Volume 29(2), August 31, 2023, 247-269
DOI: 10.17959/sppm.2023.29.2.247

Studies in Phonetics, Phonology and Morphology

ISSN: 1226-8690 / e-ISSN 2671-616X
http://www.phonology.or.kr

Probability and randomness in phonology: Deep vs. shallow stochasticity

Sam Tilsen (Cornell University)*

Abstract

This paper argues that there are two different ways that probability and randomness can be used in models of phonological cognition. One of these is termed shallow stochasticity and refers to a situation in which probabilities are cognitively represented and provide a causal basis for variation in behavior. The other is termed deep stochasticity and refers to a situation in which probabilities are not cognitively represented but merely useful due to our ignorance of the detailed dynamics of a system. The argument is made specifically in relation to the framework of Maximum Entropy Harmonic Grammar (MaxEnt) and its theoretical analysis of the phonological pattern of nasal substitution in Tagalog. A number of critiques of the MaxEnt approach are presented.

Keywords

Maximum Entropy Harmonic Grammar, MaxEnt, probability, randomness, stochasticity, Tagalog

* E-mail: tilsen@cornell.edu
Received: July 28, 2023; Revised: August 8, 2023; Accepted: August 9, 2023

1. Introduction

What role should probabilities—and the randomness they are associated with—play in our phonological analyses? In this paper I describe two different ways in which probability and randomness can be incorporated into phonological reasoning. I argue that one of these is preferable to the other. The preferred use of randomness I call *deep stochasticity*. This refers to the idea that the systems responsible for phonological cognition produce apparently random patterns because we are largely ignorant of the details of their dynamics, and so probabilistic description of phonological behavior is a practical necessity. This conception of stochasticity is "deep" because randomness is attributed a causal role prior to probabilistic description. In other words, the usefulness of probabilistic description is a consequence of underlying randomness. Note that here I use the technical sense of the word *random* in which any particular outcome might not be predictable yet relative frequencies of repeated outcomes may be; this should not be confused with the lay sense of *random*, which refers to the lack of any definite pattern.

In contrast to deep stochasticity, I call the dispreferred use of probability *shallow stochasticity*. In models with shallow stochasticity, the brain represents probabilities of behavioral variants, and these cognitive representations of probabilities are the basis for variation in behavioral patterns, including variation in novel word form production, and in acceptability intuitions. This view reifies probabilities—it entails that the nervous system in actuality calculates quantities of this sort. This conception is "shallow" because randomness is attributed a causal role subsequent to a probabilistic representation.

To elaborate on the distinction between shallow and deep stochasticity I focus here on the theoretical framework of Maximum Entropy Harmonic Grammar (Goldwater et al. 2003, Hayes and Wilson 2008, Hayes 2022), or *MaxEnt* for short. MaxEnt derives from Harmonic grammar (Smolensky and Legendre 2006), which in turn is an adaptation of Harmony theory (Smolensky 1986). It a far beyond the scope of this brief review to address additional variants of constraint-based phonological theories. Thus we will focus solely on a series of critiques of MaxEnt that relate to the issue of shallow vs. deep stochasticity. I chose MaxEnt for this purpose simply because it is a prominent phonological model that incorporates probability. Furthermore, for the sake of having a concrete empirical example to consider, we will examine the phonological pattern of Tagalog nasal substitution (Zuraw 2000, 2010), which is

analyzed extensively in the MaxEnt framework (Hayes 2022, Zuraw and Hayes 2017).

The theoretical argument that I critique is illustrated schematically with an inverted triangle in Figure 1A. Note that this representation is my interpretation of arguments presented in Hayes (2022) and in Zuraw and Hayes (2017). The left leg of the triangle represents the idea that speakers learn the weights of a MaxEnt harmonic grammar from the statistics of phonological patterns in their language. In practice the weights of the grammar are fit via maximum likelihood to proportions of morphophonologically conditioned variants counted in a Tagalog-English dictionary (Zuraw 2000). It is not clear how well the statistical patterns of word form variants in the dictionary in actuality reflect the linguistic experience of speakers, but it is reasonable to assume that there is some correspondence between the two. We will not critique this aspect of theoretical argument since it does not relate directly to issues of stochasticity and probability. However, we will examine the fit between the predictions of the grammar and the lexicon, in order to shed light on the nature of the grammar.

The top leg of the triangle represents the observation that there is a correlation between the statistical patterns in the dictionary and behavioral patterns observed in a nonce word acceptability judgement task, conducted by Zuraw (2000). Specifically, certain phonological factors that predict variation in the dictionary data have similar predictive power in relation to average differences in acceptability ratings of nonce words. We will not critique this aspect of the theoretical argument either, since it is quite sensible that nonce-word acceptability judgments are correlated with lexical statistics or a model fit thereof. However, we will examine the correlation with the purpose of assessing how strongly the lexical patterns and the grammar/model predict the behavioral ones.

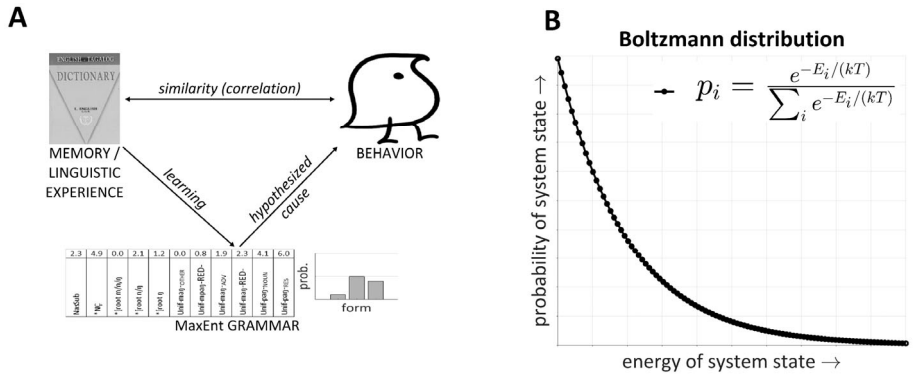


Figure 1. (A) Schematic representation of the theoretical argument critiqued in this paper. (B) A generic depiction of a Boltzmann distribution, which is similar to the equation used to generate probabilities in MaxEnt Harmonic grammars.

It is the rightmost leg of the triangle in Figure 1(A) that is the primary focus of the critique herein. This leg represents the theoretical claim that the MaxEnt grammar and the probabilities of word form variants that it generates are in some way the cause of behavioral variability. As stated by Zuraw and Hayes: "the purpose of the grammar is to generate [the] probabilities" (2017: 504) and the MaxEnt model "is intended as an account of the productive behavior of Tagalog speakers" (2017: 505). The theory thus purports to explain the correlation between lexical and behavioral patterns vis-a-vis the grammar. In my opinion, this requires an undesirable view of the role of probability and randomness in phonological cognition, one which I contend uses randomness in a shallow way.

One important aspect of my argument is based on the fact that MaxEnt makes use of a particular equation to generate probabilities of word form variants. That equation is very similar to an equation used in statistical physics, known as the Boltzmann (or Gibbs) distribution, shown in Figure 1(B). The Boltzmann distribution expresses the probability that a physical system will be in a particular state, as a function of the energy of that state. The motivation for use of the Boltzmann distribution in a cognitive context was made explicit in Harmony theory (Smolensky 1986). The crucial point here is that MaxEnt, by using the Boltzmann distribution, draws an analogy between the domain of phonological cognition and the domain of physical systems to which the Boltzmann distribution applies (i.e., closed systems in thermal

equilibrium with their surroundings). Later on we will examine and assess some of the mappings that constitute this analogy. This examination is valuable because the ways in which the analogy breaks down are related to the distinction between shallow vs. deep stochasticity.

The majority of this paper will be devoted to the critique of the role of probability and randomness in MaxEnt. For the sake of having a concrete example, Section 2 will present: (i) some details of the Tagalog nasal substitution pattern, (ii) the MaxEnt analysis of the patterns in the dictionary data, and (iii) the correlation between this pattern and nonce word acceptability judgements. Section 3 will elaborate on a general critique of the role of probability and randomness in the MaxEnt framework along with a number of specific critiques. The reader should be aware that this short paper will focus almost entirely on critiques of MaxEnt. Section 4 will summarize and contextualize the critiques, and briefly offer an alternative perspective. The broader issue surrounding this paper is the question of whether a probability-generating "grammar" (of the sort proffered by MaxEnt) is desirable to understand linguistic behavior. The critique of MaxEnt presented here is one small part of the argument that such grammars may be both undesirable and unnecessary. The in-depth exploration of alternatives is deferred for future work.

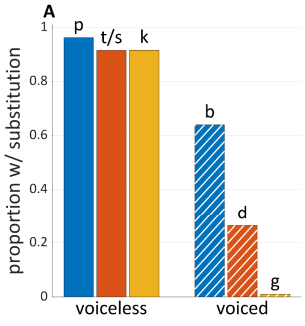
2. Tagalog nasal substitution in the lexicon and the MaxEnt analysis

The Tagalog nasal substitution pattern is a morphophonological alternation that occurs in word forms that combine certain nasal-final prefixes with obstruent-initial stems. In this environment there are two main alternants that are observed: a nasal assimilation pattern and a nasal substitution pattern. The assimilation pattern is shown in the examples of (1a), where one can see that the final nasal of the prefix assimilates in place to the stem-initial obstruent. The substitution pattern, shown in (1b), takes this one step further and replaces the stem-initial obstruent with the place-assimilated nasal. The prefix-final nasal may be analyzed as a floating [+nasal] feature (Zuraw 2010). Further, there are various phonological arguments that the substituting nasal segment is associated with the stem, rather than the prefix. For example, in reduplication constructions that target the stem, the substituting nasal is reduplicated. Note that the examples presented here and below are from De Guzman (1978) and Zuraw (2000).

(1a)	/paŋ + pitas/	>	pam- pitas	‘something for picking’
	/paŋ + súlat/	>	pan- súlat	‘writing instrument’
	/paŋ + kamot/	>	paŋ- ŋuha	‘something for scratching’
(1b)	/maŋ pitas/	>	ma- mitas	‘to distribute’
	/maŋ súlat/	>	ma- núlat	‘to write professionally’
	/maŋ kuha/	>	ma- ŋuha	‘to go collecting’

Importantly, the nasal substitution pattern is primarily a case of *lexical* variation, not *free* variation (Zuraw 2010). In free variation, a given word form may be produced in different ways by the same speaker. In contrast, in lexical variation, the members of a set of word forms that contain a common phonological environment exhibit differing, lexically determined patterns of realization—each particular word form is generally realized the same way, but the particular realization varies across the members of the set. In the case of the Tagalog nasal substitution pattern, we might characterize the set as: “word forms with a nasal-final prefix before an obstruent-initial stem,” and the pattern of realization that any particular word exhibits is either the nasal assimilation (1a) or the nasal substitution (1b). That said, there do appear to be a small number of forms produced with variation within speaker, i.e. forms in free variation (see De Guzman 1978). Similar patterns in related languages might also lead us to suspect that there are substantial between-speaker differences in cases of lexical variation (Kurniawan 2018), and generally we should expect a variety of sources of variation in naturalistic data (see Cohn and Renwick 2021).

There are two phonological factors that influence the likelihood of a word form exhibiting the substitution or assimilation pattern: the voicing of the stem-initial obstruent, and the place of the stem-initial obstruent. The effects of both factors are shown in Figure 2A, which plots the proportion of relevant word-forms with the substitution pattern from the Tagalog-English dictionary (data for this plot were extracted from visual inspection of Figure 1 of Zuraw (2010)). The *voicing effect* is a tendency for nasal substitution to occur more frequently when the stem-initial obstruent is voiceless. The *place effect* is a tendency for nasal substitution to occur more frequently with a more anterior place of articulation in the stem-initial obstruent, especially within the voiced obstruent series (b > d > g).



B

Weights:	2.3	4.9	0.0	2.1	1.2	0.0	0.8	1.9	2.3	4.1	6.0		
/paŋ$_{noun}$ + t.../	NasSub	*NÇ	*[root m/n/ŋ	*[root n/ŋ	*[root ŋ	Unif-maŋ-$_{OTHER}$	Unif-mapaŋ-RED-	Unif-maŋ-$_{ADV}$	Unif-maŋ-RED-	Unif-paŋ-$_{NOUN}$	Unif-paŋ-$_{RES}$	*Harmony*	*probability*
(SUB): [pa-n...]	0	0	1	1	0	0	0	0	0	1	0	6.2	0.73
(ASSIM): [pan-t...]	1	1	0	0	0	0	0	0	0	0	0	7.2	0.27

Figure 2. (A) Proportions of forms in the Tagalog-English dictionary with substitution, for various stem-intial obstruents (data from Zuraw (2010)). (B) Example MaxEnt analysis of the pattern, adapted from Zuraw and Hayes (2017). Base harmony constraints are shaded green; prefix-specific uniformity constraints shaded light red.

The MaxEnt grammar that Zuraw and Hayes (2017) fit to the lexical data employs two different groups of constraints, distinguished in the example tableaux of Figure 2(B), which is adapted from Zuraw and Hayes (2017). On one hand, there are constraints that contribute to the *base harmony* of an output candidate—these are shaded green in Figure 2(B). These constraints are *not* specific to the identity of the prefix. The constraint NasSub favors nasal substitution by penalizing nasal+obstruent sequences (where "+" is a morpheme boundary). *NÇ favors substitution in voiceless environments over voiced ones by penalizing a sequence of a nasal and voiceless obstruent. Constraints *[root m/n/ŋ, *[root n/ŋ, and *[root ŋ are used to create a place-dependent preference for substitution.

The other group of constraints are prefix-specific uniformity constraints, shaded light red in Figure 2(B). These constraints are responsible for prefix-dependent penalization of substitution. They accomplish this by penalizing outputs in which features from the prefix and stem correspond to the same output segment (which happens only in the substitution pattern). Note that for a given input, which specifies just one prefix, only one prefix-specific uniformity constraint is applicable.

The MaxEnt constraint weights were optimized in Zuraw and Hayes (2017) to fit the empirical proportions of nasal substitution for the six most common prefixes in the Tagalog-English dictionary, with word forms grouped by prefix identity and stem-initial obstruent. The empirical proportions from the dictionary are shown with colored points in Figure 3(A), and the MaxEnt predicted probabilities of substitution

are shown with solid lines. The base harmony—which depends on the identity of the stem-initial consonant (i.e. its place and voicing) is shown in the horizontal axis, and the harmonies associated with the prefix-specific uniformity constraints are shown in the legend. To understand the roles of these two portions of harmony, first notice that the base harmonies associated with voiceless stem-initial consonants (p, t/s, and k) are less than those associated with voiced ones (b, d, g). Similarly, notice that within the voiceless and voiced sets, harmony decreases with anteriority (p < t/s < k, and b < d < g). Second, notice that the prefix associated with the lowest weighted uniformity constraint (i.e. maŋ- other) has the strongest bias towards substitution, while the prefix associated with the highest weighted uniformity constraint (paŋ- reservational) has the weakest bias towards substitution.

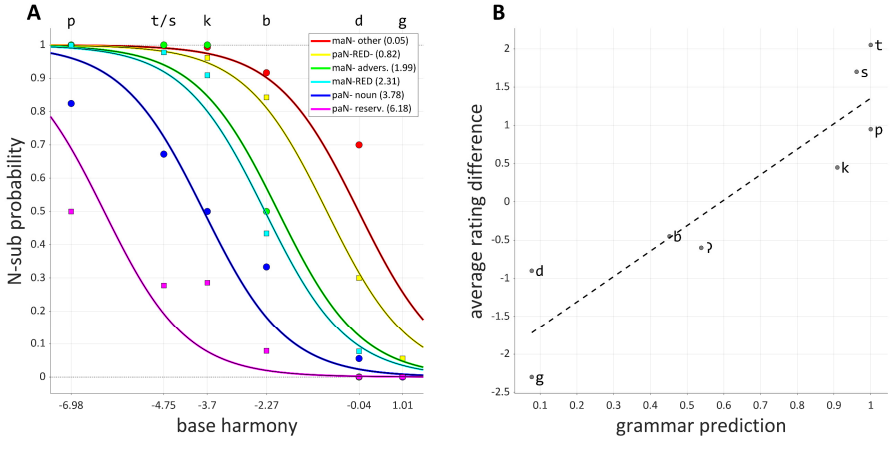


Figure 3. (A) Empirical proportions of nasal substitution from the dictionary (points) and grammar-predicted probabilities (solid lines) as a function of base harmony. Individual lines correspond to different prefixes; prefix-specific uniformity constraint weights are included in the legend. (B) Scatterplot and linear fit of average rating differences and grammar-predicted probabilities of substitution, for each stem-initial consonant in the acceptability judgement study.

Although the model fit can be quantified, it is hard to characterize the quality of the fit in a meaningful way, since it is unclear what the relevant comparisons should be (i.e. fits of lexical variation in different languages that might be accepted as "good" or "bad"). Hayes (2022) describes the fit (visually) as "reasonably good", but there

are nonetheless some sizable discrepancies, particularly when the empirical proportions are further from the floor/ceiling. In any case, we are not too concerned here with the fit of the model to dictionary data, since the aspect of the theoretical argument we are focused on is the hypothesis that the grammar accounts for the productive behavior of speakers. As stated by Zuraw and Hayes (2017), the model "is intended as an account of the productive behavior of Tagalog speakers when they apply /ŋ/-final prefixes in creating novel forms".

To that end, Zuraw (2000) conducted two tasks to investigate productive behavior of speakers, a nonce-word production task and a nonce-word acceptability judgment task. In the production task, participants "did not reliably display the voicing or place effects" (Zuraw 2010), so I will focus on the acceptability judgement task, where those effects were indeed observed. Participants in this task were shown hand-drawn pictures of farming activities along with two sentences containing a novel root. The first sentence introduced the novel root in a morphological construction that does not allow for nasal substitution, so that participants are made aware of the underlying form of the stem-initial consonant. The second sentence showed the novel root in the agentizing maŋ-red construction, in either the assimilated or substituted variant. Participants rated the sentence pair for its acceptability on a 1-10 scale. Zuraw constructed 25 novel roots for the task, and 50 cards, one with each variant of each root. During the experiment, the participants rated sentences with both the assimilated and substituted forms of each root, and therefore for each participant/novel root an acceptability rating difference score can be calculated, representing the preference the participant had for the substitution pattern in that root.

The average rating difference scores showed a voicing effect: the nasal substitution pattern was preferred to a greater extent in roots with voiceless initial consonants than ones with voiced initial consonants. The place effect was a bit less clear. To assess how strongly the behavioral pattern correlates with the empirical patterns and grammar predictions, I visually estimated the average rating differences and 95% confidence intervals from Figure 14 of Zuraw (2010). I then calculated the correlation coefficient between these averages and both the dictionary proportions and grammar predictions. A scatterplot of the latter is shown in Figure 3(B), where the horizontal axis is the MaxEnt grammar predicted proportion of substitution for the maŋ-red prefix, and the vertical axis is the rating difference averaged over participants and roots with identical stem-initial consonants. Note that positive values of the rating difference indicate a preference for substitution. The plot shows what

appears to be a strong positive correlation ($R^2 = 0.81$). The correlation strength was very similar when the dictionary proportions were used in place of the model predictions (not shown; $R^2 = 0.82$). These strong correlations would seem to support the inference that either the grammar or lexical statistics do in fact govern behavior.

However, it is important to consider that the average rating differences in the nonce-word acceptability task do not capture variation between participants or between novel roots with the same initial consonant. Indeed, there are only eight datapoints on this highly aggregated level of analysis. In order to get a sense of how strong the correlation between grammar prediction and behavior is at the level of individual roots/participants, I used the 95% confidence intervals to infer the standard deviations of the rating differences for each stem-initial consonant (i.e., from the formula: 95% c.i. = $\mu \pm 1.96\ \sigma n^{-1/2}$). I note that the confidence intervals in Zuraw (2010) are symmetric and appear to reflect the assumption that rating differences are normally distributed. Since the rating differences are constrained by the finite endpoints of the acceptability scale, this assumption is not fully tenable, and a more sophisticated analysis would incorporate random effects of speaker and root. Nonetheless, using the inferred standard deviations of rating differences, I simulated 1000 rating differences for each stem-initial consonant, and calculated the correlation between the participant/token-level rating differences and the grammar predictions/dictionary proportions. The correlation strength fell drastically in both cases, to an R^2 of 0.10. This suggests that the MaxEnt grammar or lexical statistics can weakly predict behavior in token-level data.

Although I am primarily interested in assessing the probability-related aspects of theoretical argument for MaxEnt, the weak correlation between model and behavior is noteworthy, since we might prefer a model that is able to account for more variation in behavior. Nonetheless, there does appear to be a non-negligible correlation between the dictionary patterns and behavioral ones, and so this correlation begs for an explanation. To its credit, the theoretical argument advanced by Zuraw and Hayes does offer an explanation vis-à-vis the MaxEnt grammar.

3. Critiques of MaxEnt

Here I discuss a number of specific critiques of the MaxEnt model of phonological cognition, which relate in various ways to the general distinction between shallow and deep stochasticity introduced earlier. Before elaborating on these critiques, we

need to examine the physical analogy on which MaxEnt is based, starting with the Boltzmann distribution. The reader might wonder: why does it matter that the mathematical procedure for assigning probabilities to candidates in MaxEnt is drawn from an equation used for physical systems? Indeed, one might have the opinion that it is entirely irrelevant. I believe that it matters greatly, since the analogy presupposes that our conception of phonological cognition is parallel in at least some ways to our conception of physical systems—otherwise there would be no basis for the analogy and the use of the equation would be arbitrary. Indeed, I contend that examining the ways in which the analogy breaks down can help us assess the theory itself.

The Boltzmann distribution applies to a wide range of physical systems, from ideal gases to spin glasses (e.g., the Ising model), to semiconductors. I asked ChatGPT to "give me a long list of systems that the Boltzmann distribution applies to" and it returned a list of twenty-eight examples. One of these that I think is particularly useful for expository purposes is the case of biological macromolecules (e.g. proteins, nucleic acids), which are large molecules found in all living organisms. Macromolecules can also be described more generally as polymers, which are chains of smaller molecules connected through chemical bonds. As represented schematically in Figure 4(A), a macromolecular system is often embedded in a fluid surroundings, which consists of many small molecules that are in constant motion. The system exchanges energy with its surroundings via collisions with the molecules of the fluid. A typical polymer can obtain many different configurations (i.e. states), which require different amounts of energy. For the hypothetical case shown here, there are 15 different configurations and the energy of a configuration (often called a microstate) depends on the orientation of the bonds between units, as shown in Figure 4(B). In this particular hypothetical example, the bonds between the second and third and between the fifth and sixth units are allowed to obtain one of several possible orientations, while the other bonds must be oriented at 180° angles. The Boltzmann distribution describes the probability that the system is in a particular microstate as a function of the energy of that state, with higher energy states always being less probable than lower-energy ones (Figure 4(C)). Note that this characteristic of the relation between energy and probability is a direct consequence of the negative exponential in the numerator of the equation for the Boltzmann distribution (see Figure 1B and Table 1 below). Note also that by convention the lowest energy is taken as a reference level and assigned a value of zero.

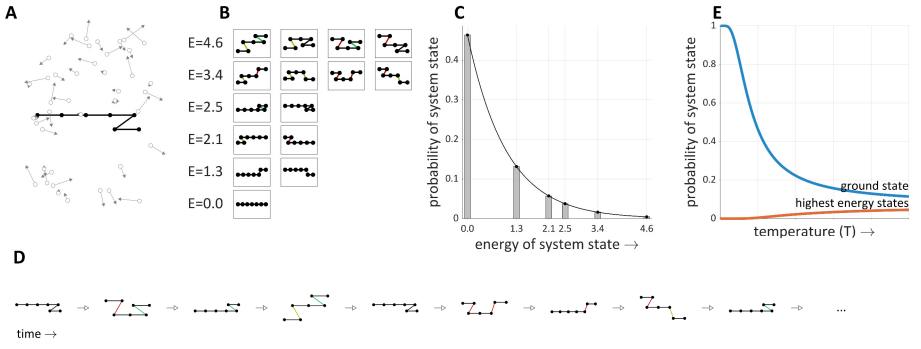


Figure 4. Illustration of properties of a hypothetical physical system (i.e. a polymer) to which the Boltzmann distribution applies. (A) The polymer exchanges energy with its surroundings. (B) Possible configurations of the system and associated energies. (C) Probabilities of configurations as a function of energy. (D) Example time evolution of the system. (E) Higher energy states become relatively more likely as temperature increases.

The metaphoric mappings associated with the MaxEnt analogy between phonological cognition and physical systems are shown in Table 1 below, which also shows the equations used to obtain probability in each domain. In the phonological domain, P_i is the probability of an output candidate *i*, and in the physical domain it is a probability of a microstate *i*. Furthermore, H_i is the harmony of a candidate and E_i is the energy of a state. Z is a normalizing factor in the phonological domain, and is the sum of the negative exponentials of all output candidates, i.e. $\boldsymbol{Z = \sum_i e^{-H_i}}$; in the physical domain Z plays the same role and is called the partition function.

Table 1. Core mappings of the MaxEnt analogy

phonological cognition	physical systems
$p_i = \frac{e^{-H_i}}{Z}$	$p_i = \frac{e^{-\frac{E_i}{kT}}}{Z} \rightarrow \frac{e^{-E_i}}{Z}$
word form representations (i.e. output candidates)	microstates (e.g. conformations of a molecule)
constraint violations	types of contributions to energy
harmony	total energy

The first and perhaps most fundamental mapping of the analogy is the metaphor that word form representations (i.e. surface candidates in harmonic grammar) are microstates of a physical system (e.g., conformations of a molecule). This mapping is implied by the indices *i* in both domains: *i* indexes word forms in the phonological domain and microstates in the physical domain. Another mapping is that constraint violations are different types of contributions to energy. This is reflected by the fact that harmony is a sum of individual terms associated with separate constraints; likewise, the energy of a physical system can be described as the sum of component energies. (In the case of the polymer example above, differently oriented bonds are associated with different energies; of course in actual physical systems the energy calculation is much more complicated than this). Finally, harmony is the total energy associated with a particular state of the system—this mapping is most clear when the constants *k* (the Boltzmann factor, which converts from units of temperature to units of energy) and *T* (the temperature) are removed from the equation (which follows from setting the product *kT* to 1). Note that *T* is constant for systems described by the Boltzmann distribution, since they are in thermal equilibrium with their surroundings.

A very important characteristic of physical systems described by the Boltzmann distribution is that they can change state from moment to moment, which is illustrated for the hypothetical polymer in Figure 4(D). This is possible because the systems are closed but *not* isolated from their surroundings, meaning that they can exchange energy but not matter with their surroundings. This is crucial because these energy exchanges are responsible for the apparently random changes from one microstate to another. For biological macromolecules such as proteins, many of these systems exist in cytoplasm, which (to oversimply a bit) consists of smaller water molecules and various ions. Collisions with these smaller molecules transfer energy to and from the macromolecule, thereby allowing it to obtain different conformations from one moment to the next.

A second very important characteristic of physical systems described by the Boltzmann distribution is that they are in thermal equilibrium with their surroundings. Hence the temperature of the system is equivalent to the temperature of the surroundings. In a fluid, temperature relates to the average kinetic energy of the molecules of the fluid, and so when the surroundings of a macromolecule have a higher temperature, energy exchanges between the molecule and its surroundings are greater, allowing higher energy states to be more probable at higher temperatures

than at lower temperatures. This is shown Fig. 4E, which plots the probabilities of the lowest energy (ground) state and any one of the highest energy microstates as a function of temperature. A third characteristic of these physical systems that applies in many cases is ergodicity: over time the system obtains all of its accessible microstates, with relative frequencies that are described by the probabilities of the Boltzmann distribution.

The key idea to have in mind when reasoning about the sorts of physical systems that are described by the Boltzmann distribution is that from one moment to the next they may change state, and the causes of such changes are numerous interactions with their surroundings that we are unable to observe. Crucially, the probabilities specified by the Boltzmann distribution are *not* the cause of these changes: the probabilities do not determine directly or indirectly what state the system will be in from one moment to the next. The probabilities are merely a description of the likelihood that the system will be in a state with a particular energy, should we choose to observe the state of the system at some time. There does not exist a miniature demon who rolls weighted dice (based on upon probabilities) to determine the state the system will obtain. Rather, the probabilistic description is useful because we (as observers of the system) cannot keep track of all of the detailed information about the system and its surroundings that would be necessary to predict the state of the system with a deterministic model.

A more familiar example of this understanding of randomness might be the casino game craps, in which players role a pair of dice on the table and place various bets on the outcome. The outcome of the dice roll is not causally determined by probabilities. Rather, the outcome is deterministic and could be predicted exactly if we had sufficient knowledge of the relevant system and surroundings—that is, the momenta of all of the air molecules in the surroundings, the exact physical structure of the table and the dice, the initial orientations and momenta of the dice upon release, etc. In fact, recent studies suggest that professional craps players are able to control outcomes to some extent by purposefully releasing the dice in a particular configuration with rotation along a specific axis (Scott III and Smith 2020)—these professionals are merely making use of the fact that the dice are a deterministic system.

Thus, when we say that the physical system is “random” or “stochastic”, this is not because there are no specific causes of changes in the state of the system from one moment to the next, but rather, because we are ignorant of the details that would be

necessary to predict those changes. This state of affairs is what I refer to as *deep stochasticity*: a probabilistic description of a system is useful to due to our ignorance of the detailed dynamics. Based on my understanding, I believe this conceptualization is appropriate for systems that are governed by classical physical laws, but perhaps not for quantum systems.

Given the above understanding of physical systems for which the Boltzmann distribution is appropriate, we can now examine the MaxEnt model of phonological cognition and critique it in various ways, many of which relate to disanalogies between phonological cognition and physical systems. One general disconnect between the MaxEnt conception of phonological cognition and physical systems relates to the actual representation of probabilities. In physical systems, probabilities are never calculated by the systems themselves, but rather, by an observer (often a physicist). In contrast, at least some practitioners of MaxEnt appear to hold the belief that the nervous system actually calculates probabilities of output candidates. For example, Zuraw and Hayes (2017) state that "the purpose of the grammar is to generate the probabilities", and Hayes states that "MaxEnt is probably an innate cognitive capacity" (2021: 29). I will refer to this notion—that the nervous system calculates probabilities (or some more or less direct representation thereof)—as the *reificationist interpretation* of MaxEnt, because it entails that something fairly abstract (probabilities) are represented concretely in the brain.

My first specific critique of MaxEnt is that it does not address the *actuation problem* in behavior, which is: given a cognitive representation of word form probabilities, how do language users actually decide which forms to produce? And how are acceptability intuitions (e.g. of the sort obtained in Zuraw (2000)) generated? I am not aware of any specific proposals in the MaxEnt framework regarding how probabilities are mapped to behaviors. In a sense, MaxEnt is missing a specific mechanism to translate from the probabilities of word forms that are generated by the grammar to the actual choices that humans make in nonce-word production or acceptability tasks. Such a mechanism cannot be too simple. In nonce-word production, one cannot simply choose the most probable form, since that would not generate variation. Perhaps the nervous system could implement some sort of random sampling mechanism, with the sampling of discrete outcomes biased by the MaxEnt probabilities. It may be a failure of my own imagination, but to me it is not clear how the nervous system would accomplish this. Whatever the solution, we would seem to have a circumstance in which probabilities dictate behavioral

outcomes, with random sampling intervening between an underlying cognitive representation of probability and behavior. This is an example of what I mean by shallow stochasticity: probabilities are used to impose randomness in behavior; probabilities somehow causally dictate behavior. It differs substantially from the deep stochasticity of physical systems, where our ignorance of detail makes a probabilistic description useful.

My second critique is that certain arguments for MaxEnt anthropomorphize phonological cognition. Anthropomorphization is when we attribute human-like behavior to non-human entities. For example, the statement "when the Earth needs a drink, Mother Nature makes it rain" conceptualizes both the planet Earth and the weather as agents with human properties. Anthropomorphization is very common rhetorically but I suggest that it can be counterproductive to deeper understanding. In Hayes (2021), a main argument for MaxEnt is that it a form of common sense reasoning: "MaxEnt [is] a mathematicized embodiment of common sense" (2021: 3) and is a consequence of adopting "principles of effective inductive reasoning" (2021: 32). Hayes (2021) lists a number of these principles (italics added): "constraints differ in their *evidential* force", "all *evidence* is considered, none thrown out", "*evidence* has a smaller effect as we approach certainty", and so on. Notice that there is a repeated reference to *evidence* in the list of principles of common sense. I contend that these principles evoke a conceptual frame that we might call the evidence assessment frame. (Note that by "conceptual frame" I mean a frame in the sense of cognitive semantics (Fillmore 1976, 2006), i.e., a set of roles, events, and relations which form a complex and co-evoke each other). In the evidence assessment frame we have the following entities/events: something uncertain, an observer, things that are observed ("evidence"), an assessment process ("weighing" of the evidence), and the outcome of that process, which is often a choice or judgement.

What is crucial to consider here is that by describing MaxEnt as "common sense," and evoking the evidence assessment frame, Hayes invites the metaphor that cognitive processes are evidence-based decisions. In other words, the physical processes that occur in the nervous system are conceptualized as the actions of a human-like observer who weighs evidence. This is clearly a metaphor, because there is not actually a miniature imp who resides in our brains and makes decisions. And yet, the argument that MaxEnt is common sense subtly evokes this idea. This is problematic because anthropomorphization of cognitive processes may discourage us

from pursuing more realistic descriptions that conceptualize cognition as the dynamics of a highly organized, many-component complex system.

My third critique relates to a number of assumptions that are necessary for defining probability in the context of phonological behavior. It is important to recognize that we do not observe probabilities directly; rather, we observe frequencies of events. To associate probabilities with those events requires additional assumptions. In the Kolmogorov set-theoretic formulation of probability theory (Kolmogorov 2018), events or outcomes of a random process are considered members of a set. The events must be mutually exclusive. Probabilities are numeric values assigned to each member of the set, such that the sum of all probabilities is 1. For probabilities to be empirically useful in a given domain, the ratio of any two probabilities should correspond to the ratio of frequencies of events observed over a long time. In order for probabilities to be well defined, therefore, it is necessary that the outcomes/events are correctly defined.

Yet there are many ways in which the set of events posited in a MaxEnt analysis could be improperly defined. What we think of as a single event might actually be two distinct ones, or vice versa, what we think of as distinct events might be not be the same. Even worse, we could be unaware that certain events are possible. These are so-called *black swans*, a reference to the idea that Europeans presumed black swans did not exist (the probability of a swan being black was zero) prior to their discovery in Australia. For domain-specific examples, consider that it is hard to know how speech errors or even subtly abnormally articulated forms should be incorporated into the set of possible outcomes.

Even worse, there is a possibility that researchers are entirely mistaken about the nature of the events themselves, in a very fundamental way. Consider that, despite decades of effort, phoneticians have failed to identify invariant physical properties of "phones" or speech "segments". Segments and the featural symbols that define them are presupposed to be empirically identifiable in the vast majority of MaxEnt analyses, and yet there is no way to verify that they "exist" in a speech signal, nor is there any method for revealing a ground truth segmentation of speech. Ultimately, one must accept that the basic elements of mainstream phonological theories—phonemes, phones, features, syllables, etc.—are hypothetical entities, rather than physical things that we observe. Indeed, many influential researchers have questioned the reality of segments, describing them as a "figments of our good scientific imaginations" (Ladefoged 2001), "practical tools" (Browman and

Goldstein 1990), or artefacts of a cultural bias to favor discrete symbolic computation (Port and Leary 2005). See Tilsen (2016) for further discussion of the contestedness of segmental representations as well as an alternative conception of phonological organization based on mechanisms for selecting and coordinating articulatory gestures.

If the segmental conception of speech is wrong, then MaxEnt analyses are mischaracterizing the set of possible outcomes/events, and thus the probabilities that are obtained might bear little correspondence to what is actually represented in the brain. To be fair, this is primarily an issue with theories of phonological representation, not MaxEnt itself. One could very well adapt MaxEnt harmonic grammar to employ any set of outcomes, and so even if we as phonologists have not identified the correct set, perhaps the brain has done so. Nonetheless, it is somewhat suspicious to assert that, despite the difficulties phonologists have in identifying empirically testable outcomes, the brain has no such problem.

My fourth critique relates to a particular disanalogy between phonological cognition and relevant physical systems. Specifically, phonological cognition does not seem compatible with the temporal dynamics of closed systems in thermal equilibrium. Recall from above that in physical systems described by the Boltzmann distribution (e.g. macromolecules), the system potentially changes state from moment to moment, due to energy exchanges with its surroundings (e.g. collisions with other molecules). Crucially, the system is only in one particular microstate at a time, and due to our ignorance of the detailed dynamics (i.e. randomness), a probabilistic characterization of the system state is the best we can do. The problem with analogizing physical systems of this sort to phonological cognition is that it does not seem plausible that the analogue of microstates—word form representations—have this same property. That is, it does not make sense to assert that, at a given time, the phonological system is always in one particular state that corresponds to one word form representation.

To see more specifically why this disanalogy is problematic, consider once again the normalizing factor/partition function (Z) of the equations in Table 1, which is necessary for these equations to calculate properly normalized probabilities. The factor Z sums over logically possible microstates in the physical domain and over output representations in the phonological domain. In the physical domain, there is no pretense that the physical universe is actually calculating Z: there is no omniscient imp who iterates through possible states to obtain a value for Z. Rather, the

calculation of the partition function is something that physicists do in order to describe a system statistically and probabilistically. In contrast, since MaxEnt *is* generating probabilities, and representations of these probabilities are held to *exist* in the nervous system, then in the phonological domain, Z must actually be calculated by the nervous system.

The necessity of calculating Z in phonological cognition is problematic for the very reason mentioned above: it requires the phonological system to calculate the harmony of all possible states. How could this be accomplished? It seems implausible that the states are visited simultaneously, since they would likely interfere with one another. Certainly, classical physical systems cannot be in different states simultaneously, and I am reluctant to consider the notion that phonological cognition behaves like a quantum system, due to the physical scale of its substrate. Perhaps the phonological system visits each state in some logical sequence, in order to calculate each component of the normalizing factor—but what would determine the order, and how would the system know what all of the possibilities are? Perhaps, the calculation relies on ergodicity—the system randomly changes state from one moment to the next, updating the normalizing factor each time a new state is visited; but at some point it would need to recognize that all states have been visited and stop the calculation. Of course, these proposed solutions are quite speculative and not without problems. Yet, if the brain were truly calculating probabilities, it would be desirable to understand how this is accomplished. The point I am trying to make here is that the use of the Boltzmann distribution to calculate probabilities in phonological cognition leads to a number of inconsistencies, if we take the analogy seriously.

4. Conclusion

In this paper I have argued that the reification of probabilities as cognitively represented quantities invites a problematic conception of randomness and probability, in which randomness plays a shallow role in the generation of behavior. This shallow stochasticity is at odds with the role of randomness in physical systems that are the analogues of phonological cognition in the MaxEnt analogy. The states of physical systems are not governed by probabilities. Probabilities are merely statistical descriptions that are useful when we are mostly ignorant about the detailed dynamics of a system and its surroundings. Furthermore, I have offered a number of more specific critiques of MaxEnt: (i) it lacks an explicit mechanism for generating

behavioral variation from probabilities; (ii) the argument that it reflects "common sense" implies a problematic anthropomorphization of cognition; (iii) its implementation is based on possibly incorrect assumptions about the nature of phonological representations; and (iv) it leads to untenable views of the dynamics of the nervous system in order to calculate probabilities.

Perhaps a general counterargument to my critiques is that, in fact, I am taking the analogy too seriously, or that it is not a problem that phonological cognition differs in some ways but not others from the analogized physical systems. Clearly we do not expect phonological cognition to be identical to those physical systems in all ways. My reply is that we should not give ourselves license to borrow an equation from some physical domain and apply it arbitrarily to another. The use of Boltzmann distribution in MaxEnt begs the question of why we would expect phonological word form probabilities to exhibit the same relations as physical system-state probabilities, if not due to similarities between the two domains.

Along those lines, consider that MaxEnt adopts the Boltzmann distribution in a way that implicitly assumes that the product of the Boltzmann constant k and the temperature T is 1; thus temperature is a hidden parameter of the MaxEnt model—one that must necessarily NOT be hidden if we want to describe a physical system accurately. The reason this matters is that T IS a quantity that very well could vary over the course of preparing an utterance. That is indeed what was proposed in Harmony theory (Smolensky 1986) and cooling dynamics play a key role in the modern descendant of Harmony theory, Gradient Symbolic Computation (Smolensky et al. 2014, Smolensky and Goldrick 2016). In gradient symbolic computation there is both a dynamic temperature parameter and a dynamic parameter that quantizes states of continuous activation. It is beyond our scope to examine this framework but I will note that unlike MaxEnt, it does not necessarily entail that the cognitive system generates probabilities.

If not with reified probabilities, then how should we explain variation in phonological behavior? Although the point of this paper has been to present a series of critiques of reified probability, here I briefly sketch the outlines of an alternative. My view is that there are dynamical, deterministic processes that lead to a speaker to particular states that govern behavior. For concreteness, consider a nonce-word acceptability task of the sort conducted by Zuraw (2000), in which unfamiliar roots are presented first in bare form and then in either the assimilated or substituted form. In order to choose a number from 1 to 10 that reflects "acceptability", speakers

presumably arrive at some intuition about how closely the form associated with the stimulus matches similar forms associated with phonologically similar words that they have memories of. This sort of "matching" process is the sort of mechanism that is described in exemplar theory (Johnson, 2006, 2007), where new stimuli activate memories of previously heard stimuli (exemplars) in a way that depends upon their similarity. This sort of mechanism could be adapted to predict acceptability judgements by incorporating a function that maps from the total activation of exemplars to values on a 1-10 scale. Alternatively, the relevant input to this function may be activation of phonological and conceptual categories whose representations are evoked more or less strongly by the novel stimuli. In either case, randomness can be incorporated in a deep way by modeling our ignorance regarding the forces that govern exemplar or category activations. These sorts of explanations are not without their own problems, but my point here is that there are ways to explain behavioral variation that do not resort to reified probabilities.

Ultimately, the goal of this paper is to elaborate why we should disprefer models of phonological cognition that explicitly resort to cognitive representations of probability. I hope that my arguments will inspire others to pursue alternative models that are more deeply stochastic.

REFERENCES

BROWMAN, C. and GOLDSTEIN, L. 1990. Representation and reality: Physical systems and phonological structure. *Journal of Phonetics* 18.3, 411-424.

COHN, A. C. and RENWICK, M. E. 2021. Embracing multidimensionality in phonological analysis. *The Linguistic Review* 38.1, 101-139.

DE GUZMAN, V. P. 1978. A case for nonphonological constraints on nasal substitution. *Oceanic Linguistics* 87-106.

FILLMORE, C. J. 1976. Frame semantics and the nature of language. Annals of the *New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280.1, 20-32.

_____. 2006. Frame semantics. *Cognitive Linguistics: Basic Readings* 34, 373-400.

GOLDWATER, S., JOHNSON, M., SPENADER, J., ERIKSSON, A. and DAHL, Ö. 2003. Learning OT constraint rankings using a maximum entropy model.

Proceedings of the *Stockholm Workshop on Variation within Optimality Theory* 111, 120.

HAYES, B. 2021. Deriving the Wug-shaped curve: A criterion for assessing formal theories of linguistic variation. Retrieved from https://linguistics.ucla.edu/people/hayes/papers/HayesWugShapedCurve.pdf

_____. 2022. Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics* 8, 473-494.

_____. and WILSON, C. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3, 379-440.

JOHNSON, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34.4, 485-499.

_____. 2007. Decisions and mechanisms in exemplar-based phonology. *Experimental Approaches to Phonology*, 25-40.

KOLMOGOROV, A. 2018. *Foundations of the Theory of Probability: Second English Edition*. Courier Dover Publications

KURNIAWAN, F. O. 2018. *Phonological variation in Jakarta Indonesian: An Emerging Variety of Indonesian*. PhD Dissertation. Cornell University.

LADEFOGED, P. 2001. *Vowels and Consonants: An Introduction to the Sounds of the World*. Blackwell Publications.

PORT, R. F. and LEARY, A. P. 2005. Against formal phonology. *Language* 927-964.

SCOTT III, R. H. and SMITH, D. R. 2020. Pair-a-Dice Lost: Experiments in Dice Control. *UNLV Gaming Research & Review Journal* 24.1, 1.

SMOLENSKY, P. 1986. Information Processing in Dynamical Systems: Foundations of Harmony Theory.

_____ and GOLDRICK, M. 2016. Gradient symbolic representations in grammar: The case of French liaison. *Rutgers Optimality Archive* 1552, 1-37.

_____, GOLDRICK, M. and MATHIS, D. 2014. Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science* 38.6, 1102-1138.

_____ and LEGENDRE, G. 2006. *The Harmonic Mind: From Neural Computation to Optimality-theoretic Grammar (Cognitive architecture), Vol. 1*. MIT press.

TILSEN, S. 2016. Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics* 55, 53-77.

ZURAW, K. R. 2000. *Patterned Exceptions in Phonology*. PhD Dissertation. University of California, Los Angeles.

_____. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language & Linguistic Theory* 28, 417-472.

_____. and HAYES, B. 2017. Intersecting constraint families: An argument for Harmonic Grammar. *Language* 93.3, 497-548.

Sam Tilsen (Professor)
Department of Linguistics
Cornell University
Ithaca, NY 14850
United States
E-mail: tilsen@cornell.edu