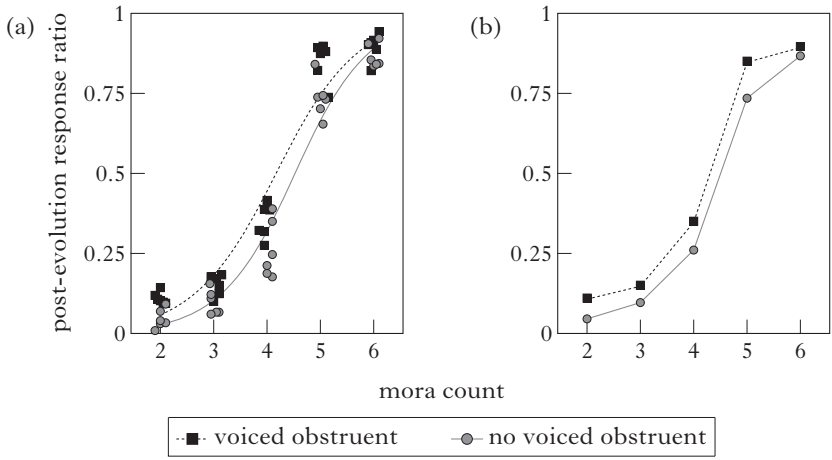

Figure 4

(a) The by-participant averages for each item. The items with a voiced obstruent are shown with black squares and those without a voiced obstruent with grey circles. To avoid overlap, the points are horizontally jittered. Logistic curves are superimposed – the dashed black line represents the condition with a voiced obstruent, and the solid grey line represents the condition without a voiced obstruent. (b) The line plots, with grand averages for each condition.

for each condition – this analysis does not presuppose that sigmoid curves would fit the data points well. The slopes between the 3-mora condition and the 5-mora condition are rather steep. On the other hand, they are not very steep between the 2-mora and 3-mora conditions or between the 5-mora and 6-mora conditions. As Hayes (2020: 3) puts it, 'certainty is evidentially expensive' – we require very strong evidence to be certain that a particular name is that of a pre-evolution or a post-evolution character. A more elaborate defence of using a wug-shaped curve to fit the data is provided in §5.3, once we have developed a full MaxEnt analysis of the data.

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

4 Discussion

The effect of mora count is an example of counting cumulativity, in that each increase in the mora-count scale contributes to the probability that a name will be judged to be that of a post-evolution character. This

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

	β	SE	z	p
Intercept	−0.76	0.107	−7.14	<0.001
Mora count	1.43	0.075	19.04	<0.001
Voiced obstruent	0.63	0.148	4.26	<0.001
Mora count ×Voiced obstruent	−0.15	0.106	−1.42	*n.s.*

Table II
Summary of the logistic linear mixed-effects model.

effect is evident both with and without a name-initial voiced obstruent. The effect of a voiced obstruent in name-initial position is manifested as a shift between the two sigmoid curves. The two effects together are an example of ganging-up cumulativity – both factors contribute to the judgement of evolvedness. Overall, the 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 with the probabilistic phonological alternation patterns discussed by McPherson & Hayes (2016), Zuraw & Hayes (2017) and Hayes (2020). See also Breiss (2020) and Kawahara (2020a) for summaries of cumulative effects in phonological alternations and in well-formedness judgement patterns of surface phonotactics.

While the results in Fig. 4 seem to provide a clear case of 'wug-shaped' curves, we might wonder if the results could have been different. The answer is positive, as multiple alternative patterns could have arisen from the experimental design. For example, the mora-count effect could have been cumulative, but linear rather than sigmoidal. Indeed, the effect of mora count in the existing Pokémon names actually looks more linear than sigmoidal (see the Appendix for discussion).

Alternatively, the results could have been non-cumulative. For example, there could have been a 'length threshold', such that any names shorter than that threshold were judged to be pre-evolution; however, the actual results did not follow such a pattern. Nor did the presence of a voiced obstruent make a name post-evolution in all cases. Instead, both mora counts and voiced obstruents gradiently increased the probabilities of each name being judged to be a post-evolution name.[7] This point is related to another important aspect of sound symbolism, its stochastic nature (Dingemanse 2018, Kawahara *et al.* 2019). More generally,

[7] A very small subset of the participants (17 out of 730; ca. 2%) showed categorical patterns with respect to the effects of mora count, in that they assigned names of all mora counts to either pre-evolution characters or post-evolution characters in all cases and showed no intermediate responses. A MaxEnt grammar, as 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 in all cases.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

Gigerenzer & Gaissmaier (2011) discuss a number of cases in which people making decision adopt 'a fast and frugal decision heuristics' approach – they take into account only the most important information, 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 provides a parallel to a growing body of evidence that many, but perhaps not all, phonological generalisations have to be stated in a stochastic or probabilistic way; for example, 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.[8] 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).[9] A 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 phonological input–output mappings (Kawahara *et al*. 2019, Kawahara 2020a). The model deploys the kind of constraints familiar from the OT tradition (Prince & Smolensky 1993). To underscore the parallel between

[8] 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. 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 McPherson & Hayes 2016 and Hayes 2017, 2020 for detailed discussion). 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.

[9] Another quantitative framework that can model stochastic generalisations 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).

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

phonological analyses and the analysis of sound symbolism developed in this paper, I adopt a particular formalism that has been used to define constraints in the OT research tradition, that of McCarthy (2003).

5.1 A brief review of MaxEnt

This section briefly reviews how MaxEnt works in the context of linguistic analyses.[10] The MaxEnt grammar is similar to OT, in that a set of candidates is evaluated against a set of constraints. Unlike OT, however, constraints are weighted rather than ranked. Consider the toy example in (2). The set of candidates to be evaluated are listed in the leftmost column, and the top row gives the relevant constraints; each constraint is assigned a particular weight (w). The tableau shows the violation profiles of each constraint – how many times each candidate violates a particular constraint.

(2)

	A $w=3$	B $w=2$	C $w=1$	$\mathcal{H}$-score	eHarmony	Z	predicted
Candidate 1	1			$1\times3=3$	$e^{-3}=0.0498$	0.0565	88%
Candidate 2		2	1	$2\times2+1\times1=5$	$e^{-5}=0.0067$	0.0565	12%

Based on the constraint-violation profiles, the Harmony score of each candidate x ($\mathcal{H}$-score(x)) is calculated using the formula in (3), where N is the number of relevant constraints, w_i is the weight of the ith constraint and $C_i(x)$ is the number of times candidate x violates the ith constraint.

(3) $$\mathcal{H}\text{-score}(x) = \sum_{i}^{N} w_i C_i(x)$$

For example, Candidate 2 in tableau (2) violates Constraint B twice and Constraint C once; its $\mathcal{H}$-score is therefore $2 \times 2 + 1 \times 1 = 5$.

The $\mathcal{H}$-scores are negatively exponentiated (eHarmony, represented as e^{-H} or $1 / e^{H}$; according to Hayes 2020, the term was introduced by Colin Wilson in a tutorial presentation at MIT), which is proportional to the probability of each candidate. Intuitively, the more constraint violations a candidate incurs, the higher the $\mathcal{H}$-score, and hence the lower the eHarmony (e^{-H}). Therefore, more constraint violations lead to that candidate having lower probability. The eHarmony values are relativised against the sum of the eHarmony values of all the candidates, Z, as in (4), where M is the number of candidates.

(4) $$Z = \sum_{j}^{M} (e^{-H})_j$$

[10] An earlier version of this section can be found in Kawahara (2020a: §3).

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

In the example in (2), Z is $0.0498 + 0.0067 = 0.0565$. The predicted probability of each candidate x_j, $p(x_j)$, is $\text{eHarmony}(x_j) / Z$.

5.2 A MaxEnt analysis of the results of the experiment

Like most phonological analyses in OT and other related frameworks, a 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 or post-evolution character names. The set of constraints employed in the current analysis is given in (5).[11] These constraints essentially correspond to OT markedness constraints, in that they evaluate the well-formedness of output structures. The definition of the constraints follows the format in McCarthy (2003).

(5) a. *LONG$_{\text{pre-ev}}$
Assign a violation mark for each mora in a pre-evolution character name.

b. *VCD$_{\text{pre-ev}}$
Assign a violation mark for each voiced obstruent in a pre-evolution character name.

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

*LONG$_{\text{pre-ev}}$ 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). It is a single gradient/scalar constraint (McPherson & Hayes 2016), in that it is a reflection of a single principle, whose violations can be assessed on a numerical scale.[12] This constraint corresponds to the scalar constraint S used to illustrate the wug-shaped curves in §1.1. *VCD$_{\text{pre-ev}}$ is a formal expression of the preference that character names with voiced obstruents should be used for post-evolution names; this corresponds to

[11] If notions like 'pre-evolution' and 'post-evolution' are considered to be too language- or culture-specific to be mentioned in OT-style constraints, which are generally taken to be universal, they can be replaced with 'small entity' and 'large entity', since Pokémon characters generally become larger after evolution. Size, together with shape, is a semantic dimension that is clearly signalled by sound symbolism in many languages (Sidhu & Pexman 2018).

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

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

the perturber constraint *P* in §1.1. *Post is a *Struc constraint (Prince & Smolensky 1993), which penalises post-evolution character names in general, and corresponds to the binary constraint *B* discussed in §1.1. We need this constraint because there has to be some constraint that favours pre-evolution character names. All three constraints are statistically motivated by a log-likelihood ratio test, to be presented below in Table III.

Hayes (2020) recommends that we conceive of constraint violations as providing evidence for which candidate should be chosen. 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 (*Long$_{\text{pre-ev}}$) or when they contain a voiced obstruent (*Vcd$_{\text{pre-ev}}$), and *Post helps us to decide on a pre-evolution name in general. The weights associated with each constraint reflect the strengths, or cogency, of each piece of evidence.

(6)

	input	*output*	*Long$_{\text{pre}}$ 1.346	*Vcd$_{\text{pre}}$ 0.49	*Post 6.10	$\mathcal{H}$-score	eHarmony	E	O
a.	2 moras: voiceless	pre	2			2.69	0.068	96.80	95.48
		post			1	6.10	0.002	3.20	4.52
b.	3 moras: voiceless	pre	3			4.04	0.018	88.72	90.39
		post			1	6.10	0.002	11.28	9.61
c.	4 moras: voiceless	pre	4			5.38	0.005	67.18	73.97
		post			1	6.10	0.002	32.82	26.03
d.	5 moras: voiceless	pre	5			6.73	0.001	34.76	26.51
		post			1	6.10	0.002	65.24	73.49
e.	6 moras: voiceless	pre	6			8.08	0.0003	12.18	13.29
		post			1	6.10	0.002	87.82	86.71
f.	2 moras: voiced	pre	2	1		3.18	0.042	94.89	89.32
		post			1	6.10	0.002	5.11	10.68
g.	3 moras: voiced	pre	3	1		4.53	0.011	82.84	85.30
		post			1	6.10	0.002	17.16	14.70
h.	4 moras: voiced	pre	4	1		5.87	0.003	55.68	65.30
		post			1	6.10	0.002	44.32	34.70
i.	5 moras: voiced	pre	5	1		7.22	0.001	24.64	15.27
		post			1	6.10	0.002	75.36	84.73
j.	6 moras: voiced	pre	6	1		8.57	0.0002	7.84	10.71
		post			1	6.10	0.002	92.16	89.29

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

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

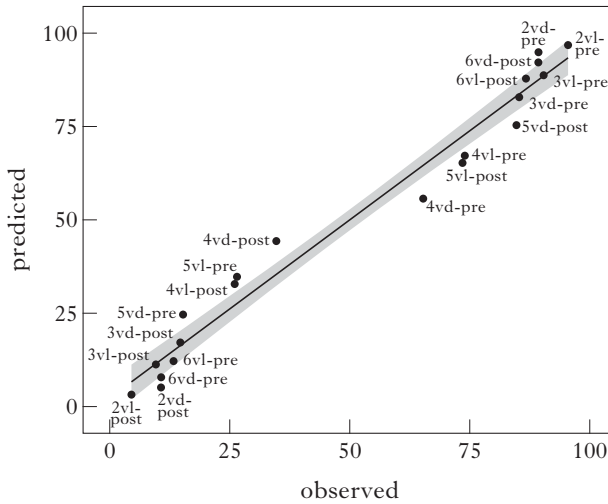

Figure 5
The correlation between the observed and the predicted percentages obtained from the MaxEnt analysis in (6).

character names *vs*. post-evolution character names. 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 Materials A). The weights obtained by this analysis are shown in the top row of the tableaux. These weights, together with the constraint profiles, allow us to calculate $\mathcal{H}$-scores, eHarmony scores and predicted percentages, using the procedure reviewed in §5.1.

The observed and predicted values are very similar. Figure 5 plots the correlation between the probabilities obtained in the experiment and the probabilities predicted by the MaxEnt model. The figure shows a good fit between the two measures, demonstrating the success of the MaxEnt analysis.

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 2020; also Hayes *et al*. 2012 and Breiss & Hayes 2020 for applications of this test in linguistic analyses). We can do this by comparing two grammatical models – for the current analysis, we compare the full model incorporating all three constraints with smaller models incorporating 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 examining their ratios, which tell us whether the full model fits the data better than the simpler models to a statistically significant degree.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

The results of these log-likelihood ratio tests are shown in Table III, which demonstrates that there is statistical justification for all three constraints playing a role in the explanation of the data (see Breiss & Hayes 2020: Appendix).

	Δlikelihood	$\chi^2(1)$	p
*$\textsc{Long}_{\text{pre-ev}}$	249.88	499.77	<0.001
*$\textsc{Vcd}_{\text{pre-ev}}$	4.10	8.20	<0.01
*$\textsc{Post}$	256.65	513.30	<0.001

Table III

The results of the log-likelihood ratio tests. The log-likelihood of the best-fitting model with the three constraints was −432.30. See Supplementary Materials A.

Next, a more complex model was tested, with a fourth constraint representing the interaction term between *$\textsc{Long}_{\text{pre-ev}}$ and *$\textsc{Vcd}_{\text{pre-ev}}$, equivalent to the locally conjoined version of these two constraints (cf. Shih 2017). The results show that addition of this constraint did not improve the model fit. The Solver actually assigned zero weight to the conjoined constraint. Even when constraint weights were allowed to be negative, the Solver assigned a weight that is very close to zero (−0.13). This is a welcome result, since the interaction of the effects of voiced obstruents and those of mora count 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 a general question regarding wug-shaped curves: whether it is possible to objectively assess if given data is best fitted with a wug-shaped curve. To reiterate, 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: (i) it consists of two sigmoid curves, (ii) the two curves are identical and (iii) they are separate. No real data would perfectly fit this mathematical definition, because it involves 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 is relatively straightforward: it can be assessed by examining the effect of the perturber. In the current analysis, the perturber corresponds to the constraint *$\textsc{Vcd}_{\text{pre-ev}}$, which was significant in the MaxEnt analysis developed in §5.2. Whether the two curves are identical can be addressed by examining

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

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 $*\text{LONG}_{\text{pre-ev}}$ and $*\text{VCD}_{\text{pre-ev}}$ were significant, we could reasonably have 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 that we can obtain two curves that are literally identical, because 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 empirical investigation, rather than being determined *a priori*. Two similar, but not 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 a question of whether we should allow interaction terms – or conjoined constraints – to play a substantial role in a MaxEnt grammar. McPherson & Hayes (2016) and Zuraw & Hayes (2017) posit no interaction terms for their analyses; Shih (2017), on the other hand, argues that constraint conjunction is required even in a MaxEnt grammar. More quantitative studies are necessary to settle this issue.

A final challenge is how to decide whether the pattern is best modelled using a sigmoid curve, 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 $p(\text{evolved}) = -0.51 + 0.228 \times \text{Mora} + 0.067 \times \text{Voiced obstruent}$. The log-likelihood of this linear model is -501.0,[13] which is worse than the sigmoidal MaxEnt model, which has a log-likelihood of -432.3. 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 fitted by the model).

However, relying on log-likelihood alone does not allow us to conclude that the sigmoid function is *the* function that underlies 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-generalisable; i.e. they would suffer from the general problem of overfitting (Good & Hardin 2006). In order to balance the goodness of the fit to the data and model

[13] See Supplementary Materials C. This model predicts that bimoraic forms without voiced obstruents should be post-evolution characters in '−5.4% of cases', 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 is one strength of MaxEnt – since harmony is negatively exponentiated, it never yields probabilities below zero.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

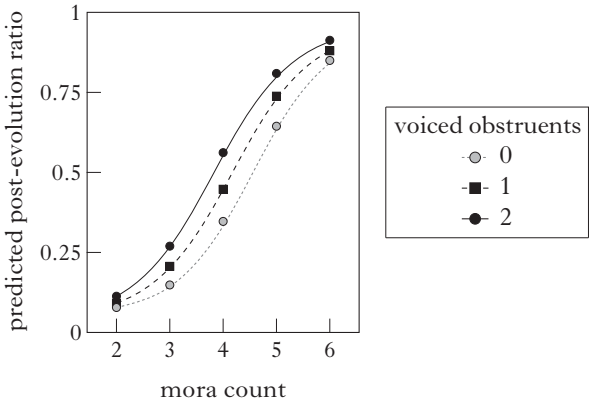

Figure 6
Predictions of the current MaxEnt model for forms with two voiced obstruents, instantiating a 'stripey wug'. See Supplementary Materials B.

complexity, additional statistical measures, such as the Akaike Information Criterion (AIC; 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 quantitative observations. For now, I am reasonably confident that mathematical functions generated by MaxEnt are suited to model cross-linguistic quantitative patterns, as reviewed in §1.1.

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 (to appear) shows that nonce names with two voiced obstruents are more likely to be judged as post-evolution character 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 $*\text{VCD}_{\text{pre-ev}}$ in (5) actually predicts this cumulative behaviour, since forms with two voiced obstruents are assigned two violation marks when they are mapped onto a pre-evolution character. Since the weights of the constraints are already calculated and the constraint-violation profiles are known, the MaxEnt model makes specific quantitative predictions.[14] These predictions are illustrated in Fig. 6, an example of a 'stripey wug' consisting of three sigmoid curves (McPherson & Hayes 2016, Zimmermann 2017,

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

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

Zuraw & Hayes 2017, Hayes 2020). While the current experiment was limited to items containing only one voiced obstruent, these predicted values can be tested in future experiments.

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

5.4 Some notes on MaxEnt and logistic regression

I note at this point that MaxEnt is mathematically equivalent to a (multinomial) logistic regression (see in particular Jurafsky & Martin 2019: ch. 5, as well as Shih 2017 and Breiss & Hayes 2020). A mixed-effects logistic regression analysis was reported in §3 as a means to test the experimental results without any particular linguistic theories or analyses in mind. On the other hand, in this section a MaxEnt analysis has been developed 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 emphasise that this MaxEnt analysis is indeed a generative phonological analysis, I employed McCarthy's (2003) OT constraint schema.

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) demonstrably offers a useful tool to discern causes and meanings in data in general, it would not be too surprising if children use logistic regression (or something akin to it) in order to find patterns in the grammar that they are learning. On this view, UG employs 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 OT (Prince & Smolensky 1993; see also Breiss & Hayes 2020), which invites the interesting question of whether UG can be reduced to a domain-general statistical tool. Providing a full 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) the nature of the vocabulary that this constraint set refers to (e.g. distinctive features such as [+sonorant] 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, and if so, to what extent (Potts & Pullum 2002, McCarthy 2003, de Lacy 2006, Crowhurst 2011,

Coetzee & Kawahara 2013, among many others). Restricting CON may be necessary to explain cases in which speakers' behaviour diverges substantially from what is predicted by the statistical patterns in the lexicon (e.g. Becker *et al*. 2011, Jarosz 2017, Garcia 2019). Additionally, UG may impose particular biases toward, for example, phonetically natural patterns, which can be formalised in the MaxEnt framework in terms of biases on constraint weights (Wilson 2006, Hayes *et al*. 2009, Hayes & White 2013). In short, UG can be a metatheory 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 in a quantitatively rigorous manner what CON consists of (Shih 2017).

6 Analyses with Stochastic Optimality Theory

Although Zuraw & Hayes (2017) and Hayes (2020) argue that patterns with wug-shaped curves cannot be modelled 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 Gaussian noise at each evaluation. Just as in Classic OT, each evaluation is computed with strict domination, predicting a single winner in each evaluation trial. The probability distributions of variable outputs are calculated over multiple evaluation cycles.

To analyse the current experimental results using Stochastic OT, the same data structure that was used for the MaxEnt analysis in (6) was fed to OTSoft (Hayes *et al*. 2014), using the Gradual Learning Algorithm (Boersma & Hayes 2001). The initial ranking values of all constraints were set at 100 (the default value). The initial plasticity and the final plasticity were set at 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 obtain the predicted probability distribution. The results of all the learning simulations presented in this section are available in Supplementary Materials D.

This learning simulation yielded the following ranking values: $*\text{LONG}_{\text{pre-ev}} = 99.6$, $*\text{VCD}_{\text{pre-ev}} = 98.1$, $*\text{POST} = 100.4$. All the constraints were active in at least one of the evaluation trials. A problem with this Stochastic OT analysis is that it was not able to model the effects of mora count at all; indeed, Stochastic OT does not handle counting cumulativity effects well in general (Jäger 2007, Hayes 2020). For all the conditions without voiced obstruents, regardless of the mora counts, post-evolution candidates were predicted to win in 40% of the cases and pre-evolution candidates in 60%. For all the conditions with voiced obstruents, post-evolution characters were predicted to win in 46.6% of the cases and pre-evolution characters in 53.4%. 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

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

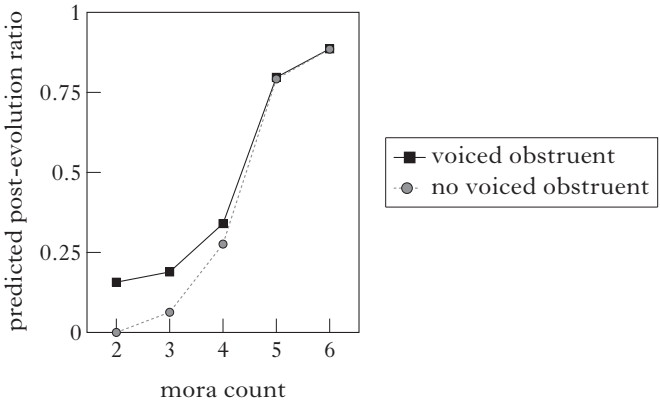

Figure 7
The probability patterns predicted by the GLA when *LONG$_{\text{pre-ev}}$ is split into a family of different constraints.

averaged across all the mora-count conditions (40.1% *vs.* 46.8%). However, it was unable to learn the mora-count effects.

The failure to model the counting cumulativity effects of mora count is due to the fact that Stochastic OT is no different from Classic OT (Prince & Smolensky 1993) at each time of evaluation. OT does not distinguish between, for example, one violation mark *vs.* two violation marks on the one hand and one violation mark *vs.* four violation marks on the other. Therefore, if *POST dominates *LONG$_{\text{pre-ev}}$ 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 *LONG$_{\text{pre-ev}}$ the pre-evolution candidate incurs. Similarly, if *LONG$_{\text{pre-ev}}$ dominates *POST, the post-evolution candidate wins, no matter how long the pre-evolution candidate is. The number of violations is irrelevant 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.

A (partial) solution to this problem involves splitting up *LONG$_{\text{pre-ev}}$ into a set of separate constraints which each penalise a pre-evolution name with a particular mora count; i.e. *LONG$(3\mu)_{\text{pre-ev}}$, *LONG$(4\mu)_{\text{pre-ev}}$, *LONG$(5\mu)_{\text{pre-ev}}$ and *LONG$(6\mu)_{\text{pre-ev}}$ (see McPherson & Hayes 2016: n. 21, as well as Boersma 1998, Gouskova 2004 and de Lacy 2006). A new learning simulation was run with the same parameter settings. With the expanded set of constraints, it learned the following values: *LONG$(3\mu)_{\text{pre-ev}} = 97.2$, *LONG$(4\mu)_{\text{pre-ev}} = 99.7$, *LONG$(5\mu)_{\text{pre-ev}} = 103.7$, *LONG$(6\mu)_{\text{pre-ev}} = 103.2$, *VCD$_{\text{pre-ev}} = 98.7$, *POST $= 101.6$. Plotting the predicted probabilities based on these ranking values results in two separate curves for the two voicing conditions, as shown in Fig. 7. However, these curves formed an 'open jaw' pattern, in which we observe the

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

convergence of the two curves at one end and divergence at the other end, with the difference between the two curves increasing monotonically toward the left (compare this pattern with Fig. 4b).

The problem comes from the fact that the ranking value of the perturber constraint – $*\text{VCD}_{\text{pre-ev}}$ – differs too much from the ranking values of $*\text{LONG}(5\mu)_{\text{pre-ev}}$, $*\text{LONG}(6\mu)_{\text{pre-ev}}$ and *POST, resulting in 'near strict domination'. As a result, $*\text{VCD}_{\text{pre-ev}}$ does not have a visible influence on 5-mora and 6-mora names. This problem is a general one (Hayes 2020): the perturber constraint can have only 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 quantitative 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 $*\text{LONG}(n\mu)_{\text{pre-ev}}$ constraints and $*\text{VCD}_{\text{pre-ev}}$ are synergistic, in that they both favour post-evolution names, while the other constraint, *POST, favours 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 Fig. 7. Citing unpublished work by Giorgio Magri, they characterise 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'. The pattern in Fig. 7 looks precisely like what Zuraw & Hayes describe, with the very minor difference that the divergence is larger on the left of the plot in Fig. 7, rather than on the right.

Bruce Hayes (personal communication) points out that Stochastic OT may perform better if the perturber constraint ($*\text{VCD}_{\text{pre-ev}}$) is reformulated in such a way that it penalises the same candidate as the binary constraint (*POST). Following this suggestion, I reformulated $*\text{VCD}_{\text{pre-ev}}$ as a constraint that penalises a post-evolution name which does not start with a voiced obstruent, as in (7).

(7) $\text{INITIALC=VCD}_{\text{post-ev}}$
Assign a violation mark for each post-evolution character name which does not start with a voiced obstruent.

This new constraint is reminiscent of a positional markedness constraint, which for example, requires a low-sonority segment in onset positions (Smith 2002). Unlike $*\text{VCD}_{\text{pre-ev}}$, it penalises a post-evolution name (rather than a pre-evolution name) when it does not have a particular property. The ranking values that the General Learning Algorithm learned with the constraint in (7) are $*\text{LONG}(3\mu)_{\text{pre-ev}} = 64.6$, $*\text{LONG}(4\mu)_{\text{pre-ev}} = 65.9$, $*\text{LONG}(5\mu)_{\text{pre-ev}} = 69.9$, $*\text{LONG}(6\mu)_{\text{pre-ev}} = 70.4$, $\text{INITIALC=VCD}_{\text{post-ev}} = 66.6$, *POST = 67.5, which yields the two curves shown in Fig. 8.

The two curves are better separated in Fig. 8 than in Fig. 7, because the ranking value of the perturber constraint, $\text{INITIALC=VCD}_{\text{post-ev}}$, is in the middle of the constraint-ranking continuum in this analysis. We can

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

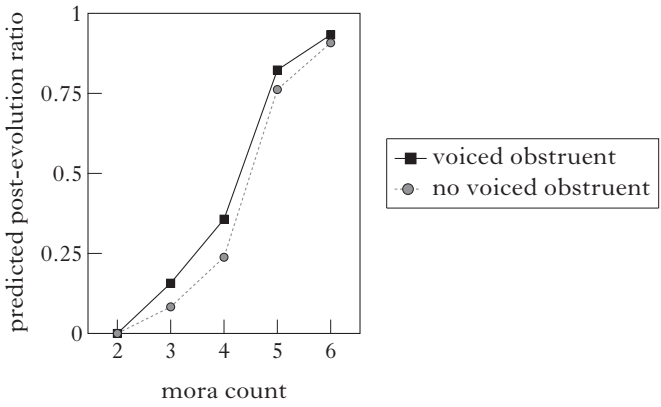

Figure 8
The probability patterns predicted by the GLA with the perturber constraint in (7).

see that the difference between the two curves is largest for 4-mora names, and 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 of the horizontal axis.

As demonstrated in this section, Stochastic OT requires that we split the scalar constraint (*LONG$_{\text{pre-ev}}$) 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 analyses here, is a general one: the perturber constraint can have only one ranking value, so its influence is localised. When it is placed in the middle of the ranking-value continuum, as in Fig. 8, we 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.

The log-likelihood – a measure of deviation between the observed data and the model predictions – of the Stochastic OT analyses was calculated. The values for the two analyses were −459.6 (Fig. 7) and −546.8 (Fig. 8).[15] 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): six *vs*. three. The AIC was therefore calculated for each model, yielding 931.2 and

[15] Since we cannot calculate the log of 0, it was replaced with $1 / 10^{-6}$.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

1105.5 for the two Stochastic OT models and 870.7 for the MaxEnt model (a model with a lower AIC makes a better prediction).[16]

7 Concluding remarks

7.1 Summary

The current project was largely inspired by the research programme 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 experiment in this paper addressed this prediction in the domain of sound symbolism, and showed that we can indeed identify wug-shaped curves when certain variables are systematically manipulated for the judgement of evolvedness in Pokémon names. To the extent that wug-shaped curves are typical quantitative signatures of MaxEnt, this shows that MaxEnt is a grammatical framework that is suitable for modelling 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 voiced obstruents) in a cumulative way, more specifically, in a way that is naturally predicted by MaxEnt.

Viewed from a slightly different – albeit related – perspective, the experiment addressed the general issue of cumulativity in sound symbolism. The effects of mora counts were an example of counting cumulativity, in that each mora count contributed to the judgement of evolvedness in a sigmoidal fashion. The overall patterns also instantiated ganging-up cumulativity, in that the effects of voiced obstruents and of mora counts additively contributed to the judgement of evolvedness. Such cumulative patterns are a natural consequence of MaxEnt.

7.2 Phonological patterns and sound-symbolic patterns

To the extent that MaxEnt is a useful tool for modelling phonological patterns such as input–output mappings and surface phonotactics judgements, 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 received hardly any serious attention from formal phonologists (but see

[16] There are two caveats. First, *Long$_{\text{pre-ev}}$ in the MaxEnt model can assign a wider range of constraint-violation marks than the set of *Long($n\mu$)$_{\text{pre-ev}}$ constraints in the Stochastic OT model does, because the former is a scalar constraint and the latter are binary constraints. Second, since the comparison between MaxEnt and Stochastic OT is based on a single case study, the arguments presented are not definitive. See Zuraw & Hayes (2017) and Breiss (2020) for other recent case studies offering quantitative comparisons of MaxEnt and Stochastic OT.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

Alderete & Kochetov 2017, Kawahara 2020b). However, the results suggest that there may be non-negligible similarities between sound–meaning mappings and phonological input–output mappings (as well as well-formedness judgements of surface phonotactic patterns). Phonological patterns and sound-symbolic patterns share two important properties, stochasticity and cumulativity, both of which follow naturally from a MaxEnt grammar. This conclusion 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, Kumagai 2019, Jang 2020, Kawahara 2020b, Shih 2020).[17]

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 answer to this general question here, but can address it partially by asking a more concrete question: whether sound-symbolic constraints of the sort used in this paper can trigger phonological changes. Alderete & Kochetov (2017) argue that such patterns do exist. Patterns of expressive palatalisation, often found in baby-talk registers, exhibit properties that are different from 'regular' phonological palatalisation processes; for example, the former can target all the coronal segments in a word without a clear trigger like a high front vowel (e.g. Japanese /osakana-saɴ/ → [oɕakaɲa-ɕaɴ] 'fish-y'). They thus argue that expressive palatalisation 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 palatalisation may thus instantiate a case in which sound-symbolic constraints coerce phonological changes. See Kumagai (2019) and Jang (2020) for other possible examples.

7.3 Closing remarks

I would like to close this paper by putting forward the following methodological thesis: phonological theory can inform research on sound symbolism. Although there is a great deal of current work on sound symbolism, most of this research has been conducted by psychologists, cognitive scientists and cognitive linguists, and few formal phonologists have paid serious attention to sound symbolism. However, the research reported in this paper has revealed important aspects of sound symbolism – its cumulative nature and how it can be modelled using MaxEnt. Hayes (2020) offers an abstract 'top-down' approach, which takes one theory seriously and considers its consequences. The research discussed here would not have been possible without this approach. More generally, then, phonological theory can inform research on sound symbolism in

[17] To this we can add studies of metrics conducted by phonologists (see, for example, Hayes *et al*. 2012 for a MaxEnt analysis of metrics). Given that metrics are a topic of phonological research, I do not see any fundamental reason to exclude sound symbolism from similar inquiry.

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

important ways. In addition, I hope to have shown that sound symbolism can offer a new testing ground for the examination of how the cumulative nature of linguistic patterns is manifested, and of how sound symbolism can inform phonological theories. More generally, the case study in this paper has shown that phonological theories and research on sound symbolism can and should mutually inform each other.

Appendix: Patterns in existing Pokémon names

We might wonder how the existing patterns of Pokémon names behave with respect to the issues discussed in the main text. To address this question, I used the dataset compiled by Kawahara *et al*. (2018), which includes all the data up to the sixth generation, about 700 characters. Some Pokémon characters do not undergo evolution at all, and those were removed from the analysis. Some others were 'baby' Pokémons, introduced as a pre-evolution version of an already existing character in a later series. While there were not many ($N = 16$), they were also excluded. Pokémons can undergo evolution twice; in the current analysis, as long as they had evolved once, they were counted as post-evolution. There was only one 6-mora name, 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 in the analysis of existing Pokémon names, Fig. 9a plots the relationship between the mora counts and the averaged probabilities of the names being used for post-evolution characters. Both a linear function (solid line) and a sigmoid curve (dashed line) were fitted to the data. There does not seem to be any 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) – shows a similar linear trend, as shown as Fig. 9b (based on Kawahara *et al*. 2018: Fig. 7).

We may tentatively conclude from Fig. 9 that sigmoid curves (and hence wug-shaped curves) emerged as a result of the experimental settings, despite the absence of such patterns in the existing names.

An anonymous reviewer raises the question of where this difference between the real names and experimental results comes from, asking if MaxEnt would force a linear pattern in the input to be converted to a sigmoidal pattern in the output. The answer is positive. Because of the mathematics that underlies MaxEnt, a scalar constraint has to result in a sigmoid curve, not a linear curve (McPherson & Hayes 2016, Zuraw & Hayes 2017, Jurafsky & Martin 2019, Winter 2019).

A question that arises is why we observe a linear pattern in the existing names, rather than a sigmoid curve. My tentative hypothesis is that, since the experiment focused on sound symbolism using nonce names, it was able to tap into how sound-symbolic knowledge is revealed in a more pure and direct form than would be the case if we had looked at the set of existing names. Sound symbolism is not the only factor that determines existing Pokémon names; other factors are also taken into consideration, such as the occasional use of real words to describe a character; e.g. *hitokage* 'fire lizard' is a kind of a lizard (*tokage*) which spits out fire (*hi*). Another complication is that the Pokémon

https://doi.org/10.1017/S0952675720000202 Published online by Cambridge University Press

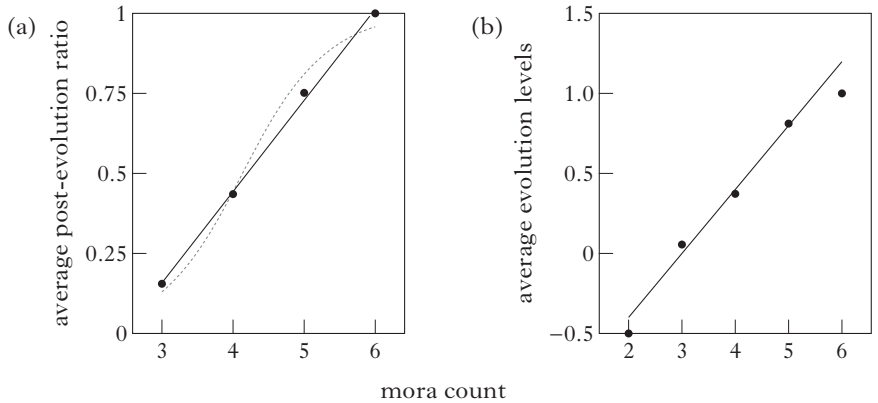

Figure 9

(a) The relationship between the mora count 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.

lexicon has evolved over a number of generations, with new characters added in each generation. The question of why the existing names show a linear pattern requires further scrutiny, but the experimental results reported here nevertheless remain encouraging, because, as we have seen, MaxEnt can have a linear input, but has to return a sigmoidal output, as confirmed by the current experiment.

REFERENCES

Akaike, Hirotugu (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petov & F. Caski (eds.) *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akadémiai Kiadó. 267–281.

Alderete, John (1997). Dissimilation as local conjunction. *NELS* **27**. 17–31.

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

Anttila, Arto & Giorgio Magri (2018). Does MaxEnt overgenerate? Implicational universals in Maximum Entropy Grammar. In Gillian Gallagher, Maria Gouskova & Sora Heng Yin (eds.) *Proceedings of the 2017 Annual Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v5i0.4260.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tilly (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. *Lg* **87**. 84–125.

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

Boersma, Paul (1998). *Functional phonology: formalizing the interactions between articulatory and perceptual drives*. PhD dissertation, University of Amsterdam.

Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *LI* **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*. London: Equinox. 389–434.
Breiss, Canaan (2020). Cumulativity by default in phonotactic learning. Ms, University of California, Los Angeles. https://ling.auf.net/lingbuzz/004747.
Breiss, Canaan & Bruce Hayes (2020). Phonological markedness effects in sentence formation. *Lg* **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**. 245–259.
Cedergren, Henrietta J. & David Sankoff (1974). Variable rules: performance as a statistical reflection of competence. *Lg* **50**. 333–355.
Chomsky, Noam (1957). *Syntactic structures*. The Hague: Mouton.
Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.
Coetzee, Andries W. & Shigeto Kawahara (2013). Frequency biases in phonological variation. *NLLT* **31**. 47–89.
Coetzee, Andries W. & Joe Pater (2011). The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.) *The handbook of phonological theory*. 2nd edn. Malden, Mass. & Oxford: Wiley-Blackwell. 401–431.
Cohn, Abigail C. (2006). Is there gradient phonology? In Gisbert Fanselow, Caroline Féry, Ralf Vogel & Matthias Schlesewsky (eds.) *Gradience in grammar: generative perspectives*. Oxford: Oxford University Press. 25–44.
Crowhurst, Megan J. (2011). Constraint conjunction. In van Oostendorp *et al*. (2011). 1461–1490.
Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann (2011). Explaining sonority projection effects. *Phonology* **28**. 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. http://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**. 367–393.
Featherston, Sam (2005). The decathlon model of empirical syntax. In Stephan Kepser & Marga Reis (eds.) *Linguistic evidence: empirical, theoretical, and computational perspectives*. Berlin & New York: Mouton de Gruyter. 187– 208.
Garcia, Guilherme Duarte (2019). When lexical statistics and the grammar conflict: learning and repairing weight effects on stress. *Lg* **95**. 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. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.
Good, Phillip I. & James W. Hardin (2006). *Common errors in statistics* (*and how to avoid them*). 2nd edn. Hoboken, N.J.: Wiley.
Gouskova, Maria (2004). Relational hierarchies in Optimality Theory: the case of syllable contact. *Phonology* **21**. 201–250.
Guy, Gregory R. (1991). Explanation in variable phonology: an exponential model of morphological constraints. *Language Variation and Change* **3**. 1–22.
Guy, Gregory R. (2011). Variability. In van Oostendorp *et al*. (2011). 2190–2213.
Haiman, John (1980). The iconicity of grammar: isomorphism and motivation. *Lg* **56**. 515–540.
Haiman, John (1984). *Natural syntax: iconicity and erosion*. Cambridge: Cambridge 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*. Cambridge, Mass.: MIT Press. 294–303.

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

Hamano, Shoko (2013). Hoogen-ni okeru giongo-gitaigo-no taikeiteki-kenkyuu-no igi. [The significance of systematic research on ideophones in Japanese dialects.] In Kazuko Shinohara & Ryoko Uno (eds.) *Chikazuku oto-to imi: onomatope kenkyuu-no shatei*. [*Sound symbolism and mimetics: rethinking the relationship between sound and meaning in language*.] Tokyo: Hitsuzi Syobo. 133–147.

Hayes, Bruce (2017). Varieties of Noisy Harmony Grammar. In Karen Jesney, Charlie O'Hara, Caitlin Smith & Rachel Walker (eds.) *Proceedings of the 2016 Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v4i0.3997.

Hayes, Bruce (2020). Assessing grammatical architectures through their quantitative signatures. Paper presented at the Berkeley Linguistics Society workshop 'Phonological representations: at the crossroad between gradience and categoricity'. Handout available (July 2020) at https://linguistics.ucla.edu/people/hayes/papers/HayesBLSTalkFebruary2020.pdf.

Hayes, Bruce & Zsuzsa Cziráky 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. *LI* **44**. 45–75.

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

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

Hayes, Bruce, Kie Zuraw, Péter Siptár & Zsuzsa Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Lg* **85**. 822–863.

Hinton, Leanne, Johanna Nichols & John J. Ohala (2006). *Sound symbolism*. 2nd edn. Cambridge: Cambridge University Press.

Itô, Junko & Armin Mester (1986). The phonology of voicing in Japanese: theoretical consequences for morphological accessibility. *LI* **17**. 49–73.

Ito, Junko & Armin Mester (2003). *Japanese morphophonemics: markedness and word structure*. Cambridge, Mass.: MIT Press.

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

Jäger, Gerhard (2007). Maximum entropy models and Stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.) *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan*. Stanford: CSLI. 467–479.

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

Jang, Hayeun (2020). How cute do I sound to you? Gender and age effects in the use and evaluation of Korean baby-talk register, *Aegyo*. *Language Sciences*. https://doi.org/10.1016/j.langsci.2020.101289.

Jarosz, Gaja (2017). Defying the stimulus: acquisition of complex onsets in Polish. *Phonology* **34**. 269–298.

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

Jurafsky, Daniel & James H. Martin (2019). *Speech and language processing*. 3rd edn. Draft available (July 2020) at https://web.stanford.edu/~jurafsky/slp3/.
Kawahara, Shigeto (2017). *A-wa i-yori ookii!? Onshouchou-de manabu onseigaku nyuumon*. [*Introducing phonetics through sound symbolism*.] Tokyo: Hitsuzi Syobo.
Kawahara, Shigeto (2020a). Cumulative effects in sound symbolism. Ms, Keio University. Available (July 2020) at http://user.keio.ac.jp/~kawahara/pdf/CumulativitySoundSymbolism.pdf.
Kawahara, Shigeto (2020b). Sound symbolism and theoretical phonology. *Language and Linguistic Compass* **14**. https://doi.org/10.1111/lnc3.12372.
Kawahara, Shigeto, Mahayana C. Godoy & Gakuji Kumagai (2020). Do sibilants fly? Evidence from a sound symbolic pattern in Pokémon names. *Open Linguistics* **6**. 386–400.
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 (to appear). What voiced obstruents symbolically represent in Japanese: evidence from the Pokémon universe. *Journal of Japanese Linguistics* **37**.
Kawahara, Shigeto, Atsushi Noto & Gakuji Kumagai (2018). Sound symbolic patterns in Pokémon names. *Phonetica* **75**. 219–244.
Kawahara, Shigeto, Kazuko Shinohara & Yumi Uchimoto (2008). A positional effect in sound symbolism: an experimental study. In *Proceedings of the 8th Annual Meeting of the Japan Cognitive Linguistics Association*. Tokyo: JCLA. 417–427.
Kawahara, Shigeto, Michinori Suzuki & Gakuji Kumagai (2020). The sound symbolic patterns in Pokémon move names in Japanese. *ICU Working Papers in Linguistics* **10**. 17–30.
Kisseberth, Charles W. (1970). The treatment of exceptions. *Papers in Linguistics* **2**. 44–58.
Kroch, Anthony S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change* **1**. 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**. 39–74.
Labov, William (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.
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*. 2nd edn. Vol. 1. Berlin & New York: de Gruyter. 6–21.
Leben, William R. (1973). *Suprasegmental phonology*. PhD dissertation, MIT.
Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman & Belver C. Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* **54**. 358–368.
McCarthy, John J. (1983). Phonological features and morphological structure. In John F. Richardson, Mitchell Marks & Amy Chukerman (eds.) *Papers from the parasession on the interplay of phonology, morphology, and syntax*. Chicago: Chicago Linguistic Society. 135–161.
McCarthy, John J. (2003). OT constraints are categorical. *Phonology* **20**. 75–138.
McCarthy, John J. & Alan Prince (1986). *Prosodic morphology*. Ms, University of Massachusetts, Amherst & Brandeis University.
McPherson, Laura & Bruce Hayes (2016). Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology* **33**. 125–167.
Martin, Samuel E. (1962). Phonetic symbolism in Korean. In Nicholas Poppe (ed.) *American studies in Altaic linguistics*. Bloomington: Indiana University Press. 177–189.

Myers, Scott (1997). OCP effects in Optimality Theory. *NLLT* **15**. 847–892.
Nishimura, Kohei (2006). Lyman's Law in loanwords. *Phonological Studies* **9**. 83–90.
Oostendorp, Marc van, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.) (2011). *The Blackwell companion to phonology*. Malden, Mass.: Wiley-Blackwell.
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 B. (2001). Stochastic phonology. *Glot International* **5**. 195–207.
Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker (2010). Harmonic Grammar with linear programming: from linear systems to linguistic typology. *Phonology* **27**. 77–117.
Potts, Christopher & Geoffrey K. Pullum (2002). Model theory and the content of OT constraints. *Phonology* **19**. 361–393.
Prince, Alan & Paul Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published 2004, Malden, Mass. & Oxford: Blackwell.
Ramachandran, V. S. & E. M. Hubbard (2001). Synesthesia: a window into perception, thought, and language. *Journal of Consciousness Studies* **8**. 3–34.
Schütze, Carlson T. (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**. 243–268.
Shih, Stephanie S. (2020). Gradient categories in lexically-conditioned phonology: an example from sound symbolism. In Hyunah Baek, Chikako Takahashi & Alex Hong-Lun Yeung (eds.) *Proceedings of the 2019 Annual Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v8i0.4689.
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). Crosslinguistic 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 & University of Chicago. Available (July 2020) at http://user.keio.ac.jp/~kawahara/pdf/PokemonasticsShihEtAl.pdf.
Sidhu, David M. & Penny M. Pexman (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin and Review* **25**. 1619–1643.
Smith, Jennifer L. (2002). *Phonological augmentation in prominent positions*. PhD dissertation, University of Massachusetts, Amherst.
Smolensky, Paul (1986). Information processing in dynamical systems: foundations of Harmony Theory. In D. E. Rumelhart, J. L. McClelland & the PDP Research Group (eds.) *Parallel Distributed Processing: explorations in the micro-structure of cognition*. Vol. 1: *Foundations*. Cambridge, Mass.: MIT Press. 194–281.
Smolensky, Paul (1995). On the internal structure of the constraint component *Con* of UG. Ms, Johns Hopkins University. Available as ROA-86 from the Rutgers Optimality Archive.
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. Available as ROA-939 from the Rutgers Optimality Archive.

Thompson, Patrick D. & Zachary Estes (2011). Sound symbolic naming of novel objects is a graded function. *Quarterly Journal of Experimental Psychology* **64**. 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**. 945–982.
Winter, Bodo (2019). *Statistics for linguists: an introduction using R*. New York: Routledge.
Zimmermann, Richard (2017). *Formal and quantitative approaches to the study of syntactic change: three case studies from the history of English*. PhD dissertation, University of Geneva.
Zuraw, Kie & Bruce Hayes (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Lg* **93**. 497–548.