Deriving the Wug-shaped curve: A criterion for assessing formal theories of linguistic variation*

Bruce Hayes
UCLA

July 2021

To appear in copy-edited form in *Annual Review of Linguistics*

Abstract

I assess a variety of constraint-based formal frameworks that can treat variable phenomena, such as well-formedness intuitions, outputs in free variation, and lexical frequency matching. The idea behind this assessment is that data in gradient linguistics fall into natural mathematical patterns, which I will call **quantitative signatures**. The key signatures treated here are the **sigmoid curve**, going from zero to one probability, and the "**wug shaped curve**," which combines two or more sigmoids. I argue that these signatures appear repeatedly in linguistics, adducing examples from phonology, syntax, semantics, sociolinguistics, phonetics, and language change. I suggest that the ability to generate these signatures is a trait that can help us choose between rival frameworks.

*I would like to thank Scott AnderBois, Adrian Brasoveanu, Joan Bresnan, Volya Kapatsinski, Shigeto Kawahara, Tony Kroch , Mark Liberman, Beatrice Santorini, Benjamin Storme, Richard Zimmermann, and the audience members at the 2020 Berkeley Linguistic Society and the UCLA Phonology Seminar for helpful input and comments on this project.

1. Introduction: some probabilistic phenomena in linguistics

This article addresses three linguistic phenomena in which we need to characterize variability and gradience in the analysis. First, we frequently need to model cases where **alternative surface forms** are generated, at varying probabilities, from the same underlying form. This is a research focus in sociolinguistics (§5.2), phonology (§5, §5.1.2) and syntax (§5.4). Second, speakers of a language can **frequency-match** statistical patterns in the lexicon. For instance, when Hungarian speakers undertake a nonce-probe task testing their intuitions about the principles of vowel sequencing, their responses statistically match the pattern of the Hungarian lexicon (Hayes et al. 2009); while in syntax, speakers statistically track the selectional properties of verbs and use this information in sentence perception (Jurafsky 2003, Linzen et al. 2016). Third, **native speaker judgments**, which include phonological well-formedness judgments (Scholes 1965, Hayes and Wilson 2008) and grammaticality judgments in syntax (Lau et al. 2017), are characteristically gradient and can be modeled probabilistically.[1]

To treat these cases in generative grammar, we need frameworks that can generate outputs on a probability scale. The key framework to be covered here will be **Maximum Entropy Harmonic Grammar** (Goldwater and Johnson 2003, Wilson 2006) — for short, "MaxEnt" — which is a probabilistic version of Optimality Theory (Prince and Smolensky 1993/2004). The apparatus in MaxEnt that assigns probabilities is identical to the statistical procedure of **logistic regression**, and I will alternatingly use "MaxEnt" and "logistic regression" below to refer to the same math, depending on context.

I will also evaluate MaxEnt against alternative approaches to constraint-based probabilistic linguistics. The strategy adopted uses simple math to locate the quantitative patterns characteristically generated under each theory, patterns which are visually identifiable when we plot them on a graph. I will call such patterns **quantitative signatures**. My work follows up on earlier studies of this kind (Jesney 2007, Zuraw and Hayes 2017, Hayes 2017, Smith and Pater 2020). I extend this research by offering a way to visualize the signatures that I believe is informative; and by applying the method to all areas of grammar.

I will address two related signatures. For each, I will describe the pattern, cite real-world cases, and demonstrate mathematically which frameworks possess these signatures; this in turn is taken to reflect on the empirical adequacy of these frameworks. In pursuing this inquiry I examined about 25 different cases in various fields. This brief paper cannot accommodate them all, yet rigor compels me to report them. To this end, I have created a web site, the *Gallery of Wug-Shaped Curves*, (linguistics.ucla.edu/people/hayes/ GalleryOfWugShapedCurves/). For each case, the site includes an illustrative graph as well as the spreadsheet calculations that generated it. The site also includes a longer version of the present article, covering matters omitted here for reasons of space.

[1] The three cases just given do not exhaust the set of gradient phenomena in linguistics; I omit the research program of modeling *physical* gradience in the phonetic output of the grammar, as in Liberman and Pierrehumbert (1984). Some recent examples that approach this problem using math similar to that discussed here are Flemming and Cho (2017) and Hayes and Schuh (2019).

2. MaxEnt

I begin with an exposition of MaxEnt. This will be more than an overview, because developing a close *intuitive* understanding of MaxEnt helps with the task of assessing quantitative signatures, hence theory-comparison.

2.1 MaxEnt as a species of Optimality Theory

In linguistics, MaxEnt is a version of Optimality Theory (OT; Prince and Smolensky 1993/2004). In OT, one analyzes a language system using a set of **inputs**, sets of candidate **outputs** for each input, and a set of **constraints** used to choose from among candidates. The theory derives outputs not with a serial derivation, but by defining in advance the set of all possible outputs (GEN), and employing a metric (EVAL) that selects the best one. The metric used for candidate selection is this: constraints are strictly ranked, as part of the language-specific grammar. Between any pair of candidates for a given input, the decision is made by the highest-ranking constraint that prefers (assigns fewer violations to) one of them. Similar decisions made across the whole candidate set determine a unique overall winner, which is the output of the grammar.

In probabilistic versions of OT, selection of a unique winner is replaced by assignment of probability to every member of GEN. In variable phenomena, often more than one candidate receives non-negligible probability, and these numbers serve as the predictions of the model, testable against corpus or experimental data.

2.2 The MaxEnt math and its intuitive rationale

MaxEnt replaces the strict-winner selection system of classical OT with the mathematics of logistic regression, with constraint violations taking the role of predictors. In this section, I take apart this math, step by step, showing that each step is intuitive and sensible. This will help later as we examine how the math behaves in language examples.

A goal of this discussion is to portray MaxEnt as a *mathematicized embodiment of common sense.* The key idea is to think of MaxEnt as a decision procedure. The constraint violations are, in essence, *evidence* bearing on which candidates should be assigned high or low probability. We start by looking at the whole formula, given in (1).[2]

(1) *The MaxEnt formula*

$$\Pr(x) = \frac{\exp(-\Sigma_i w_i \mathrm{f}_i(x))}{Z}, \text{ where } Z = \Sigma_j \exp(-\Sigma_i w_i \mathrm{f}_i(x_j))$$

The formula calculates **Pr(*x*)**, the probability of candidate *x* for some input. The formula includes everything needed to calculate this probability, including the set of output candidates, the constraints, a violation count for each constraint/candidate pair — and one other item, the set of

[2] The formula appears in, e.g., Goldwater and Johnson (2003, ex. (1)); or the logistic regression chapter of Jurafsky and Martin (2020).

constraint weights, discussed below. We will now reconstruct the formula in stages, starting from its smallest parts.

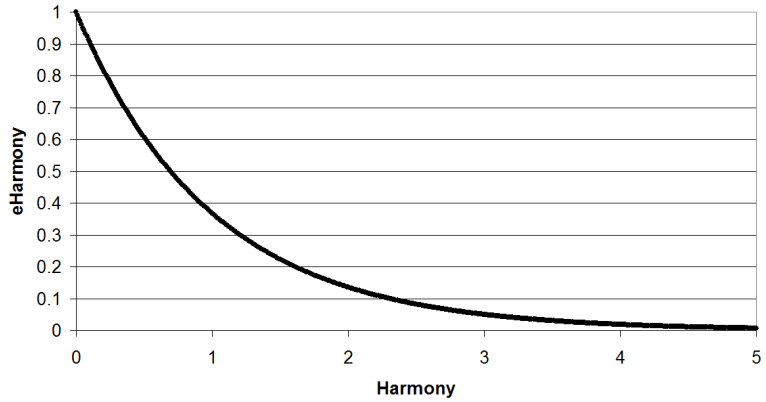
Constraint weights. In a MaxEnt grammar, the **weight** of a constraint is a nonnegative number that, intuitively, tells you how strong it is; or more specifically, how much it lowers the probability of candidates that violate it. In (1), this is $\boldsymbol{w_i}$ for each constraint i. Assigning weights to constraints is intuitive, because reasons differ in cogency.

Multiple violations. In (1), we see the expression $\boldsymbol{w_i f_i(x)}$, where x is the candidate being evaluated, $f_i(x)$ is the number of violations that candidate incurs for the ith constraint, and w_i is the weight of the ith constraint. Thus, weights are multiplied by violation counts. This is intuitive in the sense that two violations are plausibly "twice the evidence" of one.

Harmony. Once weights and violations have been multiplied, we calculate a sum across all constraints for each candidate. This sums acts as an penalty score for the candidate, and it is often called the **Harmony** (Smolensky 1986). In (1), Harmony is represented by $\boldsymbol{\Sigma_i w_i f_i(x)}$, where Σ_i represents summation across constraints. The use of summation is intuitive because when we make rational decisions, we find it appropriate to weigh *all* of the evidence. In this respect, classical OT is bravely counterintuitive, because the choice between two candidates is made solely by the highest ranked constraint that distinguishes them, ignoring the testimony of all lower-ranked constraints (Prince and Smolensky 1993:§5.2.3.2). The view taken here is that Prince and Smolensky's move to discard evidence was brave, but emerges in the end as empirically wrong.

eHarmony. The Harmony values are next converted to what Wilson (2014) has called **eHarmony**.[3] This is done by negating Harmony, then taking e (about 2.72) to the result. In formula (1), the term for eHarmony is: $\boldsymbol{\exp(-\Sigma_i w_i f_i(x))}$, where $\exp(x)$ is an abbreviation for e^x. The eHarmony function is plotted in (2) below.

(2) *eHarmony plotted against Harmony*



[3] Wilson was joking (eHarmony is a dating website), but the mnemonic seems useful.

The eHarmony function *rescales* the evidence: if Harmony is increased from an already-large value, then the eHarmony, being already close to zero, and gets only slightly smaller; whereas if Harmony is not very big in the first place then small differences in Harmony result in large differences of eHarmony.

I suggest that this rescaling reflects intuitively sensible decision making. For suppose we are trying to predict output probability for a candidate for which we know, as a rough guess, that the probability is going to be about .5. In such a case, we are quite uncertain, and additional information to inform our choice is welcome and taken seriously. If on the other hand, if a candidate is heavily penalized by information we already have (e.g. probability .001), than even a great deal of evidence may shift probability by only a small amount; say, to .0005. And for most people, I suspect, to become *absolutely* certain requires a vast, perhaps infinite, amount of evidence. A slogan that may be useful to remember is: *certainty is evidentially expensive*: to move probability around when it is already close to zero or one requires large infusions of evidence. The use of eHarmony implements this intuition mathematically.

Probability. There are two more steps: (1) We sum up eHarmony for all the candidates assigned to a given input, calling this sum Z. In formula (1), Z is expressed as: $\boldsymbol{\sum_j \exp(-\sum_i w_i f_i(x_j))}$, where *j* is the index intended to denote candidates. (2) We calculate the *probability* of a candidate by dividing its eHarmony by Z; i.e. we calculate its share in Z. This division appears in the complete formula for P(*x*) in (1). The addition-then-division procedure is intuitive, since it says that a candidate is *less likely if it has strong rivals*. Further, we see now that the probability of any candidate is proportional to its eHarmony; hence the discussion in the preceding section, showing how exponentiation makes certainty evidentially expensive, carries through to the final probability relations.

Summing up, the MaxEnt computation is claimed here to be intuitive at every stage:

(3) *MaxEnt and common sense*

a. Constraints differ in their evidential force
b. Multiple violations of the same constraint make a candidate less probable.
c. All evidence is considered, none thrown out.
d. Evidence has a smaller effect as we approach certainty.
e. Candidates are less probable when they compete with powerful rivals.

To the extent that these five properties reflect sensible principles for arriving at conclusions from evidence, MaxEnt (or any framework that has these properties) can be said to have an *a priori* claim on our attention.[4]

[4] Obviously, there is much more to say about MaxEnt/logistic regression from the technical point of view. For logistic regression as a statistical inference technique, with applicable methods of significance testing, see the textbooks by Johnson (2008) and Baayen (2008). On logistic regression in computer science, with the standard method of calculating the best weights to fit the data (and the proof of its convergence), see Jurafsky and Martin (2020). For MaxEnt specifically applied as a method of analysis in generative grammar, see Goldwater and Johnson (2003), Jäger (2007), and Hayes and Wilson (2008).

3. First quantitative signature: the sigmoid curve

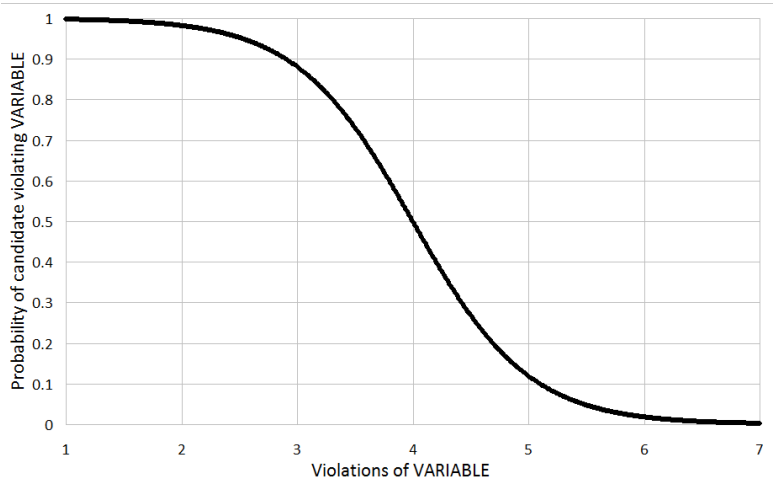
With this background we can turn to the main topic: quantitative signatures, their derivation under different theories, and their distribution in the real world. We focus on simple cases in which for each input, there are just two viable output candidates. In OT, including MaxEnt, this means that all other conceivable candidates are ruled out by powerful constraints. This is normal in OT, and I will not bother with formulating the necessary constraints below.

The two viable candidates compete on the basis of less-powerful constraints. Suppose that one of these constraints may be violated either *once* or *not at all*; call it ONOFF. Let the other be a constraint, or a set of constraints, defining a **scale**. Scales are familiar in constraint-based linguistics (Prince and Smolensky 1993/2004 §5.1; de Lacy 2004); and linguists have developed analyses in which the scale is formalized either with a single, multiply-violable constraint, or with families of related constraints.

Let us deal first with the simplest case, where the scale involves multiple violations of a single constraint. We call this constraint VARIABLE and assign it violation levels ranging (for concreteness) from 1 to 7. In the candidate competition, one of the two viable candidates for each input obeys VARIABLE and violates ONOFF, while the other obeys ONOFF and violates VARIABLE some specified number of times, depending on the input. Adopting this setup, we calculate the output probabilities using (1), and plot a function: the horizontal axis gives the number of violations of VARIABLE across inputs, and the vertical axis gives the probability that the candidate violating VARIABLE wins. For clarity, I will plot this function for *all* values on the horizontal axis, not just the 1-7 that would occur for particular input forms.

The curve that MaxEnt derives under these conditions is a **sigmoid** (S-shaped) function, illustrated in (4).

(4) *A sigmoid curve generated in MaxEnt*



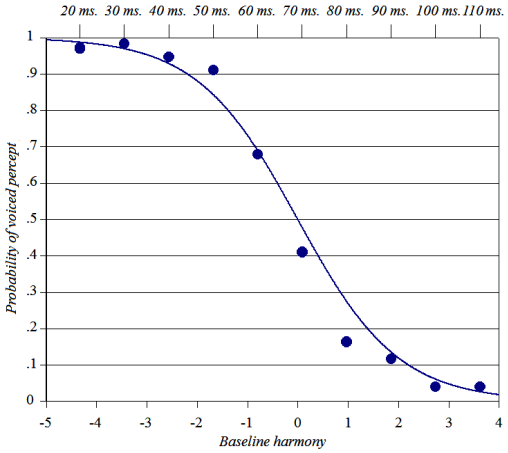
Here are crucial properties of the MaxEnt sigmoid, often called the **logistic** function. (a) It is symmetrical, and the symmetry point falls where probability crosses 50%. (b) It asymptotes on either end at 1 and 0. (c) It is steepest at the symmetry point, and becomes more level as one

proceeds in the positive or negative direction. (d) The uphill/downhill orientation depends on whether the constraint weight of VARIABLE is positive or negative; and its steepness is greater when the weight of VARIABLE is larger. (e) The relative right/left position of the curve is determined by the weight of ONOFF.[5] These properties must be kept in mind when we later assess whether an empirically-observed curve is properly to be considered as a sigmoid. Markedly asymmetrical curves, or curves that asymptote at a value other than one or zero, or curves that at some point reverse their slope, would not qualify. On the other hand, the language under study might not provide a full range of values for how many times VARIABLE is violated, so in the empirical domain we will often find truncated sigmoids.

3.1 Illustration: a sigmoid from a phonetic experiment

It is helpful to start with a case in which the horizontal axis of the sigmoid is uncontroversial, being a physical quantity rather than an analytic construct. Such cases arise frequently in phonetics, in the context of speech perception experiments. Suppose, for instance, that we plot on the horizontal axis a phonetic parameter like stop closure duration, varied in synthesized experimental stimuli. On the vertical axis we plot the probability that an experimental participant will experience a certain percept, such as [b] as opposed to [p]. Kluender et al. (1988) report such an experiment, and their data indeed emerged as an approximate sigmoid. A subset of the data is replotted in (5); the narrow line behind the data points represents the predictions of a MaxEnt model fitted to the data.

(5) *Sigmoid curve relating closure duration to voicing percept, adapted from Kluender et al. (1988)*



I assume the reader's agreement that the sigmoid curve superimposed on the data in (5) is a decent fit, and that small deviations may be attributed to sample or measurement error; the same holds for the remaining graphs in this article.[6] We turn, then, to the reanalysis of the data in MaxEnt terms.

[5] For discussion of these and other properties see McPherson and Hayes (2016).

[6] Of course, as we move from exploration (the goal here) to demonstration (the long-term goal), it becomes essential to assess model fit quantitatively. For standard techniques, see Johnson (2008) and Baayen (2008).

For present purposes it will be useful to adopt a stance proposed by Boersma (1998): that speech perception be regarded as a form of grammar. Boersma sets up a constraint-based, probabilistic theory in which the grammar inputs the acoustic signal and outputs a probability distribution for the set of possible phonemes (or words, etc.) that are inferred from the signal. The particular framework he uses to do this (not MaxEnt) is discussed below in §6.2.1.

Pursing Boersma's imperative in MaxEnt terms, we can arrange our grammar as a simple target-and-penalty system. The grammar inputs closure duration values and selects between the percepts [b] and [p]. As before, we exclude all other percepts by fiat; in a full grammar, they would violate highly-weighted constraints, resulting in essentially zero probability. Let the constraint VARIABLE penalize the percept of [b] to the extent that closure duration deviates from the extreme value of 20 ms. (which we adopt as the idealized target for [b]). VARIABLE assesses a penalty for every millisecond by which a [b] candidate exceeds this target. We also include a baseline ONOFF constraint, which simply penalizes all [p] candidates. VARIABLE and ONOFF conflict, and the computed [b]-probability will depend on the state of this conflict for a particular number of milliseconds of closure duration in the stimulus.

Using a spreadsheet, it is easy to find the weights that produce the most accurate model for the Kluender et al. data.[7] These turn out to be 0.088 for VARIABLE and 4.34 for ONOFF. From these, we can then use the MaxEnt formula (1) to calculate the probability of the voiceless candidate for all values; these are plotted as the narrow line in (5).

The units of the lower horizontal axis in (5), labeled "Baseline harmony," require comment. By simple math, applied to the MaxEnt formula, it turns out that in a system with just two viable candidates, we can recapitulate all the information needed to calculate their probability with a single number, the *difference of Harmony* between the two rival candidates. For how this works, see the long version of this article. The use of differences is helpful because we can encapsulate the relevant analytical information as a single value on the *x* axis.

Summing up so far: MaxEnt applied to the simple VARIABLE + ONOFF constraint system yields a sigmoid as its quantitative signature, and this signature emerges empirically in a speech perception experiment. We will return to this experiment, and similar ones, below.

4. Second quantitative signature: the wug-shaped curve

Assume as before an ONOFF constraint and a VARIABLE constraint, but this time let us double the input set, adding a new batch of inputs identical to the first except that they violate a constraint we will call the PERTURBER: a constraint defined on an independent dimension.

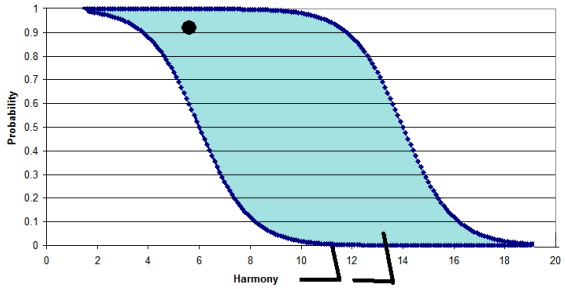
Let us first establish the MaxEnt predictions. The subpopulation of candidates that violate PERTURBER will have their Harmony values increased or decreased, depending on whether PERTURBER is "allied" with ONOFF or with VARIABLE. Other than that, these candidates will

[7] In brief, one locates the constraint weights that maximize the product of the predicted probabilities of all data points; i.e., one maximizes likelihood (Goldwater and Johnson 2003, (2)). In Excel, the Solver utility does well for this purpose on modest-size data sets. For the particular calculations done throughout this paper, see the spreadsheets posted in the Gallery.

behave just like their counterparts that do not violate the PERTURBER. Hence, if in a graph similar to (4), we plot the two populations of candidates separately, we will get a second sigmoid, *shifted over from the first* by an amount corresponding to the weight of the PERTURBER; see (6).

As Dustin Bowers suggested to me, it is not hard to imagine in this double-sigmoid shape the perky creature who in recent years has been adopted as the emblematic animal of linguistics; hence we can call it the **wug-shaped curve**, honoring its inventor, Berko (1958). I have artistically embellished (6) to emphasize the resemblance.

(6) *The wug-shaped curve*



The weight of the Perturber can be read off the graph; it is the horizontal distance between sigmoids; high and low Perturber values are thus represented graphically as fat and skinny wugs.

In some cases there will be more than one Perturber constraint. When this happens, we will obtain multiple parallel sigmoids, spaced as the weights dictate. It is tempting to think of this as a "stripey wug," but I will use "wug-shaped curve" for these cases as well.

5. Prospecting the linguistics literature for wug-shaped curves

My involvement with wug-shaped curves arose from participation as second author on Zuraw and Hayes (2017), a paper which adduces three wug-shaped curves in phonology and from them makes arguments about frameworks, some of them repeated below in §6. Subsequently, editor Mark Liberman helpfully suggested that I generalize these findings by addressing other fields of linguistics. Thus I embarked on a sort of intellectual hiking trip, browsing through classic works of probabilistic linguistics and replotting their data as arrangements of Baseline and Perturbers. To preview the outcome: this process repeatedly uncovered wug-shaped curves. I will give examples from several fields.

My criteria for choosing cases were as follows. First, the probability of candidates must approach one at one end, or zero at the other, or ideally both. Otherwise we only see vaguely parallel lines that are uninformative. Second, examples must be abundant enough so that each data point represents multiple observations, preventing random fluctuations from obscuring the pattern. Further, when setting up the analysis with Baseline and Perturber constraints, I favored a Baseline set that would yield a broad probability range. I also favored arrangements that gave the set of Perturber constraints (where possible, both sets) a unified, intuitively distinct rationale.

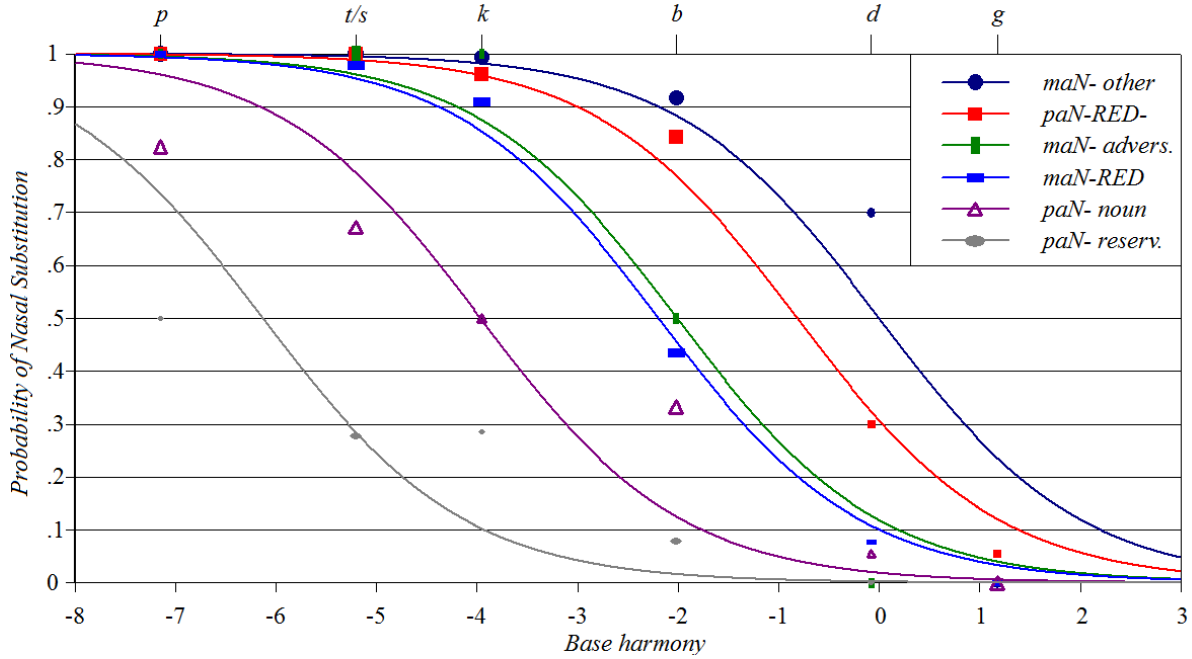
5.1 Phonology

5.1.1 Tagalog Nasal Substitution

Zuraw (2000, 2010), working on Tagalog, was the first phonologist to observe wug-shaped patterns and treat them in a probabilistic framework. In Tagalog, the sound [ŋ], when prefix-final, often *merges* with a following consonant, creating an output that blends the place of the consonant with the nasality of the [ŋ]; thus /ŋ+p/ → [m], /ŋ+t/ → [n], etc. The process is lexically optional, applying on a word-by-word basis, and the wug-shaped pattern of application rates emerged when Zuraw calculated these rates from a language-wide corpus, supported by a nonce-probe study.

In the presentation of these findings by Zuraw and Hayes (2017), a family of Baseline constraints forbids NC clusters with various features (place, voicing). This family, all of whose members receive different weights in the best-fit analysis, distinguishes six categories: {p, t/s, k, b, d, g}. These categories can be identified by the labels just above the graph in (7). Further, as Zuraw showed, each [ŋ]-final prefix of Tagalog has its own propensity to induce mutation; these differences are formalized with a family of prefix-specific Perturber constraints. The horizontal axis in (7) plots the baseline Harmony resulting from the consonant-specific constraints, and the Perturbers are represented by giving each its own sigmoid. Point sizes reflect the number of cases from which the probability is calculated. The plot is essentially the same as in Zuraw and Hayes (2017), except that the horizontal axis is scaled to reflect Baseline harmony.

(7) *The wug-shaped curve in Tagalog Nasal Substitution, after Zuraw and Hayes (2017:Fig. 10)*



The visual fit of the wug-shaped curve to the data strikes me as reasonably good; for quantitative testing of model fit, see Zuraw and Hayes (2017:§2.7).

An important aspect of the wug-shaped curve is that the magnitude of the effect of a Perturber depends on where we are located on the baseline scale: it is maximal in medial position and diminishes gradually toward the peripheries; see the vertical spacing of the dots in (7). This pattern, pointed in Zuraw and Hayes (2017) and Smith and Pater (2020), is (as can be calculated) a consequence of the MaxEnt formula. From the perspective of §2.2, the pattern is intuitive: the evidence from a Perturber buys you a lot in the middle, where you are uncertain, but will buy little at the peripheries, where you are already close to certain.

The remaining two cases discussed in Zuraw and Hayes (2017), French Liaison and Hungarian vowel harmony, when replotted using the format described here, again yield wug-shaped curves; these plots and the calculations supporting them may be viewed in the online Gallery.

5.1.2 Other work in phonology

In the Gallery, I give my replottings (with Baseline and Perturbers) of the following studies: Anttila's (1997) pioneering demonstration of constraint-based modeling of variable outputs, with data from Finnish genitive plurals; Ernestus and Baayen's (2003) modeling (including MaxEnt) of the ability of Dutch speakers to project the underlying forms of finally-devoiced consonants on the basis of the phonological properties of stems; Ryan's (2019) study of stress placement in Hupa; and Smith and Pater's (2020) study of vowel-zero alternations in French. All four cases yielded patterns reasonably interpreted as wug-shaped curves. Wug-shaped curves are also clearly found, and labeled as such, in Kawahara's studies (2020, in press) of sound symbolism in the names of Pokémon characters.

5.2 Sociolinguistics

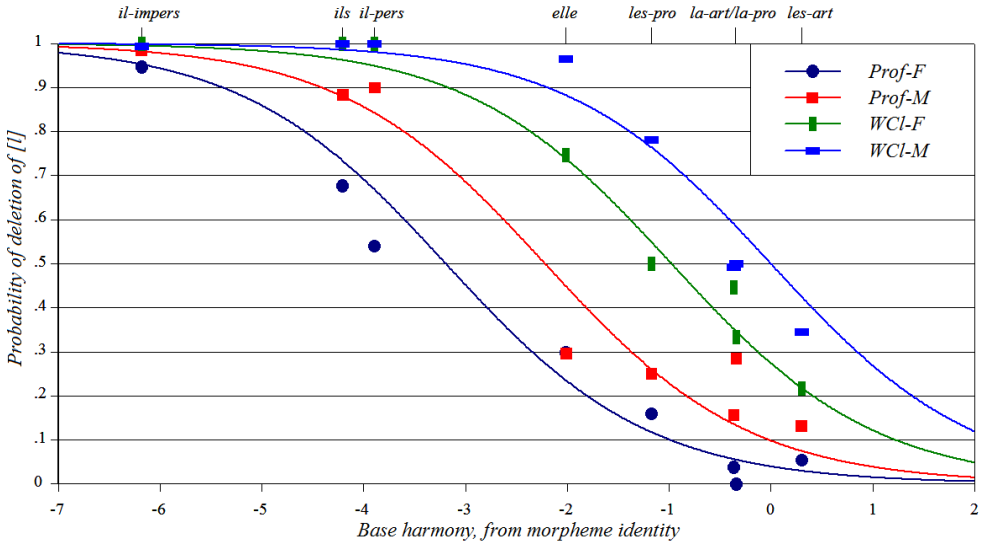
The essential theoretical concepts discussed above — MaxEnt analysis, Perturbers, and wug-shaped curves — all appear in research done by sociolinguists in the years around 1970. Labov's (1969) study of Black English copula deletion established the systematicity of linguistic variation; it also demonstrated the existence of Perturbers and their ability to affect output probabilities across the Baseline range. MaxEnt was later introduced (under the label of logistic regression) by researchers centered on D. Sankoff (Cedergren and Sankoff 1974, Rousseau and Sankoff 1978, Sankoff and Labov 1979). In this and later sociolinguistic work, the MaxEnt system was treated as a kind of triggering mechanism: each phonological rule has its own attached MaxEnt grammar telling it whether or not to apply.[8]

The illustration given here reanalyzes the data from which Bailey (1973) first deduced the presence of a wug-shaped curve. The data were taken by Bailey from G. Sankoff (1972), and involved optional deletion of [l] in function words in Québec French. In my MaxEnt reconstruction, the Baseline constraints are (1) a general Markedness constraint disfavoring the realization of [l], and (2) lexically-specific MAX constraints, militating against [l]-loss in

[8] For general background on quantitative modeling of variable phonology in sociolinguistics since the 1970s see §6.3 below, as well as Chambers and Schilling (2013) and Mendoza-Denton et al. (2003).

particular function words.[9] The Perturbers are — in superficial terms — further MAX constraints based on the sex and socioeconomic status (professional/working class) of the speaker. I doubt that such factors actually appear in the grammars of individual speakers; it seems more reasonable to suppose that speakers set the weight of MAX(l) differently in various social contexts, in ways that respond to sex and social class.[10] Thus in the present case, sex and social class are treated as proxies for the varying weight of MAX(l). The wug-shaped curve I obtained is in (8).

(8) *The wug-shaped curve in Quebec French [l]-deletion*



I also recalculated and plotted wug-shaped curves for several other classic studies, including those just mentioned: Labov (1969) (covering both contraction and deletion), Wolfram (1969) on Cluster Simplification in Detroit Black English, and three studies from Cedergren and Sankoff (1974): *que*-dropping in Québec French, [r] spirantization in Panamanian Spanish, and (with Labov's data) [r]-Dropping in New York City English. All of these may be found in the Gallery.

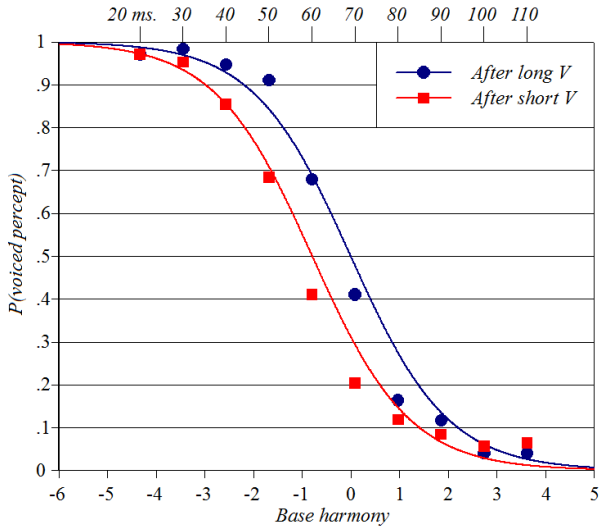
5.3 Phonetics

We return to the sigmoid from Kluender et al. (1988), discussed in §3.1. For simplicity, graph (5) plotted only one of the two data series from this paper. The authors' actual research interest was in a Perturber, the length of the vowel preceding the [b]/[p]. Their hypothesis was that, since vowels are normally longer before voiced stops, the presence of a longer vowel would bias perception in favor of [b]. That this hypothesis panned out is shown by (9) below.

[9] MAX, penalizing deletion, is a key constraint family in the standard theory of phonological Markedness constraints (McCarthy and Prince 1995). The tendency of function words to have morpheme-specific behavior has long been known; see Kaisse (1985).

[10] The response of phonology to social context is a vast research area, and the essays in Part III of Chambers et al. (2013) offer a useful guide. For an applicable MaxEnt proposal see Coetzee and Kawahara (2013).

(9) V*oicing percept by closure duration under two conditions (Kluender et al.1988)*



The MaxEnt grammar I set up for (9) is like the one for (5), except that it includes a Perturber, *VOICED PERCEPT AFTER SHORT VOWEL; this penalizes the [b] candidate when in this context. When I fit the full Kluender data, including both long and short vowels before [b] and [p], this constraint received a weight of 0.84, and the result was a clear if skinny wug.

Plots like (9) frequently appear in the work of phoneticians and psycholinguists, who use MaxEnt (under the logistic regression rubric) to quantify the influence of the Perturber.[11] Following the MaxEnt math, it is straightforward to rescale Perturber harmony as actual milliseconds. In the present case it emerges that the ms. value for *VOICED PERCEPT AFTER SHORT VOWEL is about 9.5 msec, in rough agreement with what Kluender et al. found using a different method.

5.4 Syntax

A number of studies in syntax have engaged with gradience of the types described in §1, using MaxEnt or similar models; see, for example, Velldal and Oepen (2005), Bresnan et al. (2007), Bresnan and Hay (2008), and Irvine and Dredze (2017). The particular research addressed here, by Bresnan and colleagues, focuses on a microdomain: instances in which the same communicative intent can be expressed with two different syntactic encodings. An example is the two ways that English offers to express the arguments of a verb of giving: NP NP (*Mary gave John a book*) and NP PP (*Mary gave a book to John*). In such cases, it has proven possible to identify probabilistic factors that favor one or the other outcome. In analyzing such cases, Bresnan et al. have used MaxEnt and similar tools. Their studies show that choices like *NP NP* vs. *NP PP* are, as it were, *semipredictable*, provided one uses a MaxEnt or similar model. The refined distinctions predicted by their constraint weights are supported empirically in that they show up as clear if modest distinctions between dialects, such as New Zealand and American

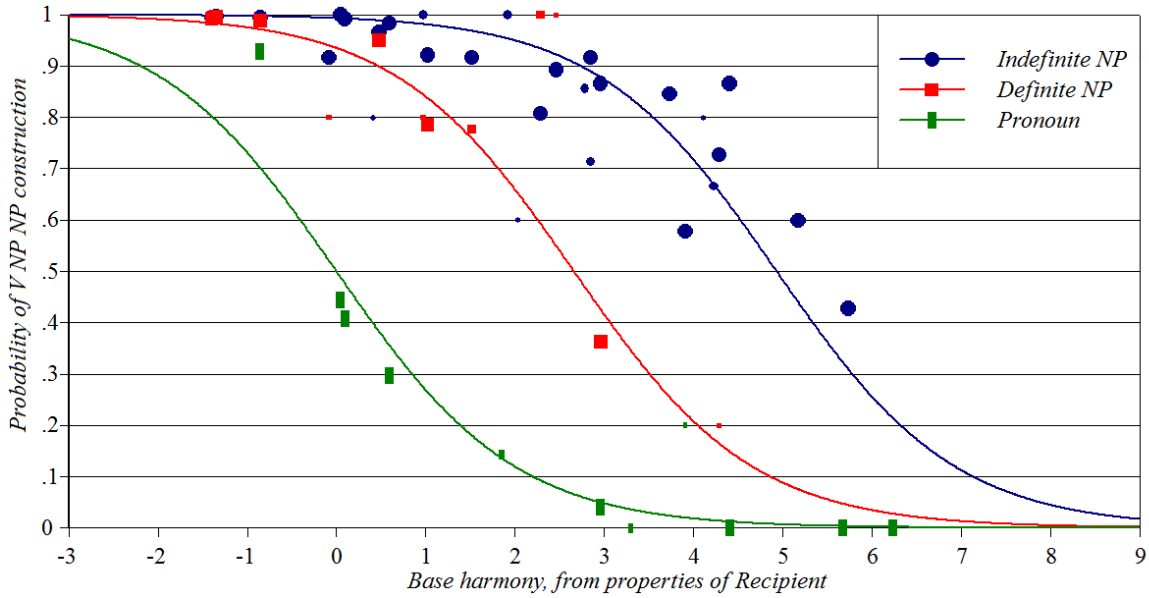
[11] Some classic papers employing this method include Ganong (1980) and Massaro and Cohen (1983); for helpful overviews see McMurray et al. (2003) and Morrison (2007).

English. These distinctions are attested both in experimentation (Bresnan and Ford 2010) and in corpus work.

Working in this tradition, Szmrecsanyi et al. (2017) uncovered dialect-specific patterns for four varieties of English (US, UK Canada, New Zealand) for two syntactic choices; the dative one just mentioned as well as the genitive choice for, e.g., *the king's palace* vs. *the palace of the king*. In my replottings, I abstracted away from these differences and merged the data from all four dialects.

For the datives, we can take as Baseline constraints the following: (1) those which depend on Szmrecsanyi et al.'s taxonomy of verb semantics, distinguishing "transfer," "communication," and "abstract"; (2) those dependent on properties of the *recipient* NP, such as animacy, definiteness, and pronounhood; (3) a constraint based on relative length (in words), which prefers placing longer phrases second. This array of constraints produces a rich baseline with multiple values (so, for details the reader should consult the original paper and the Gallery). For Perturbers, I selected the constraints and data series that single out three categories of the *theme* NP (that which is given): indefinite full NP, definite full NP, and pronoun. The wug-shaped curve that emerged under this re-plotting is shown in (10).

(10) *The wug-shaped curve in English dative constructions, after Szmrecsanyi et al. (2017)*



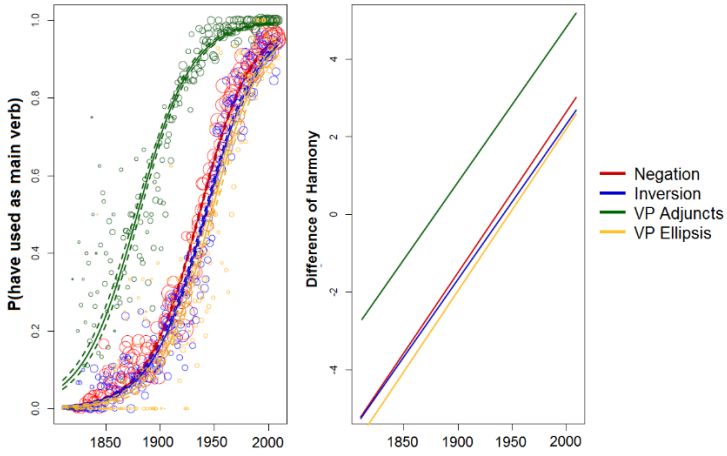
The Gallery contains replottings of the other findings in Szmrecsanyi et al. (2017). Their genitive data are of special interest because the numbers suffice to inspect the effect of one single gradient constraint, which favors the *N of NP* construction when the possessor NP is long, as measured in words. The replotting of these data demonstrates (at least in a limited range) the form of wug-shaped curve that arises (like (5)) from a system in which a single constraint with multiple violations forms the Baseline.

5.5 Historical linguistics[12]

The pioneer here is Kroch (1989), who inspected old texts across time, tracking the relative frequencies of competing syntactic variants as a language gradually changes. Employing the MaxEnt math, he made an important discovery: *when plotted as Harmony*, the rate of syntactic change is *constant* across the centuries. This constancy is obscured when the change is plotted as simple probability, where it appears as a sigmoid. Kroch also invoked Perturbers: additional constraints that interact with the change, and whose weight is constant over time. When Kroch's cases are sorted by Perturber and plotted in the manner given here, we get a wug-shaped curve.

A meticulous application of Kroch's ideas is Zimmermann (2017), addressing the evolution of English *have* from an auxiliary to a main verb. This change is manifested in four contexts: *negation* ("I (haven't/don't have) any"); *inversion* ("(Have you/Do you have) a penny?"), *ellipsis* ("You have a flair; you really (have/do)") and *adverb placement* ("He (has already/already has) the approval of the nation."). Each context is assumed to be affiliated with a Perturber constraint. The Variable is the diachronically-shifting constraint governing whether *have* functions as an Aux or main verb. Tracing each phenomenon across two centuries, Zimmermann obtains the wug-shaped curve in (11). The left panel depicts the four sigmoids that were found, with error bars and circles indicating the size of the text from which each datapoint derives.

(11) *A wug-shaped curve in syntactic change, adapted from Zimmermann (2017:107)*



The right panel implements a practice developed by Kroch: the four sigmoids are plotted not as observed proportions, but as the Harmony differences in the MaxEnt model. These lines are straight and parallel, illustrating clearly what is meant by the "constant-rate hypothesis."

Following up on Kroch, Blythe and Croft (2012:279-280) list dozens of language changes involving sigmoid curves. There is also an intriguing literature addressing *why* language change should typically show a constant rate; for discussion see the full version of this paper.

[12] For general background on probabilistic modeling of language change, including more detailed discussion of the material treated here, see Zuraw (2003).

5.6 Semantics/Pragmatics

Quantifier scope ambiguities occur in sentences like *A student saw every professor*. AnderBois et al's (2012) corpus study suggests that an effective system for predicting quantifier scope is most likely a probabilistic one: scope responds to a blend of conflicting factors. For full discussion and a wug-shaped curve from their data, see the long version of this article.

6. What formal models can generate Wug-shaped curves?

With the whirlwind tour of linguistics complete, we turn to the other goal of this article, framework assessment. This involves critiquing models that demonstrably fail to generate wug-shaped curves, and asking about models whose behavior is yet undiagnosed.

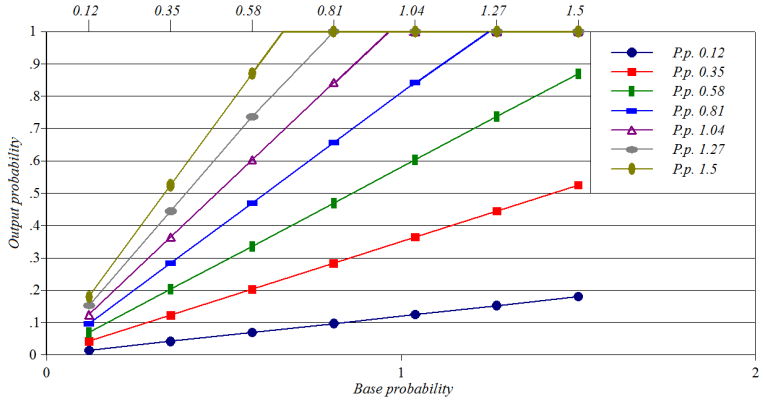
In inspecting the results of various frameworks applied to the same data, I have found a consistent pattern: often a defective framework gets lucky in fitting a particular batch of data. To evaluate a framework properly, we need to examine its performance in a variety of situations.

6.1 Some simple alternatives to MaxEnt

MaxEnt is not the only way to map from constraint violations and weights to probability, nor is it the simplest. The two alternatives discussed below were considered seriously in the early days of quantitative sociolinguistics, before the field shifted toward MaxEnt (Cedergren and Sankoff 1974, Sankoff and Labov 1979).

In a **Multiplication-cum-Cutoff** model, every constraint violation has the effect of multiplying candidate probability by the weight of the constraint. Constraints are allowed to bear weights greater than one, so they can increase as well as reduce probability. Since probabilities cannot go above one, this model prevents impossible values by imposing a ceiling of one by fiat. In (12) I give a schematic quantitative signature of this approach, assuming seven Baseline probabilities and seven Perturber probabilities (P.p.).

(12) *Quantitative signature of the Multiplication-cum-Cutoff model*



As can be seen, the prediction is that probabilities for particular Perturbers will converge in one direction, diverge in the other, up to the point where the cutoff prevents further divergence. I have never encountered data patterns like this and would be curious to know if they exist.

A second quantitative signature of the Multiplication-cum-Cutoff is obtained when it is applied to cases where a single constraint is violated a variable number of times What we find is a curious shape: a sigmoid with a sharp curve on one side but a gentle curve on the other (for graph, see long version of this article). I sense this has wrong typological implications — to my knowledge, nowhere in the speech perception literature is it claimed that the curves from identification experiments are systematically asymmetrical in this way; and the same would hold for the literature on historical-change sigmoids.

In an **Addition-cum-Cutoff** model, probability is linearly related to the violations of Variable, with each Perturber adding to, or subtracting from, the base probability by a constant. To avoid impossible probabilities, we impose cutoffs at 0 and 1. In this theory, the counterpart of the MaxEnt sigmoid is a "Z-shaped curve," and the wug is diamond-shaped, with parallel diagonals and horizontal lines at top and bottom (see long version of this article for sample graph). In actual model-fitting, this shape can often do fairly well, because the noise present in almost any data means that it is hard to distinguish smooth curves from sharp angles. However, the data from AnderBois et al. (2012) given above are poorly fit under this model; see full version of paper for illustration.[13]

The models just covered can be addressed more broadly, in terms of the ways that a constraint-based framework could express a rational inductive procedure. §2.2 showed that MaxEnt varies in how strongly evidence (here, constraint violations) bears on probability: in the middle of the probability range, violations are influential; at either periphery, less so; and this embodies the sensible principle that certainty should be evidentially expensive. Neither Multiplication-cum-Cutoff nor Addition-cum-Cutoff does this. Multiplication-cum-Cutoff says that the value of evidence is strongly asymmetrical with respect to the defining scale. Addition-cum-Cutoff says evidence is equally informative throughout the zone between cutoffs, then suddenly becomes 100% uninformative.

6.2 Frameworks originating in Optimality Theory

Two further approaches I will discuss have an ancestry in Optimality Theory; both are like MaxEnt in attempting to render OT probabilistic.

6.2.1 Stochastic Optimality Theory

In Stochastic OT (Boersma 1998), the key idea is that the content of the grammar is itself probabilistic: constraints comes with a number ("ranking value") that expresses how highly ranked they are in general, but each time the grammar is deployed ("evaluation time"), these ranking values are adjusted by a small random noise factor. The adjusted values are then used to

[13] The quantitative signature of the Addition-cum-Cutoff model would also poorly model an important finding in speech perception: small *differences* in the physical signal become progressively more informative to the hearer as one approaches the category boundary. This is a natural consequence of the MaxEnt sigmoid, whose derivative, depicting sensitivity, is a mountain-like shape. It would not be expected under Addition-cum-Cutoff, whose derivative is an all-or-nothing block, predicting that small differences should be uniformly effective in the local zone, useless outside it. The curves obtained by McMurray et al. (2008), for instance, strongly support MaxEnt under this interpretation.

sort the constraints, and at this point the choice of winning candidate follows classical OT. Repeated application of this procedure will yield a probability distribution through sampling.

Applied in the cases discussed here, Stochastic OT exhibits two failings. First, it cannot treat cases of variation from single VARIABLE constraints that vary in their violation count. This is because the classical-OT decision procedure is indifferent to the "margin of victory," caring only about relative differences. Such indifference is problematic for dealing with speech perception (§3.1, §5.3), word-count effects in syntax (§5.4) or syllable counts in sound symbolism (Kawahara (2020, in press).[14] Second, in Stochastic OT a Perturber can only perturb "within its own zone"; that is, when its own ranking value is within shouting distance of the constraints that it interacts with. But in the cases we have seen, the effect of a Perturber is *across the board*: it interacts with constraints that are mutually very far apart on the ranking scale. This point is covered in detail in Zuraw and Hayes (2017; §2.6). Both failings are rooted in traits of the classical OT on which Stochastic OT is based: contrary to principles (3b,c) above, it ignores relevant data, either in the form of violation counts, or of dominated constraints.

6.2.2 Noisy Harmonic Grammar

The primary reference is Boersma and Pater (2016). Like MaxEnt, this is a species of Harmonic Grammar, and the procedure for assigning probabilities to candidates likewise involves the computation of Harmony for each candidate. Noisy Harmonic Grammar resembles Stochastic OT in that we suppose a series of evaluation times at which the grammar gets altered by random shifts, chosen from a Gaussian distribution. The framework comes in several varieties, which differ in just which part of the calculation gets tweaked: we can alter constraint weights, violations, tableau cells, or the Harmony scores of candidates. Space does not permit detailed analysis of these varieties here, but see Hayes (2017), Anttila and Magri (2018), Anttila et al. (2019) and Kaplan (in press). Details aside, the framework is fully capable of generating wug-shaped curves, and as such is a plausible competitor to MaxEnt.

6.3 Other models

Among the probabilistic descendants of Optimality Theory, MaxEnt differs in origin from Stochastic OT and Noisy Harmonic Grammar in that it was not home-grown; it imported its math from existing work in statistics. However, from the viewpoint of statistics itself, MaxEnt is a bit retrograde, representing the avant garde of the 1970's (Cramer 2002). In more recent decades, it has become normal for experimental and corpus work that uses logistic regression to employ the *mixed-effects* version of the model (Baayen 2008, Johnson 2011), which controls for the idiosyncrasies of individual words or participants.[15] There are other models more elaborate than MaxEnt, such as neural network models (Goldberg 2017), or random forest models

[14] A proposal made by Boersma (1998: §6, §8.4) actually *can* derive sigmoids from variable constraints in Stochastic OT. The idea is to replace single gradient constraints with *bundles* of constraints, each having equivalent effect but a slightly different target value. The complexity of implementing this approach has perhaps been a factor in its still being underexplored.

[15] Indeed, the balancing of lexical vs. general preferences is a current live issue in phonological theory, and we are seeing efforts to set this balance appropriately (Moore-Cantwell and Pater 2016), including by using mixed-effects regression (Zymet 2018).

(Tagliamonte and Baayen 2012). Some of the authors whose empirical work is surveyed here have made use of these more sophisticated statistical approaches; e.g. Zimmermann (2017) employs mixed-effects regression and Szmrecsanyi et al. (2017) employ random forests. For the evolution of sociolinguistic modeling beyond simple logistic regression see Johnson (2009).

All of these developments are welcome, since there is every reason to think that more statistically sophisticated models will continue to be incorporated into linguistic theorizing, to the benefit of linguistic theory. However, unlike MaxEnt, these approaches do not (to my knowledge) have a simple analytic solution for when they would generate wug-shaped curves, and I would not venture to say anything here about their behavior in this connection.

7. Conclusions

7.1 Is the pattern found here meaningful, and if so, how?

To put a brave slant on the content of this paper: I raise the possibility that there exist general quantitative principles, along the lines of MaxEnt, that establish the normal patterning of variation in human languages, and that this is what leads to the repeated appearance of wug-shaped curves when we plot data from the various fields of linguistics. Of course, it is unlikely that with further scrutiny, *all* observed patterns of variation will line up as prettily as the ones seen here, and indeed I have found a few cases (all posted in the Gallery) where the presence of a wug-shaped curve is less compelling than described here. However, there is one fact that encourages me in thinking that a broader inquiry would confirm the basic pattern, namely that *logistic regression has proven popular wherever it has been adopted in linguistics*. This suggests that, on the whole, it has made possible accurate modeling of variation data; which, given the math discussed above, means that if we partition the constraints into Baseline and Perturber families, we will probably find more wug-shaped curves. That this is a non-trivial finding emerged from §6.1 and §6.2, which explored alternative approaches: MaxEnt and similar theories generate wug-shaped curves, others don't.

7.2 Is MaxEnt part of the language faculty?

When we ponder whether some principle pervasive in language is part of the "language faculty," there are two senses in which this is meant. One is, "an innate principle specific to human language"; the others "an innate cognitive capacity possessed by humans, employed in language." I suggest that if MaxEnt is part of the language faculty, it is probably in the latter, broader sense.

The crucial point is that MaxEnt is broadly used, under other names, elsewhere in cognitive science. The MaxEnt sigmoid and other curves that approximate it have been common currency in cognitive science for a very long time, often under the label "psychometric function" (Fechner 1860, Treutwein and Strasburger 1999). Multiple sigmoids (i.e., the wug-shaped curve) are likewise used by other cognitive scientists (for instance, applications to vision appear in Beaudot 1996 and Battista et al. 2011). One seminal work in modern cognitive science, Smolensky's (1986)'s proposal to use Harmony Theory in connectionism, included all of the MaxEnt math — without any intent to apply it specifically to language. Thus, I suspect many well-informed cognitive scientists would regard it as odd to consider the MaxEnt math as specific to language.

My own view (which others share) is that this is nothing that should trouble us; it is entirely sensible to seek general cognitive principles that illuminate the structure of language, and the MaxEnt principles may be among them.

References

AnderBois, Scott, Adrian Brasoveanu, and Robert Henderson (2012) The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*, vol. 16, no. 1, pp. 15-28.

Anttila, Arto (1997) Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory*, ed. Frans Hinskens, Roeland van Hout, and Leo Wetzels. Amsterdam: John Benjamins.

Anttila, Arto, and Giorgio Magri (2018) Does MaxEnt overgenerate? Implicational universals in maximum entropy grammar. In *Proceedings of the Annual Meetings on Phonology*, vol. 5.

Arto Anttila, Giorgio Magri, and Scott Borgeson (2019) Equiprobable mappings in weighted constraint grammars. *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*

Baayen, R. Harald (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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

Battista, Josephine, David R. Badcock, and Allison M. McKendrick (2011) Migraine increases centre-surround suppression for drifting visual stimuli. *PLoS ONE* 6(4): e18211.doi:10.1371/journal.pone.0018211.

Beaudot, William H. A. (1996) Adaptive spatiotemporal filtering by a neuromorphic model of the vertebrate retina. *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 1, pp. 427-430. IEEE.

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

Blythe, Richard A., and William Croft (2012) S-curves and the mechanisms of propagation in language change. *Language* 88:269-304.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy, eds. (2003) *Probabilistic Linguistics*. Cambridge: MIT Press.

Boersma, Paul (1998) *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. Ph.D. dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, Paul and Joe Pater (2016). Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy and Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen (2007) Predicting the dative alternation. *Cognitive foundations of interpretation*, ed. by G. Boume, I. Krämer, and J. Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

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

Bresnan, Joan, and Marilyn Ford (2010) Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86.168–213.

Cedergren, Henrietta J., and David Sankoff (1974) Variable rules: Performance as a statistical reflection of competence. *Language* 50:333-355.

Chambers, Jack K., Peter Trudgill, and Natalie Schilling-Estes, eds. (2013) *The handbook of language variation and change*. Oxford: Wiley-Blackwell.
Coetzee, Andries W., and Shigeto Kawahara (2013) Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 31:47-89.
Cramer, Jan S. (2002) The origins of logistic regression. *Tinbergen Institute Discussion Papers* No 02-119/4, Tinbergen Institute.
de Lacy, Paul (2004) Markedness conflation in Optimality Theory. *Phonology* 21:145–199.
Ernestus, Mirjam and R. Harald Baayen (2003) Predicting the unpredictable: interpreting neutralized segments in Dutch. *Language* 79:5–38.
Fechner, Gustav (1860, tr. 1966) *Elements of psychophysics*, tr. by H. E. Adler. Amsterdam: Bonset.
Flemming, Edward (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18:7-44.
Flemming, Edward and Hyesun Cho. 2017. The phonetic specification of contour tones: Evidence from the Mandarin rising tone. *Phonology* 34:1-40.
Ganong, Francis (1980) Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6:110-125.
Goldberg, Yoav (2017) *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan and Claypool.
Goldwater, Sharon and Mark Johnson (2003) Learning OT constraint rankings using a Maximum Entropy model. *Proceedings of the workshop on variation within optimality theory, Stockholm University*, 2003.
Hayes, Bruce (2017) Varieties of Noisy Harmonic Grammar. In Karen Jesney, Charlie O'Hara, Caitlin Smith and Rachel Walker (eds.), *Proceedings of AMP 2016*.
Hayes, Bruce and Russell Schuh (2019) Metrical structure and sung rhythm of the Hausa rajaz. *Language* 95:e253-e299.
Hayes, Bruce and Colin Wilson (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.
Hayes, Bruce, Kie Zuraw, Peter Siptar, and Zsuzsa Londe (2009) Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85: 822-863.
Irvine, Ann and Mark Dredze (2017) Harmonic Grammar, Optimality Theory, and syntax learnability: An empirical exploration of Czech word order. arXiv preprint arXiv:1702.05793.
Jäger, Gerhard (2007) Maximum entropy models and stochastic Optimality Theory. *Architectures, rules, and preferences. Variations on themes by Joan W. Bresnan*, ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 467-479. Stanford: CSLI Publications.
Jesney, Karen (2007) The locus of variation in weighted constraint grammars. Paper given at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford, CA.
Johnson, Keith (2011) *Quantitative methods in linguistics*. John Wiley & Sons.
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, Dan (2003) Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod et al. (2003), pp. 39-96.
Jurafsky, Dan and James H. Martin (2020) *Speech and Language Processing* (3rd ed. draft), web.stanford.edu/~jurafsky/slp3/

Kaisse, Ellen M. (1985) *Connected speech: The interaction of syntax and phonology*. San Diego: Academic Press.

Kaplan, Aaron (in press) Categorical and gradient ungrammaticality in optional processes. To appear in *Language.*

Kawahara, Shigeto (2020) A wug-shaped curve in sound symbolism: The case of Japanese Pokémon names. *Phonology* 37:383-418.

Kawahara, Shigeto (to appear) Testing MaxEnt with sound symbolism: A stripy wug-shaped curve in Japanese Pokémon names. To appear in *Language* (*Research Report*).

Kluender, Keith R., Randy L. Diehl, and Beverly A. Wright (1988) Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics* 16:153-169

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

Labov, William (1969) Contraction, deletion, and inherent variability of the English copula. *Language* 45:715-762.

Lau, Jey Han, Alexander Clark, and Shalom Lappin (2017) Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41: 1202-1241.

Liberman, Mark and Janet Pierrehumbert (1984) Intonational invariance under changes in pitch range and length. In Mark Aronoff and Richard T. Oehrle (eds.) *Language sound structure*. Cambridge, Mass.: MIT Press. 157-223.

Linzen, Tal and T. Florian Jaeger (2016) Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40:1382-1411.

McCarthy, John and Alan Prince (1995). Faithfulness and reduplicative identity. In Jill Beckman, Suzanne Urbanczyk and Laura W. Dickey (eds.) *University of Massachusetts occasional papers in linguistics 18: Papers in Optimality Theory*. 249–384.

McMurray, Bob, Michael K. Tanenhaus, Richard N. Aslin and Michael J. Spivey (2003) Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research* 32:77–97.

McMurray, Bob, Richard N. Aslin, Michael K. Tanenhaus, Michael J. Spivey, and Dana Subik (2008) Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance* 34:1609-1631.

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

Massaro, Dominic W., and Michael M. Cohen (1983) Phonological context in speech perception. *Perception and psychophysics* 34: 338-348.

Mendoza-Denton, Norma, Jennifer Hay, and Stephanie Jannedy (2003) Probabilistic sociolinguistics: Beyond the variable rule. In Bod et al. (2003), pp. 97-139.

Moore-Cantwell, Claire, and Joe Pater (2016) Gradient exceptionality in Maximum Entropy Grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15:53-66.

Morrison, Geoffrey S. (2007). Logistic regression modelling for first- and second- language perception data. In M. J. Solé, Pilar Prieto, Joan Mascaró (Eds.), *Segmental and prosodic issues in Romance phonology*, pp. 219–236. Amsterdam: John Benjamins.

Prince, Alan & Paul Smolensky (1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical report, Rutgers University Center for Cognitive Science. [Published 2004; Oxford: Blackwell]

Rousseau, Pascale and David Sankoff (1978) Advances in variable rule methodology. In David Sankoff, ed., *Linguistic variation: Models and methods*, pp. 57-69.

Ryan, Kevin (2019) *Prosodic Weight: Categories and Continua*. Oxford: Oxford University Press.

Sankoff, David, and William Labov (1979) On the uses of variable rules. *Language in Society* 8: 189-222.

Sankoff, Gillian S. (1972) A quantitative paradigm for studying communicative competence. Paper given at the Conference on the Ethnography of Speaking, Austin, Texas.

Scholes, Robert (1965) *Phonotactic Grammaticality.* The Hague: Mouton.

Smith, Brian W. and Joe Pater (2020) French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5:24.

Smolensky, Paul (1986) Information processing in dynamical systems: Foundations of Harmony Theory. In James L. McClelland, David E. Rumelhart and the PDP Research Group. *Parallel distributed processing.* Cambridge: MIT Press. 390-431.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte, and Simon Todd (2017) Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa* 2: 86.1-27.

Tagliamonte, Sali A. and Baayen, R. Harald (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 34:135-178.

Treutwein, Bernhard and Hans Strasburger (1999) Fitting the psychometric function. *Perception and psychophysics* 61:87-106.

Velldal, Erik & Oepen, Stephan (2005) Maximum entropy models for realization ranking. *Proceedings of the 10th Machine Translation Summit*, ed. by Jun-ichi Tsujii. Asia-Pacific Association for Machine Translation.

Wilson, Colin (2006) Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30.945–982.

Wilson, Colin (2014) Tutorial on Maximum Entropy models. Lecture given at the Annual Meeting on Phonology, Massachusetts Institute of Technology, Cambridge, MA, September 19.

Wolfram, Walt, and Ralph W. Fasold (1974) The study of social dialects in American English. Englewood Cliffs, NJ: Prentice Hall.

Zimmermann, Richard (2017) Formal and quantitative approaches to the study of syntactic change: Three case studies from the history of English. Ph.D. dissertation, University of Geneva.

Zuraw, Kie (2000) Patterned exceptions in phonology. Ph.D. dissertation, UCLA.

Zuraw, Kie (2003) Probability in language change. In Bod et al. (2003), pp. 139-176.

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

Zuraw, Kie and Bruce Hayes (2017) Intersecting constraint families: an argument for Harmonic Grammar. *Language* 93:497-548.

Zymet, Jesse (2018) *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning.* Ph.D. dissertation, UCLA.