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

Bruce Hayes
UCLA

July 2021

This is the longer version of a paper to be published
in shorter 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, the late Anthony Kroch, Mark Liberman, Beatrice Santorini, Benjamin Storme, Richard Zimmermann, and the audience members at the 2020 Berkeley Linguistic Society Workshop and the UCLA Phonology Seminar for helpful input and comments on this project.

## 1. Introduction: some probabilistic phenomena in linguistics

This article addresses linguistic phenomena in which we need to characterize variability and gradience in the analysis. There are at least three types:

- 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 their language. 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 (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 have examined about 25 different cases in various fields. This paper cannot accommodate them all,

[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 work that approaches this problem using math similar to that discussed here includes Flemming (2001), Flemming and Cho (2017), Lefkowitz (2017), and Hayes and Schuh (2019).

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.

## 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 adopted from the statistical method of logistic regression. The constraint violations take the role of predictors. In this section, I take apart the MaxEnt 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]

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

(1) *The MaxEnt formula*

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

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

### *2.2.1 Constraint weights*

In a MaxEnt grammar, the **weight** of a constraint is a nonnegative[3] real 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.

### *2.2.2 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 *i*th constraint, and $w_i$ is the weight of the *i*th constraint. Thus, weights are multiplied by violation counts. This is intuitive in the sense that two violations are plausibly "twice the evidence" of one.

### *2.2.3 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 $\boldsymbol{\sum_i w_i f_i(x)}$, where $\sum_i$ represents summation across constraints.

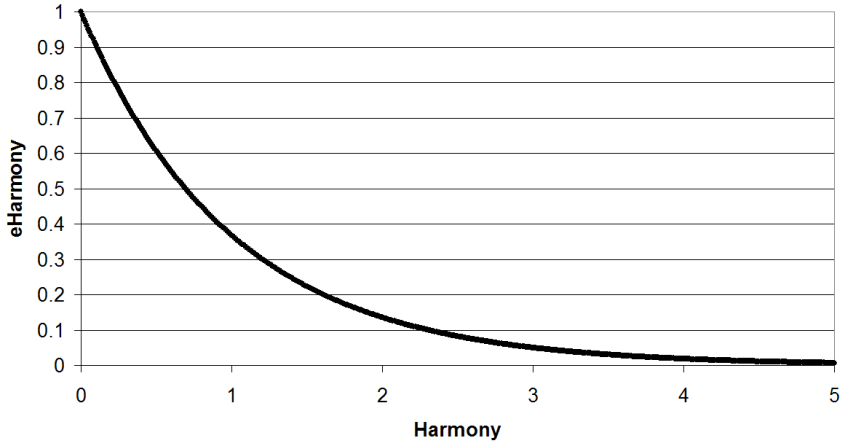
The use of summation is intuitive because when we make rational decisions, we find it appropriate to weigh *all* of the evidence. In this respect, classical OT is bravely counterintuitive, because the choice between two candidates is made solely by the highest ranked constraint that distinguishes them, 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.

[3] It is not unknown to use negative weights, in which case a constraint rewards rather than penalizing candidates; see Kaplan (2018) for a recent proposal. This method must be employed with care, since it risks assigning unlimited rewards to ever-longer candidates, wrecking the calculations.

### *2.2.4 eHarmony*

The Harmony values are next converted to what Wilson (2014) has called **eHarmony**.[4] This is done by negating Harmony,[5] then taking *e* (about 2.72) to the result. In formula (1), the term for eHarmony is: $\mathbf{exp(-\sum_i w_i f_i(x))}$, where exp(*x*) is simply a rewritten version of $e^x$. The eHarmony function is plotted in (2) below.

(2) *eHarmony plotted against Harmony*




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. The same reasoning applies when the probability of a candidate approaches 1, except that we are concerned instead with the evidence that penalizes its competitors.

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.

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

[5] The literature in MaxEnt varies concerning where this negation is carried out. In the usage adopted here, negation takes place as part of the calculations, but other papers use other methods, in particular opposite signs for the constraint weights, or opposite signs for the violations. All methods lead to the same result in the end.

### *2.2.5 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: $\mathbf{\Sigma_j \exp(-\Sigma_i w_i f_i(x_j))}$, 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.[6]

## 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.

[6] 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 (2011) 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).

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). The weights that generate this particular sigmoid are: $w_{\text{VARIABLE}} = 2$, $w_{\text{ONOFF}} = 8$.

(4) *A sigmoid curve generated in MaxEnt*


Here are some crucial properties of the MaxEnt sigmoid, often called the **logistic** function.[7] (a) It is symmetrical, and the symmetry point falls where probability crosses 50%. (b) It asymptotes on either end toward 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 proportional to the weight of ONOFF.[8] 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

[7] The logistic function was named in 1845 by its discoverer, the Belgian mathematician Pierre-François Verhulst. No one knows why he chose this name. For helpful discussion of the history of the function and of logistic regression, see Cramer (2002).

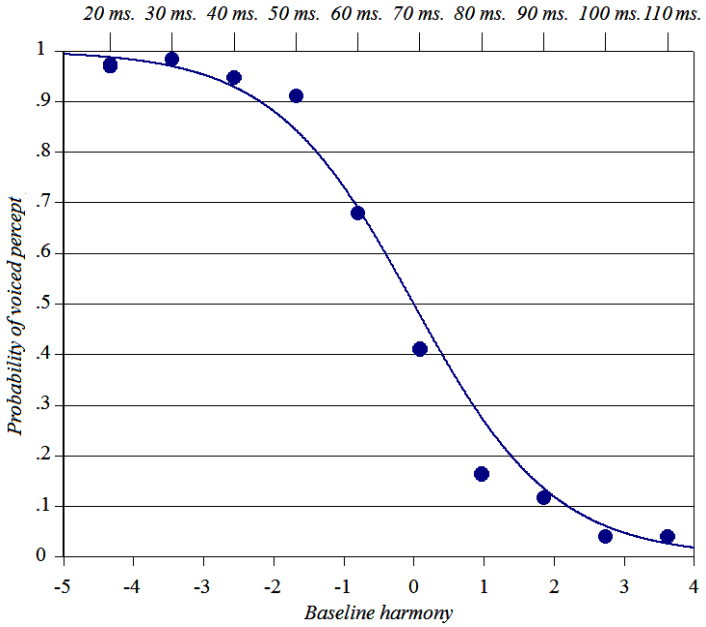
[8] More precisely, it is $w_{\text{ONOFF}}/w_{\text{VARIABLE}}$, hence in this case 4. Basic discussion of these and other properties of the MaxEnt sigmoid is given in McPherson and Hayes (2016).

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.[9] We turn, then, to the reanalysis of the data in MaxEnt terms.

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

[9] 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 (2011) and Baayen (2008).

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.[10] 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) above. Readers who wish to track these calculations may temporarily switch to reading Appendix A below.

On the lower axis of graph (5), the units plotted are **Baseline harmony**; these require comment since all the graphs below work in the same way. 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, which is what gets plotted on the *x* axis. For how this works, see Appendix B below.

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 cases below.

## 4. Second quantitative signature: the wug-shaped curve

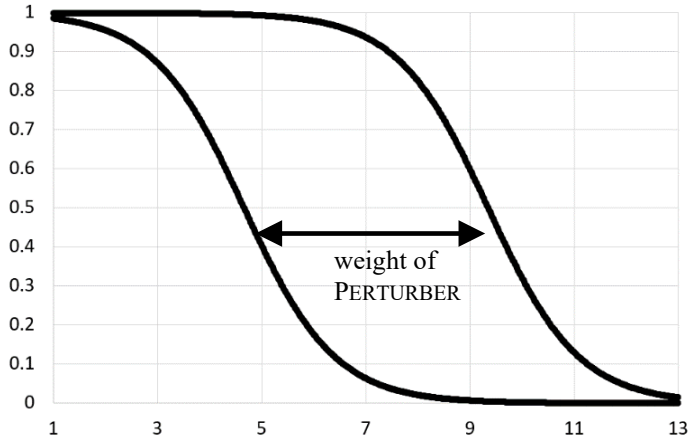
Assume as before an ONOFF constraint and a VARIABLE constraint, but this time let us double the input set, adding a new batch of inputs identical to the first except that they violate a constraint we will call the PERTURBER: a constraint defined on an independent dimension. In the analysis of cases in the real world, this situation often arises, since independent factors often bear simultaneously on the evaluation of candidates.

Let us first establish the MaxEnt predictions. The subpopulation of candidates that violate the 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

[10] 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 for this paper, see the spreadsheets posted in the Gallery.

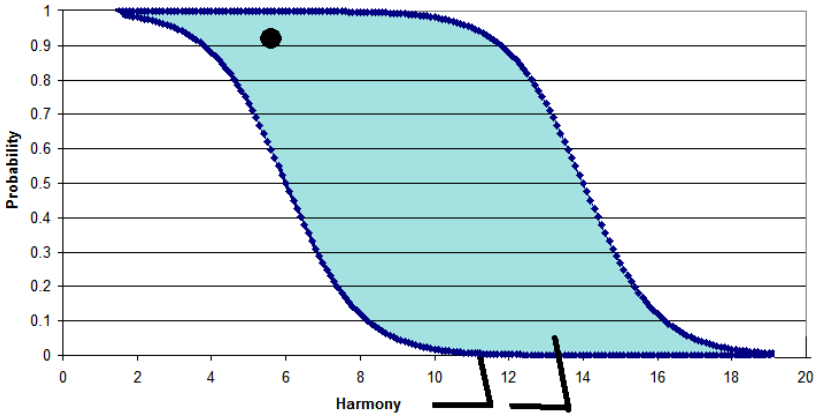
behave just like their counterparts that do not violate the PERTURBER. Hence, if in a graph similar to (4), we plot the two populations of candidates separately, we will get a second sigmoid, *shifted over from the first* by an amount corresponding to the weight of the PERTURBER, as in (7).

(6) *The double sigmoid resulting from the presence of a PERTURBER constraint*




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 (7) to emphasize the resemblance.

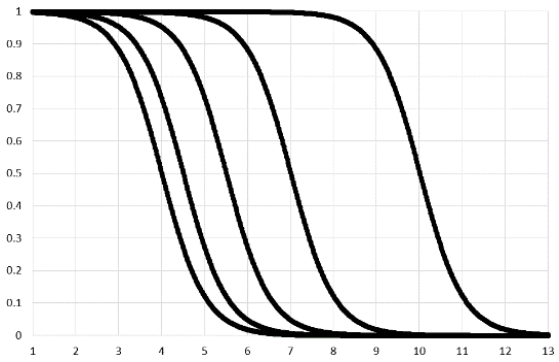
(7) *The wug-shaped curve*




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, as in (8):

(8) *The wug-shaped curve with multiple sigmoids*




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, Mark Liberman helpfully suggested that I generalize these findings by addressing other fields of linguistics. 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. and 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.

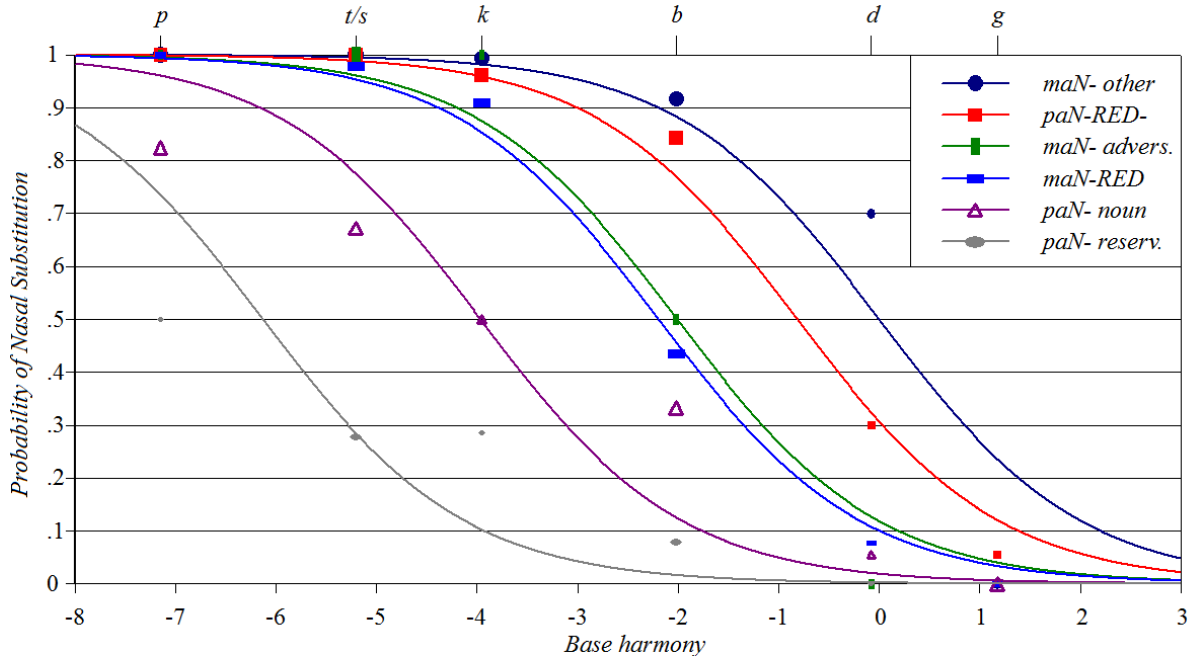
### *5.1 Phonology*

#### *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 (9). 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 (9) 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.

(9) *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 (9). 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; Smith and Pater's (2020) study of vowel-zero alternations in French; and Storme's (forthcoming) account of the same phenomenon from a different perspective. All of these 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 phonological 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.[11] 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.[12]

The illustration given here reanalyzes the data from which Bailey (1973) first deduced the presence of a wug-shaped curve.[13] 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 particular function words.[14] The Perturbers are — in superficial terms — further MAX

[11] 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).

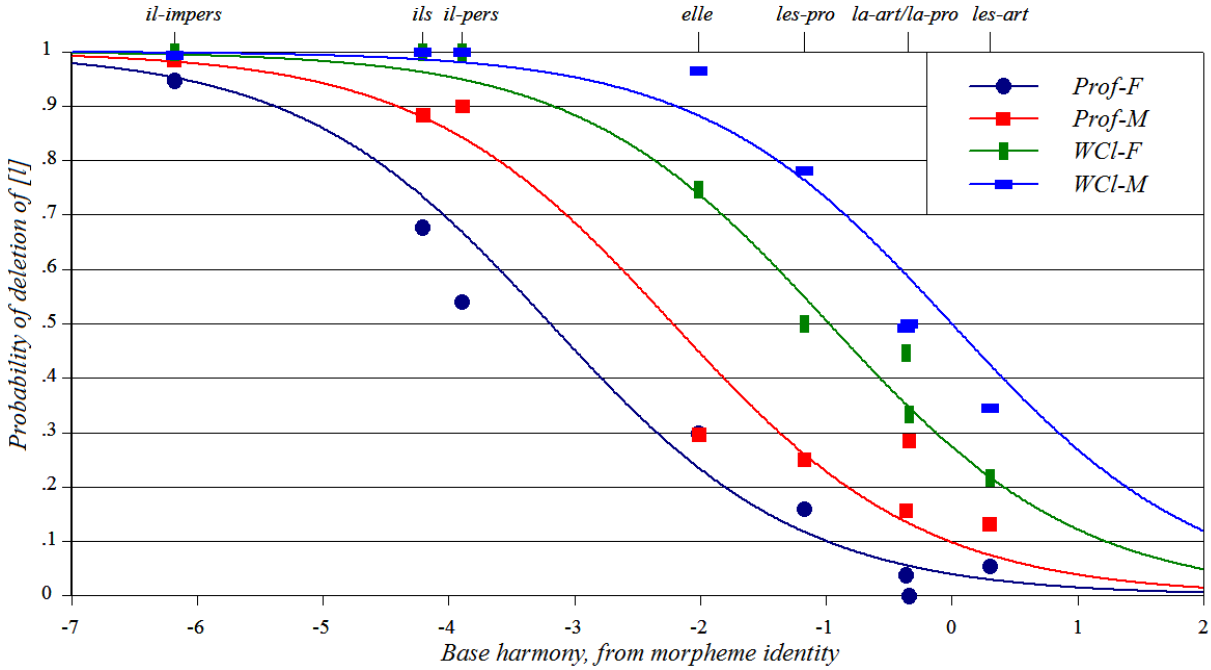
[12] In this theory, the bundle consisting of [*SPE*-style rule + MaxEnt controller] is called a **variable rule**. I feel that the variable rule system was a good idea in the 1970s, given the theoretical resources available at the time. In light of later developments, however, variable rules strike me as both extravagant (in MaxEnt, constraints alone suffice) and unconstrained (the same constraint can have different weights depending on which variable rule it resides in). The variable-rule system might be defended if cases can be adduced showing that each phonological process really does need a distinct MaxEnt grammar to guide it.

[13] Bailey's method of plotting the curve, using contour lines, strikes me as infelicitous, but his verbal description of cross-classifying Baseline and Perturber effects does capture the key point: "the statistics are more bunched in the bottom and top percentages and more spread out toward the middle percentages" (p. 106).

[14] 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).

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.[15] 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 (10).

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



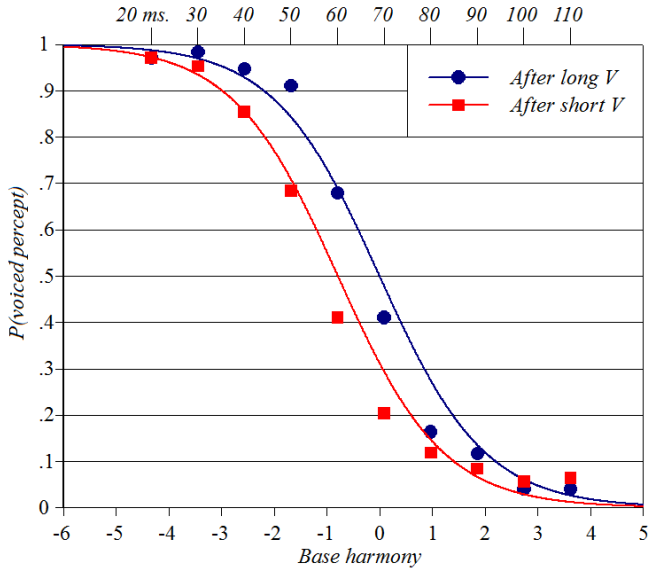

I also recalculated and plotted wug-shaped curves for several other classic sociolinguistic studies, including those just mentioned: Labov (1969) (covering both contraction and deletion), Wolfram (1969) on Cluster Simplification in Detroit Black English, and three studies from Cedergren and Sankoff (1974): *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, however, 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 (11) below.

[15] 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).

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




The MaxEnt grammar I set up for (11) 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 (11) frequently appear in the work of phoneticians and psycholinguists, who use MaxEnt (under the logistic regression rubric) to quantify the influence of the Perturber.[16] 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 comes from 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

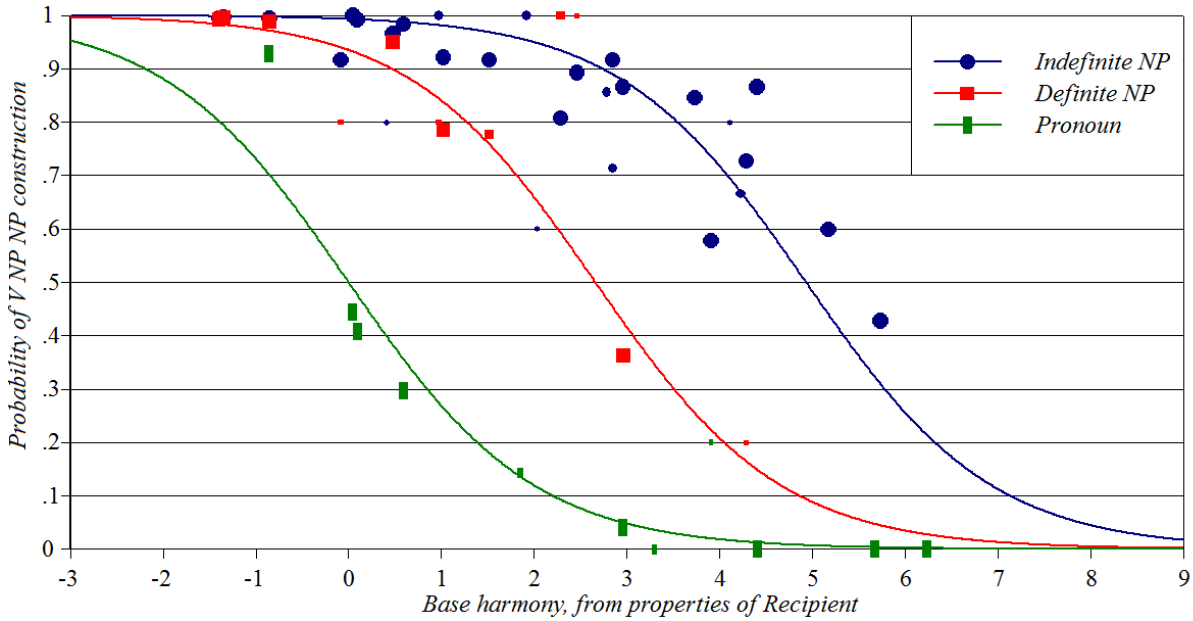
[16] Some classic papers employing this method include Ganong (1980) and Massaro and Cohen (1983, plotted in Gallery); for helpful overviews see McMurray et al. (2003) and Morrison (2007).

American 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 (U.S., U.K., Canada, New Zealand) for two syntactic choices; the dative one just mentioned as well as the genitive choice of, 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.

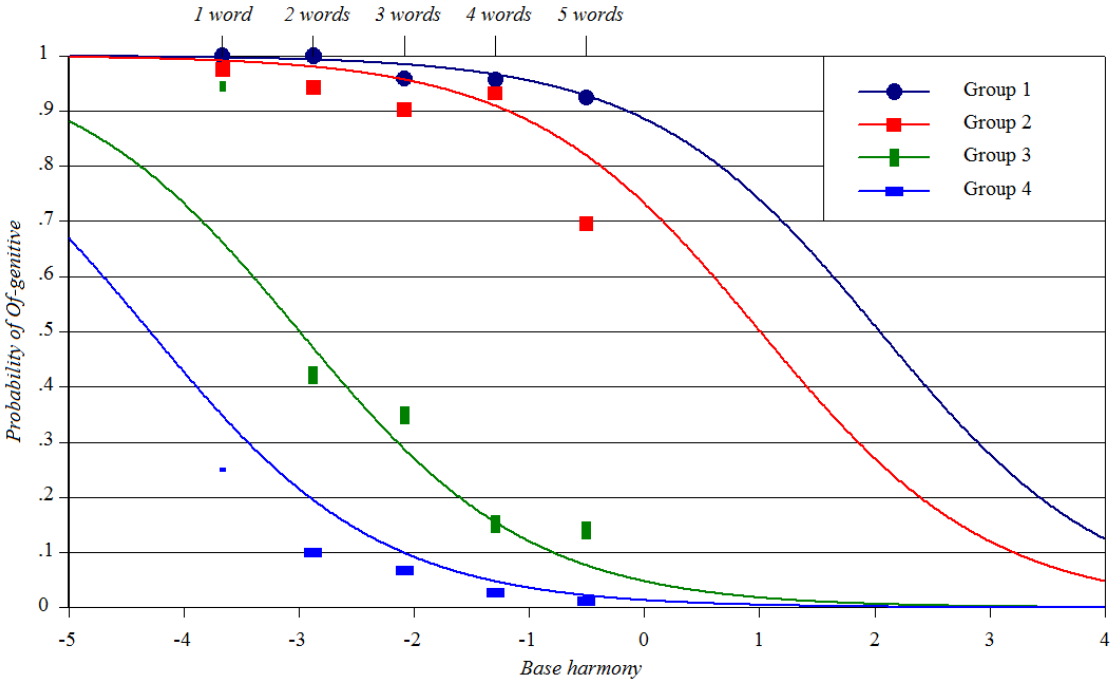
For the datives, we can take as Baseline constraints the following: (1) those which depend on Szmrecsanyi et al.'s taxonomy of verb semantics, distinguishing "transfer," "communication," and "abstract"; (2) those dependent on properties of the *recipient* NP, such as animacy, definiteness, and pronounhood; (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 (12).

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



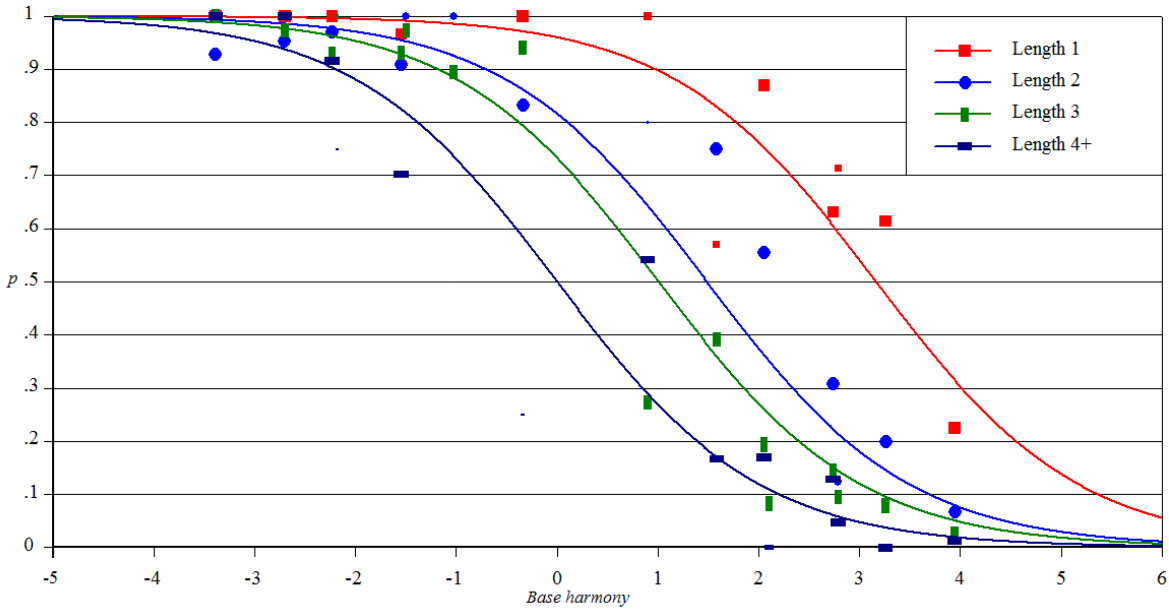

The Szmrecsanyi et al. genitive data are of special interest because the numbers suffice to inspect the effect of one single gradient constraint, which favors the *N of NP* construction when the possessor NP is long, as measured in words. The replotting of these data demonstrates (at least in a limited 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.

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




The key point is that, although the gradient length constraint lacks the force to induce a full-size sigmoid at any one level of Baseline harmony, it does seem that we are seeing "sigmoid snippets," each reflecting the gradient constraint as it shows its effects against a particular Baseline Harmony level. The same data yields a figure similar to the dative figure (12) if we choose to replot the data in the opposite way, with the various lengths treated as sources of Perturber harmony; see (14).

(14) *The wug-shaped curve in English genitive constructions, with alternative choice of Baseline*

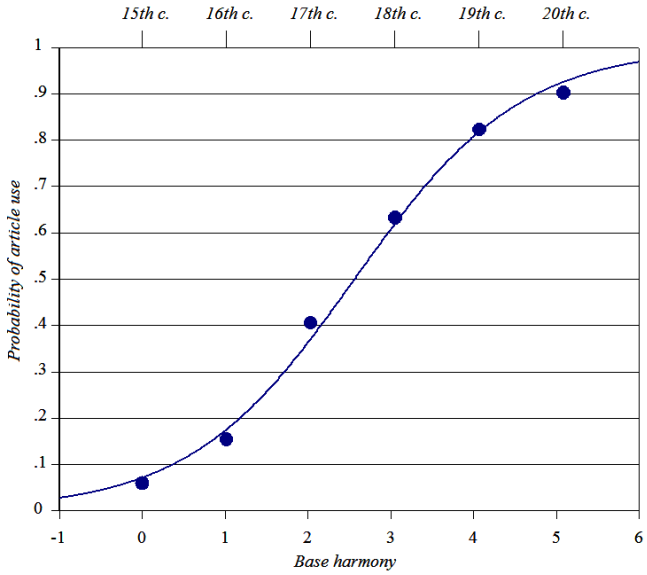



## *5.5 Historical linguistics*

The urtext for the application of MaxEnt/logistic regression to language change[17] is Kroch (1989), whose empirical method was to inspect old texts across time, tracking the relative frequencies of syntactic variants as a language gradually changes. For example, one of Kroch's data series, taken from Oliveira e Silva (1982), documents a centuries-long syntactic change in Portuguese: over time, the syntax of Portuguese noun phrases changed such that NP including a possessor would also include a definite article. Thus over time, e.g. *seus livros* 'his books' was gradually replaced by *os seus livros* '(the) his books'. In (15) below is a chart, adapted from Kroch, that shows the frequency with which the definite-article variant was employed.

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

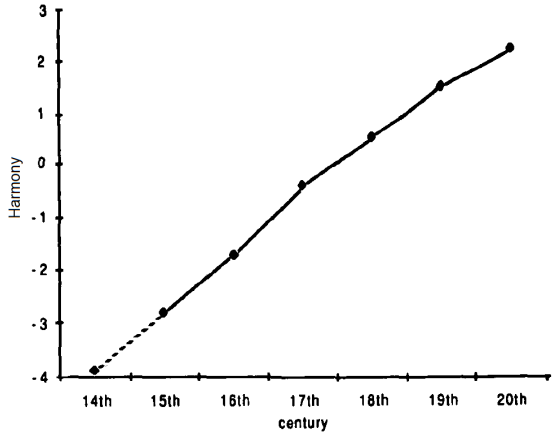
(15) *Use of the definite article in Portuguese possessed noun phrases, after Kroch (1989)*




What kind of mathematical curve is this? Kroch's suggestion, put in our terms, is that it is a MaxEnt sigmoid, with Base harmony in a linear relationship with time. In my own recalculation/replotting I make this explicit by including both time on the upper horizontal scale and the crucial harmony-difference values ($H_{article} - H_{noarticle}$) from my recalculations on the lower scale. The good fit of the MaxEnt sigmoid means that this historical change follows a clean generalization, namely that in the Harmony domain, the propensity to use a novel construction increases at a constant rate. For the Portuguese case, the rate of increase turns out to be about 1.02 Harmony units per century. Kroch's idea is now known as the "constant-rate hypothesis," and the pattern has been frequently observed since — Blythe and Croft (2012:279-280) list dozens of language changes involving similar sigmoid curves.

Kroch made his point graphically in a way that I have not yet done, namely by replotting the data on a log scale. This undoes the exp() step of MaxEnt; §2.2.4, yielding a quantity proportional to Harmony. In this format, the data plot more or less in a straight line. This is shown in (16) below, an edited version of Kroch's Fig. 1.

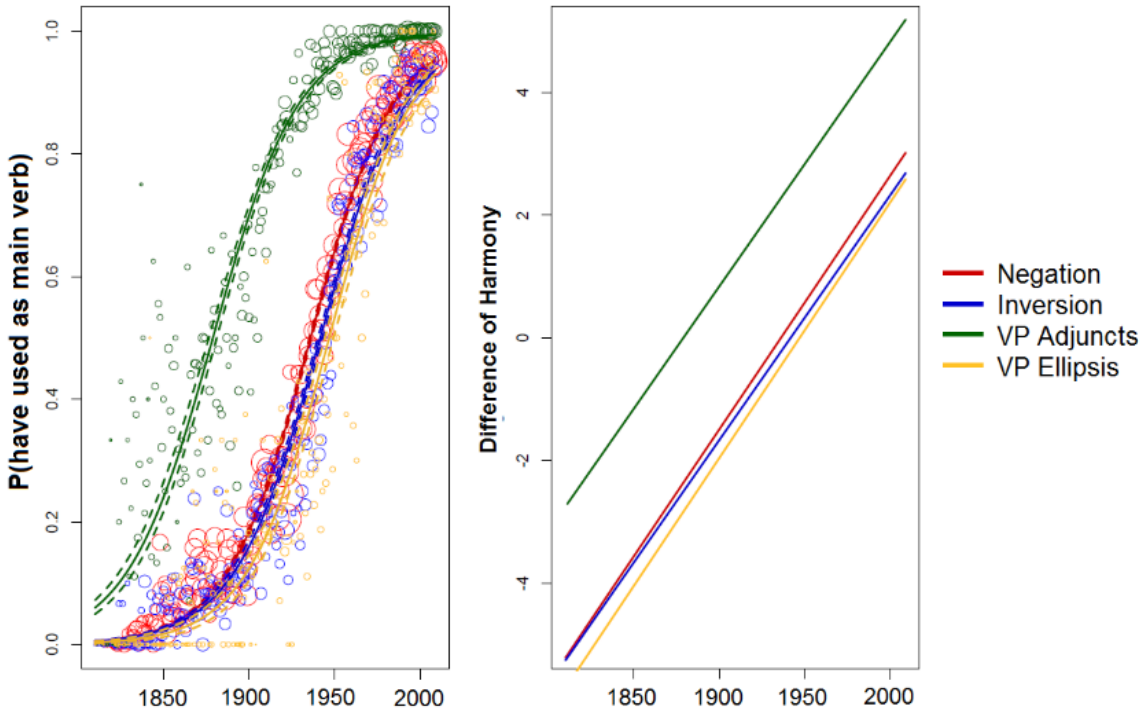
(16) *Use of the definite article in Portuguese possessed noun phrases, replotted as linearly-rising harmony (Kroch 1988)*




The constant-rate hypothesis becomes even more interesting when we include, as Kroch did, some Perturbers. Stated in present terms, the idea is that the Variable constraint is steadily increasing its weight over time, but the Perturbers are diachronically stable. Such conditions will give rise to a diachronic wug-shaped curve, with identical sigmoids spaced apart by an amount corresponding to the differences in the weights of the Perturbers. This is what Kroch and his co-workers have repeatedly found in their historical studies. The result is a pleasing one: the variation in rates of change across contexts may look nonsensical when measured directly as probability, but they are coherent and orderly — parallel lines, in fact — when measured in their natural linguistic units of Harmony.

An intensive illustration of Kroch's ideas is provided by Zimmermann (2017), who addresses 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 may be assumed to be affiliated with an appropriate 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 obtained the wug-shaped curve in (17). The left panel depicts the four sigmoids he found, with error bars and circles indicating the size of the text from which each datapoint derives.

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




The right panel implements the same practice as seen above in (16): 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."

In the present context, the constant-rate hypothesis leads to two important followups. First of all, it simply begs the question, "why should the rate be constant?". Some interesting progress has been made on this issue; for references and discussion, see Appendix C.

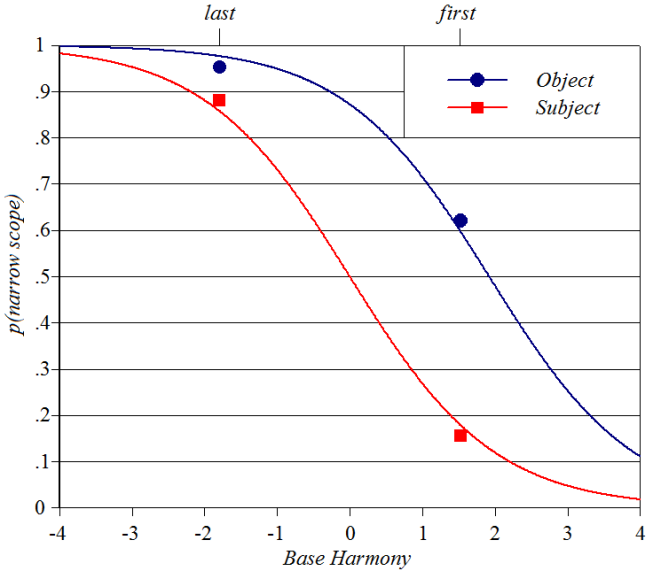
In addition, I feel that the discussion here, set out as it is in a uniform framework, may help to affirm a meaningful connection between the research tradition in diachronic syntax established by Kroch and the work on synchronic syntax by Bresnan and colleagues (§5.4). It is tempting, at least to me, to place the same theoretical gloss on both research traditions: they are both accessing a fundamentally sound theoretical conception in which grammatical patterns, including probabilistic ones, emerge from simple interaction of conflicting constraints (Legendre et al. 1990, Bresnan 1998). Within the native speaker's grammar, the constraints interact abstractly in the Harmony domain. In the observable world of speech output, this results in output frequencies that plot as wug-shaped curves, either diachronic (Kroch) or synchronic (Bresnan).

### *5.6 Semantics/Pragmatics*

**Quantifier scope ambiguities** occur in sentences like *A student saw every professor*. Corpus study (e.g. AnderBois et al. 2012) suggests that an appropriate system for predicting quantifier scope is likely to be a probabilistic one: judgments reflect a blend of conflicting factors, and my sense is that they are gradient, not categorical.

I offer here a tiny curve based on a subset of AnderBois et al.'s data. The Baseline reflects linear order (leftward favors broad scope) and the Perturber reflects grammatical relations (subjects prefer broad scope). The fit seems good (unimpressively so, given that they are few data points; but for a failed rival model see §6.1). The graph does display the narrowing of the influence of the Perturber at the periphery that MaxEnt predicts.[18]

(18) *A wug-shaped curve in quantifier scope, adapted from AnderBois et al. (2012:107)*




## 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.

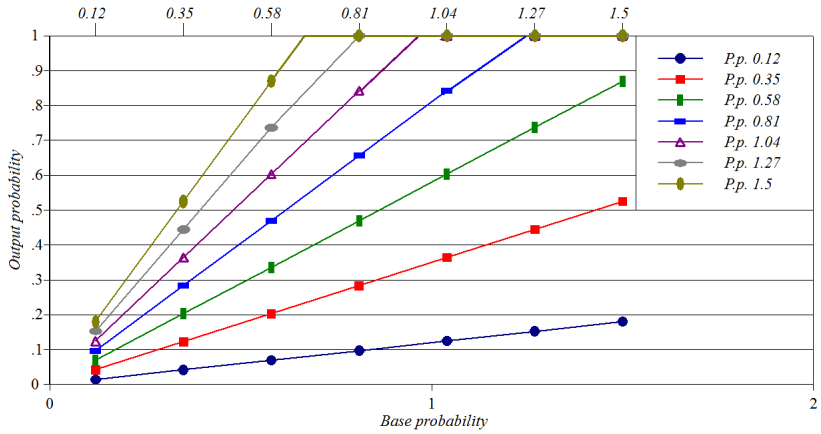
### *6.1 Some simple alternatives to MaxEnt*

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/logistic regression (Cedergren and Sankoff 1974, Sankoff and Labov 1979).

[18] AnderBois and Brasoveanu's paper make an important further point about scope judgments; namely that they are sensitive to real-world knowledge (e.g. *A doctor lives in every city*, influenced by the fact that people seldom live in more than one city). A successful predictive system for scope judgments therefore must probabilistically integrate linguistic principles with broader cognitive ones; see §7.2 for further discussion.

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 (19) I give a schematic quantitative signature of this approach, assuming seven Baseline probabilities and seven Perturber probabilities (P.p.).

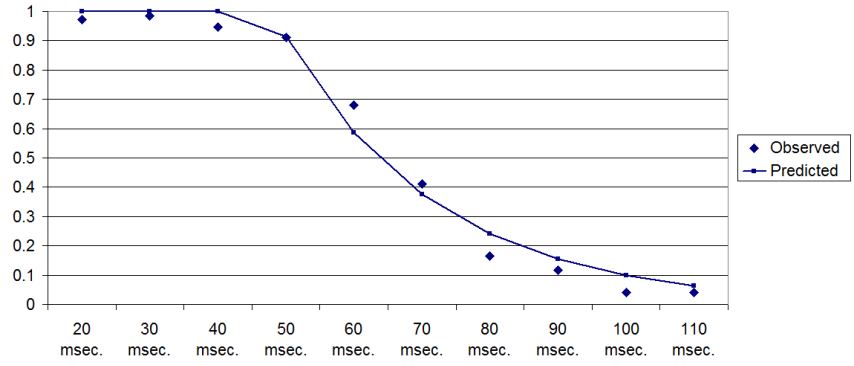
(19) *Quantitative signature of the Multiplication-cum-Cutoff model*




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, as in (11). What we find is a curious shape, with a sharp shoulder on on side but a gentle curve on the other. The curve as fitted to the Kluender et al. data (post-long-vowel series) looks like (20).

(20) *Fitting the Multiplication-cum-Cutoff model to the data of Kluender et al. (post-long vowel series)*

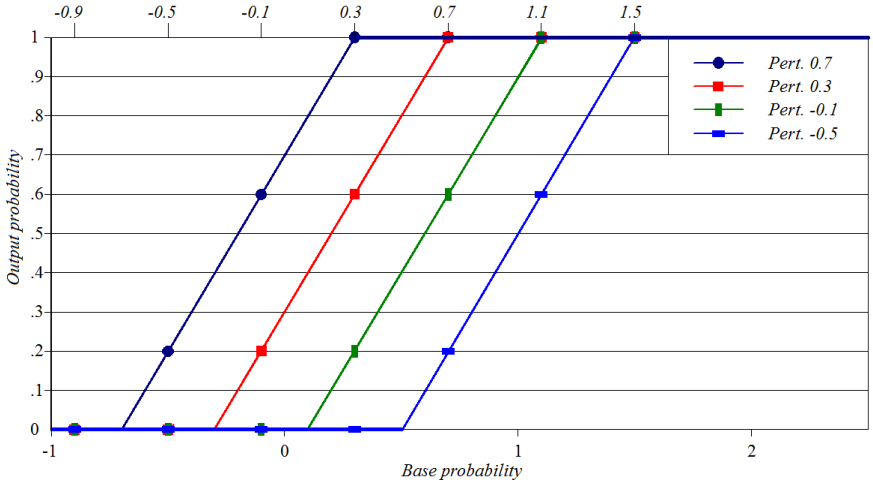



The data misfit is admittedly modest, but the typological implication seems wrong — to my knowledge, nowhere in the speech perception literature is it claimed that the curves from identification experiments are systematically asymmetrical in this way; and the same would hold

for the literature on historical-change sigmoids. For further discussion of the problems with this model, see Sankoff and Labov (1979).

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. This model was put forth as a straw man by Cedergren and Sankoff (1974). Its quantitative signature is the "Z-shaped curve," with parallel lines going diagonally between the cutoffs and sharp angles at the transition.

(21) *Quantitative signature of the Addition-cum-Cutoff model*




Where MaxEnt smoothes off the ends, gradually leveling the slope, Addition-cum-Cutoff crashes into the limits at zero and one. In actual model-fitting, this difference often produces only trivial differences in accuracy, because the noise present in almost any data means that it is hard to prove that the ends of the sigmoid really are smooth rather than angular.[19] However, even the very simple data of (18) are modeled poorly in Addition-cum-Cutoff, as the two lines "want" to have different slopes.[20]

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. Section 2.2 above 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,

[19] The closest thing I have seen to a data pattern that looks like (21) is the Pokémon data, as originally created by company employees, given in Kawahara (2020). However, the fit of the Addition-cum-Cutoff model in this case is only marginally better than the fit of MaxEnt model.

[20] 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.

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*

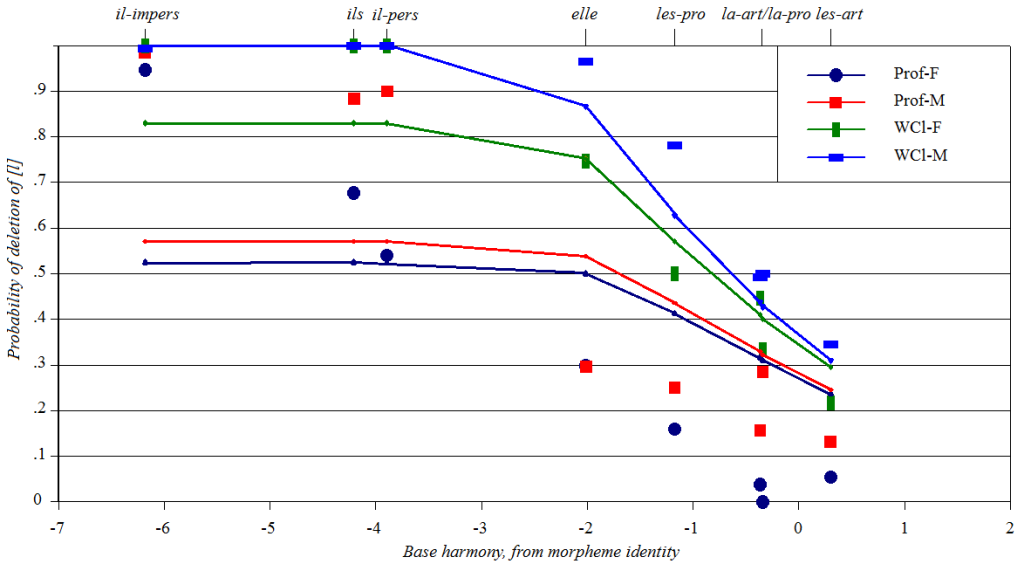
In Stochastic OT (Boersma 1998), the key idea is that the content of the grammar is itself probabilistic: constraints comes with a number ("ranking value") that expresses how highly ranked they are in general, but each time the grammar is deployed ("evaluation time"), these ranking values are adjusted by a small random noise factor. The adjusted values are then used to sort the constraints, and at this point the choice of winning candidate follows classical OT. Repeated application of this procedure will yield an estimated probability distribution.

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).[21]

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 far apart on the ranking scale. This point is covered in detail in Zuraw and Hayes (2017; §2.6), but we can observe here that one of the cases discussed above makes the same point: Stochastic OT offers a poor fit to Bailey/Sankoff Québec French data (§5.2). Figure (22) below is the same as our (10), except with the Stochastic OT predictions superimposed instead of MaxEnt.

[21] 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.

(22) *The Sankoff/Bailey Québec French pattern with Stochastic OT predictions*




The ill-fitting flat regions are cases in which a perturber simply lacks the range to interact with constraints that have greatly different ranking values.

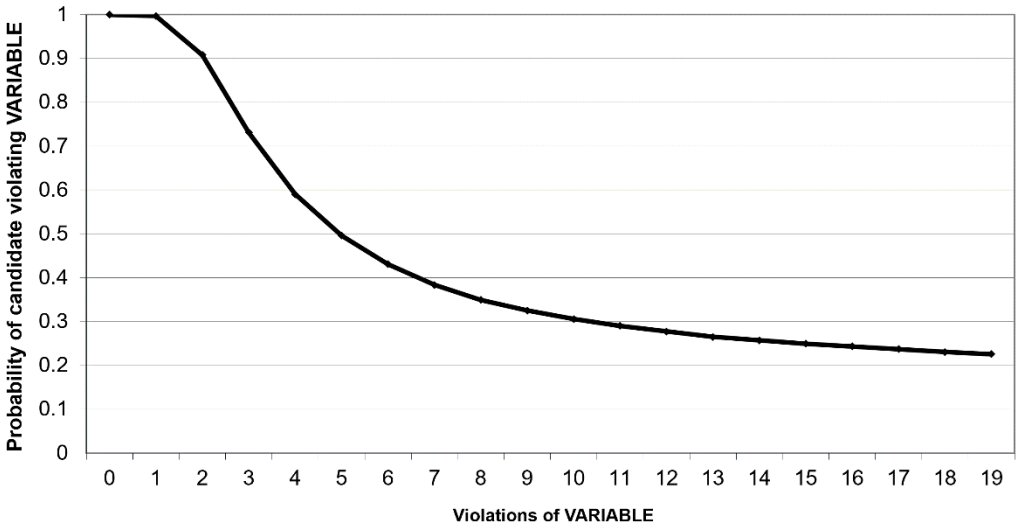
Both of these failings of Stochastic OT are rooted in traits it inherits from classical OT: contrary to principles (3b, c) above, it ignores relevant data, either in the form of violation counts, or of dominated constraints.

### *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 (Hayes 2017), which differ in which part of the calculation gets randomly tweaked: we can alter constraint weights, violations, tableau cells (violations times weight; as in Goldrick and Daland 2009) or the Harmony scores of candidates.

These types differ in their quantitative signatures. In the classical form of the model, we add the noise to the constraint weights, before any other calculations are done. This version is perhaps problematic, for its quantitative signature for variable violation-count cases has the same odd asymmetry that we saw above ((20)) for the Multiplication-cum-Cutoff model; this is given for a schematic set-up in (23) below.

(23) *An asymmetrical sigmoid generated in "classical" Noisy Harmonic Grammar*




The reason this happens, and an example of inferior fit to empirical data, is given in McPherson and Hayes (2016:156). The situation may actually be somewhat worse than for the Multiplication-cum-Cutoff, since it can be shown that the curve asymptotes on side to a positive value, never reaching zero.[22] I suspect this is not a good prediction to make in general.

A very different variety of Noisy Harmonic Grammar is obtained if we let the Harmony calculations be carried out non-stochastically, then at the end add a random noise value to the computed Harmony scores of the candidates. As Flemming (2017) demonstrates, this "late noise" theory is extremely close to MaxEnt, and the sigmoid curves it generates as a quantitative signature are virtually indistinguishable from the MaxEnt sigmoid (see figure (41) below for a comparison). When there is a Perturber constraint, we get multiple sigmoids, separated from each other by the weight of the Perturber. So this theory passes the test of being able to generate wug-shaped curves.

What clouds the picture for evaluating versions of Noisy Harmonic Grammar is that they also differ in a property important to many theorists: the classical version of the theory, when suitably deployed, gives zero probability to so-called "harmonically bounded" candidates, defined as candidates that have a proper superset of the violations of some rival candidate. Work such as Anttila and Magri (2018), Anttila et al. (2019) and Kaplan (in press) have argued for the classical variant on restrictiveness grounds. Neither "late noise" Noisy Harmonic Grammar nor MaxEnt respect the principle of harmonic bounding.[23]

[22] When calculated, the asymptote turns out to be the probability that a random Gaussian variable will have a value less than $-w_{Variable}$. Since in making chart (23) I used a value of 1 for this weight, probability will never go below 0.159, no matter how many times VARIABLE is violated.

[23] I seem to be one of just a few linguists who advocate MaxEnt analyses that actually *depend* on assigning nonzero probability to harmonically bounded candidates; see Hayes and Wilson (2008), Hayes and Moore-Cantwell (2011), and Hayes and Schuh (2019).

### *6.3 Other models*

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 logistic regression (Baayen 2008, Johnson 2011), which controls for the idiosyncrasies of individual words or participants.[24] There are other models more elaborate than MaxEnt, such as neural network models (Goldberg 2017) or random forest models (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) and Storme (forthcoming) employ 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 are likely eventually to be incorporated into linguistic theorizing, to the benefit of linguistic theory.[25] 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 concrete about their behavior in this connection.

There is one factor leads me to suspect that more sophisticated models are likely to generate wug-shaped curves in general. What I suspect is special about MaxEnt is not its particular formula, but that this formula renders in concrete mathematical form a set of plausible and intuitive principles of inductive reasoning, as I tried to show in §2.2. *Any* mathematical system that likewise formalizes plausible inductive principles is likely to generate something like the MaxEnt sigmoid. For further discussion of this point, see Appendix D.

### *6.4 Interaction effects and conjoined constraints*

These are discussed in Appendix E.

## 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

[24] 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).

[25] For more on the connection between linguistics and statistics see Appendix F.

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) account of Harmony Theory, includes in its text 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.

# APPENDICES

## Appendix A: Deriving a single data point in Kluender (1988)

This example is meant to display the MaxEnt calculations with full explicitness as applied to one single data point from the Kluender et al. (1988) experiment discussed in §3.1. The grammar I set up for explicating this experiment assigns probabilities to the percepts [b] and [p] depending on the stop closure duration of the stimulus. We set up the calculations as follows.

*Input:*

- 60 ms of closure duration

*Candidate being evaluated:*

- The percept [b]

*Constraints:*

- VARIABLE penalizes the percept of [b] to the extent that closure duration deviates from the target value of 20 ms. In this case the deviation is 60 − 20 = 40.
- ONOFF penalizes the percept [p].

*MaxEnt formula:*

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

*Calculating the probability of percept [b]:*

| | |
|---|---|
| 60 − 20 = **40** | Violations of VARIABLE for the input 60 msec., candidate [b]: deviation from target at 20 msec. |
| **1** | Violations of ONOFF for candidate [p] |
| $H_b$ = .088 × 40 = **3.53**<br>$H_p$ = 4.34 × 1 = **4.34** | Multiply violations by weights to obtain Harmony. |
| $eH_b$ = exp(−3.58) = **.029**<br>$eH_p$ = exp(−4.34) = **.013** | Take *e* to the minus Harmony to obtain eHarmony. |
| $Z = eH_b + eH_p$ = .029 + . 013 = **.042** | Add eHarmony across viable candidates for the same input, obtaining Z. |
| Pr([b]) = $eH_b$/Z =.029/.042 = **.69** | Divide eHarmony for [b] by Z, obtaining its predicted probability |

The value obtained, .69, is felicitously close to the value .68 observed by Kluender et al.; the predictions for other data points turn out to be decent but not always as close. Such deviations from prediction are expected, given that the number of participants in the experiment was only 16 (sampling error) and modest errors arising from the experimental setup are inevitable.

## Appendix B: Coding Harmony as a single value in two-candidate systems

Throughout the graphs in the main text, Harmony is plotted on the x axis as a *difference* in the Harmony values of the two viable candidates. The key idea is that in a two-candidate system it is possible to allocate all the Harmony to just one of the two candidates and give the other a

constant Harmony of zero, yet still derive the correct output probabilities. Under these conditions, we can create simple graphs in which the one single value of Harmony is plotted on the x-axis.

For simplicity, let us work with a concrete example, the phonetic perception experiment of Kluender et al. (1988) covered in §4.1 and Appendix A above. The two viable candidates in this case are the percepts [b] and [p]. We calculate the probability of [b] as in (24).

(24) *How to allocate all the Harmony to a single candidate in a two-candidate competition*

| | |
|---|---|
| $\Pr(\text{b}) = \frac{\exp(-\Sigma_i w_i \text{f}_i(\text{b}))}{Z}$, where $Z = \Sigma_j \exp(-\Sigma_i w_i \text{f}_i(x_j))$ | The MaxEnt formula, from (1) |
| $= \frac{\exp(H_b)}{Z}$, where $Z = \Sigma_j \exp(-\Sigma_i H_j))$ | Let the result of the calculations of Harmony be depicted using a single cover symbol *H*. |
| $= \frac{\exp(H_b)}{\exp(H_b) + \exp(H_p)}$ | Approximate Z, the sum of the eHarmony of all candidates. There are just two viable candidates [b] and [p]; all others contribute essentially zero eHarmony to the total. |
| $= \frac{1}{1 + \exp(H_p)/\exp(H_b)}$ | Divide top and bottom by $\exp(H_b)$. |
| $= \frac{1}{1 + \exp(H_p - H_b)}$ | Division of exponentiated quantities is the same as subtracting exponents. |

Given the perceptual constraints adopted in §4.1, the harmony difference $H_p - H_b$ will be a simple function of the closure duration of any particular input (i.e., value of closure duration in msec.); and, as the derivation shows, this number suffices for computing *P*(b). *P*(p) is also obtained, since it is (infinitesimally close to) $1 - P(\text{b})$. Thus, by putting $H_p - H_b$ on the *x* axis of graph (5), we can obtain a simple two-dimensional figure that gives the analytic basis of the MaxEnt sigmoid; and the same holds for all the analyses in this paper.

**Appendix C: Why does language change occur at a constant rate?**

The key finding discussed in §5.5, due primarily to Kroch (1989), is that syntactic changes take place at a constant rate, provided we measure them according to Harmony rather than raw frequency. Assuming this is true, it is a really remarkable fact, since the speakers involved in a change can span several centuries, forming a chain of people unknown to one another except at any particular stage. What would be a non-miraculous explanation?

Useful work has been done on this problem by Blythe and Croft (2012) and Stadler et al. (2016). A perhaps overly-free paraphrase of their proposal is as follows:

> *Language changes that progress steadily are the work of adolescents, who are in the process of fixing their grammars into adult form. They take as their role models speakers somewhat older than themselves, and they exaggerate to some degree the ways in which these slightly-older speakers themselves differ from full-grown adults. Since all generations of adolescents are roughly alike in this respect, the rate of change tends to be constant.*

My restatement is qualitative, but the authors just cited back up their proposals with quantitative modeling.

In the context established in the main text, it is tempting to suggest that exaggeration takes place on a scale based on the natural units of grammar, Harmony. To be more concrete:

> *The younger adolescent, perceiving the outputs of older adolescents and adults, detects the Harmony difference between the two groups for each constraint, and extrapolates this difference to her own age in determining constraint weight she uses.*

Thus, to be far more concrete than any facts can yet tell us: if a 14-year-old notices that 17-year-olds employ a weight of 4.1 for Constraint V (our Variable constraint) and that adults employ a weight of 4, then the 14-year-old would employ a weight of 4.2 in her own speech. This would increase the weight of V at a rate of 2.5 units per century; i.e. a constant-rate change over time. If in contrast, for constraint P, a Perturber, both 17-year-olds and adults have a weight of 3, then the 14-year-old selects 3 as well, maintaining stability.

Stadler et al. are aware of MaxEnt/logistic models, but in their work they employ instead the Multiplicative-cum-Cutoff model described (as problematical) in §6.1. However, they note that it doesn't seem to matter much what particular imitation mechanism they employ in their model. The issue of whether adolescents extrapolate in the Harmony domain remains one to be settled by research.

## Appendix D: Deriving the sigmoid curve from first principles

The main body of the paper attempted to demonstrate that the wug-shaped curve, a combination of two or more sigmoid probability functions, is a pervasive pattern seen in all areas of linguistics, and moreover is a natural consequence of adopting MaxEnt Harmonic Grammar. However, we are entitled to ask, "*if* the MaxEnt math is so pervasively appropriate, *why* should this be so?" The shortcut answer is to say that apparatus for implementing the MaxEnt math is innate in humans, and we should try to spot this apparatus somehow in our chromosomes. Perhaps this is true, but it begs the question of how an innate capacity for MaxEnt could ever have become genetically encoded in the first place. From this perspective, it seems potentially helpful pursue the discussion of §2.2 above from another perspective: to what extent does adopting principles of effective inductive reasoning like (3) lead us, more or less inevitably, to MaxEnt-like sigmoids and to wug-shaped curves? The idea pursued here is to search through the set of conceivable mathematical procedures, narrowing them down to the ones that are inductively sensible.

*D.1 The sensibleness of Harmonic Grammar per se*

To keep the discussion concrete, I will assume that all theories under consideration start by computing Harmony for each candidate. Recall that Harmony is the sum of the products of violations and weights for each constraint. This embodies two abstract principles of inductive reasoning:

(25) a. Reasons to justify a conclusion differ in how convincing they are.
b. A subset of a body of valid evidence is less convincing than the full set.

In Harmonic Grammar, principle (25a) is implemented by having the grammar include a specific weight for each constraint, as already discussed. Principle (25b) is implemented in two ways: multiplication of violations by weights (more instances of the same reason are more convincing than fewer),[26] and summation of the contributions of all constraints (no constraint has its testimony ignored).

Plainly, systems other than Harmonic Grammar could also implement (25), but I will not explore them here. Instead, the focus is on the second part of the problem, the mapping from Harmony to probability.

*D.2 Exploring the function mapping from Harmony to probability*

Let the function that maps from Harmony to probability be called $P(\mathrm{H})$. For MaxEnt, whose formula is repeated in (26), $P(\mathrm{H})$ is computed by calculating the eHarmony of every candidate, then assigning probability as the share of a candidate's eHarmony in Z, the total eHarmony of all candidates for an input.

(26) *The MaxEnt formula (repeated)*

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

The function we will focus on (for the particular candidate $x$) is obtained by simplying (26), leaving out the parts that compute Harmony and instead just list the Harmony values as elements of the formula, as in (27).

[26] A caution: in the case of gradient (multiply-violable) constraints, we must be careful about deciding what constitutes "more evidence". For instance, in assessing that a stop is [p], not [b], long Voice Onset Time is a strong cue favoring [p] over [b], and indeed many perceptual experiments show that the probability of hearing [p] increases as VOT increases. Yet after a certain point we will surely reach saturation, with any further increase of VOT no longer helpful (a VOT of 2000 msec., for instance, will probably heard more as respiratory distress than any particular speech segment). Such facts, being phenomenon-specific, would plausibly be entered into the pattern of constraint violations, not into the MaxEnt math itself.

(27) *The MaxEnt formula simplified (expressed as a function of Harmony)*

$$P(H_x) = \frac{\exp(-H_x)}{\sum_j \exp(-H_j)}$$

As was shown in §3, the MaxEnt version of *P*(H) always creates a sigmoid curve for two-candidate systems, when we plot their difference in Harmony on the horizontal axis. The MaxEnt sigmoid was displayed above in (4).

Next, instead of deriving the sigmoid from principles of MaxEnt, let us consider the conditions on *any* function *P*(H) that will produce a sigmoid of similar character.

*D.3 Reasoning from intuitive principles to mathematical properties of P*(H)

The first point at hand is almost too obvious to notice: the very choice of *probability* as the means of measuring of our confidence in an inductive conclusion automatically satisfies a plausible intuitive principle, already discussed in §2.2.5.

(28) A candidate should become less probable when it competes with powerful rivals.

This is because probabilities form an exhaustive set of alternatives must sum to one.

Returning to (25b), "A subset of a body of evidence is less convincing than the full set," we have the first in a list of "consequences"; i.e. implications from common sense principles to properties of *P*(H):

(29) *Consequence of (25b)*

> *P*(H) must be monotonic, i.e. have an always-positive or always-negative slope.

This is because we have quantified evidence as Harmony; only a monotonic function of Harmony will satisfy (25b). Whether the slope of *P*(H) is positive or negative depends on which of the two candidates is having its probability plotted.

We next appeal to another principle discussed in §2.2.4:

(30) Evidence should be scaled to make it have less effect as we approach certainty.

This tells about the *derivative* of *P*(H):

(31) *Consequence of (30)*

> The slope of *P*(H) must be greatest for some medial value, and it must steadily become more shallow toward the peripheries.

Since probability is constrained to be between zero and one, the slope must in fact tend toward zero; else a monotonic function would break out of the 0-1 range and fail to form an asymptote.

I further suggest the following principle as sensible:

(32) Evidence in ever-larger quantities is, eventually, decisive.

For every decision, there exists, in principle, a quantity of evidence that would count as persuasive. It may be vast, or not available in the real world, but at least we can *imagine* what evidence would prove to be decisive. If (4) is correct, this tells us more about the requirements on the function (H): not only must its slope approach zero toward the ends, but the actual values must approach their limits.

(33) *Consequence of (32)*

The values to which *P*(H) asymptotes must be zero and one.

Another principle that seems intuitive to me is (34):

(34) Infinitesimal differences in evidence may only yield infinitesimal difference in probability.

In more technical terms, we say:

(35) *Restatement of (34)*

*P*(H) must be continuous, in the mathematical sense

Taking this a step further, we might suppose that *sensitivity to small differences in harmony*, mentioned in fn. 20 in connection to speech perception, is likewise continuous. This means that infinitesimal differences in harmony cannot create supra-infinitesimal differences in sensitivity. In mathematical terms:

(36) The derivative of *P*(H) must be continuous.

At this point, we are in a position to critique some concrete choices for *P*(H): any model that implements the probability minima and maxima at 0 and 1 with a cutoff (e.g. Addition-cum-Cutoff, Multiplication-cum-Cutoff, §6.1) will have a discontinuous derivative at the point when the function gets truncated, and will fail this criterion.

Lastly, we go further out on a limb, groping for a reason to make our curve *symmetrical.* I suggest (37):

(37) Nothing other than proximity to certainty may influence the probability function.

From this tenuous hypothesis the following conclusion would follow:

(38) *Consequence of (37)*

*P*(H) is symmetrical.

where "symmetry" is defined by corresponding upward or downward deviations from 50% probability at equal distances from the symmetry point. I note that here we are probably on the weakest ground empirically; it is not easy to demonstrate that the right sigmoid is always symmetrical.

Assuming (38), the point of symmetry would have to be the point of maximum slope established earlier in (31), since otherwise one putative "half" of the symmetrical pattern would include a region whose slope is unmatched in the other putative half; thus we have (39):

(39) *Consequence*: The point of symmetry for *P*(H) is the point of maximum slope.

*D.4 Result*

Putting all of these principles together, we end up with a fairly narrow criterion for *P*(H), the function that maps Harmony to probability:

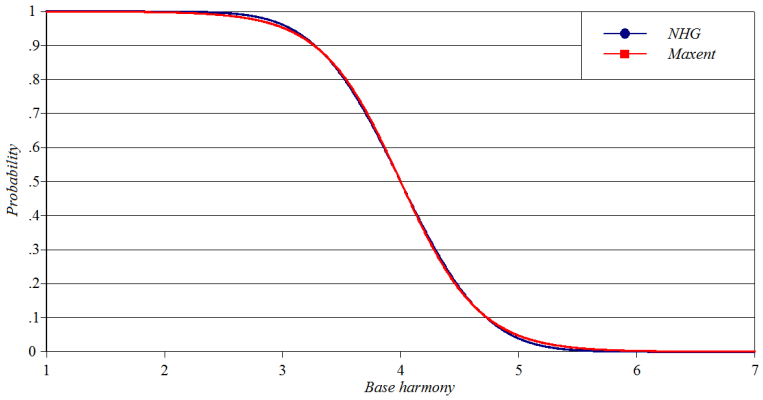
(40) *Plausible limitations on P*(H)

a. *P*(H) must be monotonic ((29)).
b. *P*(H) must have maximum slope at some medial value, with monotonically diminishing slope toward either periphery ((31)).
b. *P*(H) must asymptote at zero and one ((33)).
c. *P*(H) must be continuous with a continuous derivative ((35), (36)).
e. *P*(H) is (with greater doubt and qualms here) symmetrical about the point of maximum slope ((38), (39))

This is a fairly stringent description, which is satisfied by the two candidates for *P*(H) advocated in the main text. For MaxEnt, *P*(H) is the logistic function, plotted in (4) and shown here to have all the relevant properties. However, the *P*(H) function obtained in late-noise Noisy Harmonic Grammar (§6.2.2), which is the cumulative distribution function of the normal distribution, also has all of these properties.[27] In fact, the two functions employed in these theories look very similar to the eye, despite their completely different mathematical origin. Graph (41) shows the superposition of a MaxEnt and a Noisy Harmonic Grammar sigmoid whose parameters were set to match each other.

[27] See, for example, https://reference.wolfram.com/language/ref/NormalDistribution.html, which includes among other graphs a wug-shaped curve based on this function.

(41) *The sigmoid curves of MaxEnt and Noisy Harmonic Grammar compared*




However, it is not at all a trivial matter to find a *P*(H) function that satisfies all the criteria of (40). Most do not, and I have taken pains to critique some actually-employed theories along these lines. In particular, *P*(H) for the Additive Model (§6.1) does not have a continuous derivative, nor is its slope monotonically diminishing toward the periphery. *P*(H) for the Multiplicative model (§6.1) shares both traits and in addition is asymmetrical. *P*(H) for classical Noisy Harmonic Grammar, in cases with multiple violations of the Variable constraint ((23)), is asymmetrical and asymptotes above zero.

Are the criteria discussed above adequate to the task of defining sensible inductive reasoning? I am not sure, and am curious to know if there are other candidates for *P*(H) that satisfy the criteria of (40) yet are empirically implausible.[28]

To finish the story, we can observe that once we establish that a candidate function *P*(H) generates a MaxEnt-like sigmoid (a "quantitative signature" of the underlying theory), it trivially follows that it will generate our other quantitative signature, the wug-shaped curve. This follows because, by fiat, we are only considering models that compute a Harmony value. In such a system, a Perturber will always create an identical sigmoid, shifted over by a constant amount.

## Appendix E: Interaction terms and conjoined constraints

It is standard for scholars using MaxEnt/logistic regression for purposes of statistical testing to check *interaction terms*. For example, given two choices A vs. B, C vs. D, it can turn out to be the case that the choice between C and D comes out differently when A is true than when B is. Testing for the interaction of the A/B and C/D factors can inform us how likely it is this scenario is in effect.

Something similar to interaction terms is also employed in Optimality Theory, under the title of "conjoined constraints" (Smolensky 1995); i.e. constraints that are composed of two preexisting constraints X and Y, and are violated only when both X and Y are violated. Deployed

[28] I am aware of several other established mathematical functions that (suitably rescaled, in some cases) satisfy the criteria of (40). These include the arctangent, the hyperbolic tangent, the Gauss error function, and the unnamed function $y = x / \text{sqrt}(1+x^2)$; see https://psychology.wikia.org/wiki/Sigmoid_function.

in an OT model in which the selection procedure is MaxEnt, these are close to being conjoined constraints.[29] An argument that conjoined constraints are needed specifically in MaxEnt is given by Shih (2017); and I have checked the examples studied in this paper (using the bayesglm() function in R; Gelman et al. 2010) for whether a model with improved fit can be obtained by adding interaction terms to the analysis, and this is indeed sometimes the case. Given that interaction terms are useful, do they vitiate my claim about MaxEnt deriving wug-shaped curves?

To be clear up front: free use of local conjunction will indeed wipe out the possibility of making strict predictions about quantitative signatures. For any of the data points in any of the cases covered in this paper, we could add a conjoined constraint that covers exactly that data point, shifting the prediction made by the model to any value we please; hence at some level the framework ends up saying nothing.

However, the point just made does not match up well with ordinary practice in linguistic data analysis. Everywhere, we seek explanations for linguistic phenomena, under the assumption that these phenomena have causes. To group together two arbitrary factors into a conjoined constraint does not seem very helpful in the search for explanation. Rather, if we are to use conjoined constraints, we ought to be able to provide a *substantive reason* why the presence of (say) C makes A better relative to B.

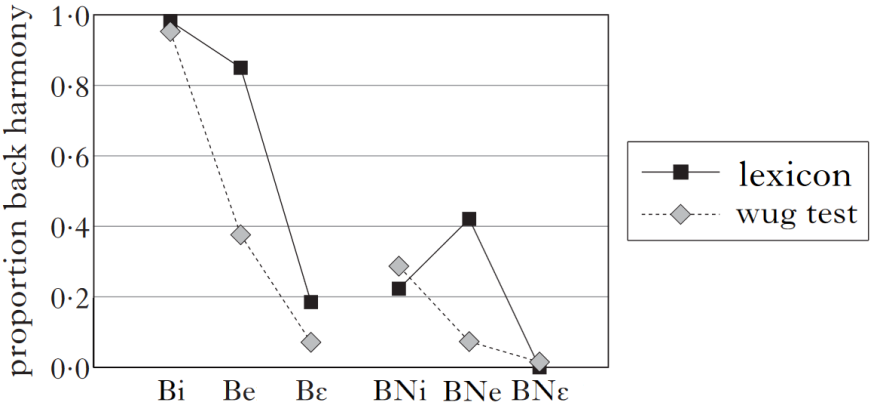
To give an example, in Japanese, there is a strong tendency to avoid voiced obstruent geminates (double consonants). Is this because the goal is to avoid two bad things at once? Many languages do indeed avoid voiced obstruents, and many languages avoid geminates, so the interaction-term approach has some appeal. Yet closer scrutiny of the phonetics of voicing difficulty, as in Westbury and Keating (1986), gives a *substantive* reason, based on airflows and vocal tract properties, to think that voiced obstruent geminates pose a special extreme of difficulty in maintaining voicing, which has been documented experimentally by Kawahara (2006). In other words, we would have good phonetic justification for setting up the "conjoined" constraint *[+voice, +obstruent, +long] even in the absence of any information that [+voice, +obstruent] and [+long] are themselves avoided to a lesser extent. *[+voice, +obstruent, +long] is plausible as a *simplex* constraint with richer internal specification, and this fits the facts of phonetics more directly than an appeal to constraint conjunction would.[30]

Other than this, I think it pays to insist that for any pattern putatively illustrating the effects of constraint conjunction, we should show that the pattern is actively internalized by language learners, rather than being a mere blip in the lexicon. A case in support of this view arises in Hayes and Londe's (2006) study of Hungarian vowel harmony; see their Fig. 5, repeated below as (42).

[29] However, the OT version of conjunction is usually *local* conjunction: *A & B is violated only when the violations of A and B are in the same location in the string, in some formally defined sense.

[30] For a much earlier insistence on substantive support when positing interaction terms, see Sankoff and Labov (1979:205).

(42) *Smoothing a sharp corner in a Hungarian vowel harmony wug test (Hayes and Londe 2006, Fig.5)*




Without going into excessive detail, we can say that the symbols B, [i], [e], [ɛ], and N represent vowels or classes of vowels that determine or influence the outcome of Hungarian vowel harmony. The key data point to observe is the black square in the "BNe" column, which reflects a gross asymmetry obtained when we look at the stems of the Hungarian lexicon that fit this description. The deviation is statistically significant, and is capturable only in a model that includes a conjoined constraint reflecting the two fundamental factors present (that is, [e] vs. other vowels, BX vs. BNX).

The striking fact is in Hayes and Londe's wug-test study with Hungarian native speakers, the outlier disappeared entirely — see the corrresponding gray data point in (42). The wug-test data are well fitted by using *simplex* constraints that reflect the distinctions just given (the result was replicated in a later wug test; Hayes et al. 2009). It seems that the blip represented in the lexical data, which presumably is accessible to Hungarian learners during the period of language acquisition, is ignored by them entirely. The conclusion we might draw is that Hungarian language learners have implicitly chosen to eschew constraint conjunction and deploy only simplex constraints, even though this leads to a less accurate characterization of the pattern of their language.

To summarize the discussion: liberal use of interaction terms/constraint conjunction produces a theory that has no quantitative signature and thus (at least from one point of view) is uninteresting. But careful attention to cases where putative conjoined constraints are justified on external grounds (essentially, to be treated as simplex), and insistence that putative cases of interaction be backed by evidence from psycholinguistic testing, may result in a more coherent and optimistic picture.

**Appendix F: Do findings drawn from the field of statistics fall within the province of linguistic theory?**

When I was a linguistics student in the 1970's, I would have found it unimaginable that statistics could become part of linguistic theory. This was partly the consequence of the statistics that was taught to undergraduates at the time, which seemed (to me at least) to consist solely of a

set of methods experimentalists might use to avoid error — a matter of good scientific hygiene but hardly of theoretical interest.

However, statistics kept evolving; in the intervening decades it seems to have become a far more lively, exploratory research activity.[31] Nowadays, statistics does better at extracting valid conclusions from noisy data than it used to; the real world, even baseball, attests to this. And extracting valid conclusions from noisy data is precisely what young human language learners must do. If there exist rational and effective mathematical modes of inductive reasoning, then it is not unreasonable to suppose that natural selection has equipped human beings to use some analogue of these principles.

On the other side of the gap, linguistics has developed in important ways that let it engage more closely and more naturally with statistical methods. A clear example is Optimality Theory and its probabilistic descendents, in which linguistic knowledge is deliberately "atomized" to the point that statistical principles can engage with it effectively. Such theories leave plenty for the linguists to do, for instance in understanding the families of constraints and their origin or in understanding the form of linguistic representations. But I think we should be happy to import our forms of probabilistic reasoning from other fields, if that is what turns out to work best scientifically.

Approaching the same issue from the other side, I feel there is also a need to incorporate linguistic theory into statistical data analysis. Some of the research I have read for purposes of writing this article strikes me as unusually agnostic with regard to theory: it is quite easy in practice to adopt purely-empirical classifications of the facts, plug them into a good statistical model, and obtain accurate results. I that our understanding will advance more rapidly if we try to use constraints that are themselves the result of extensive theoretical development and typological testing.

## References

AnderBois, Scott, Adrian Brasoveanu, and Robert Henderson (2012) The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung* 16:15-28.

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.

[31] A popularization I have enjoyed is McGrayne (2011), which conveys some of the liveliness of modern statistical inquiry as well as a sense of drama about how the field came to evolve into its present form.

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 (1998) Morphology competes with syntax: Explaining typological variation in weak crossover effects. Is the best good enough? Optimality and competition in syntax, ed. by Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky, 59–92. Cambridge, MA: MIT Press and MIT Working Papers in Linguistics.

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 (2017) Stochastic harmonic grammars as random utility models. Paper given at the 2017 Annual Meeting in Phonology.
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.
Gelman, A., Su, Y., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., & Zheng, T. (2010). Package 'arm'. Available at:http://cran.r-project.org/web/packages/ar.
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 Zsuzsa Londe (2006) Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23:59-10.
Hayes, Bruce and Claire Moore-Cantwell (2011) Gerard Manley Hopkins's sprung rhythm: corpus study and stochastic grammar. *Phonology* 28:235–282.
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 (2018) Positional licensing, asymmetric trade-offs and gradient constraints in Harmonic Grammar. *Phonology* 35:247-286.

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

2006

Kawahara, Shigeto (2006) A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language* 82:536-574.

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.

Lefkowitz, Michael (2017) Maxent harmonic grammars and phonetic duration. Ph.D. dissertation, Department of Linguistics, UCLA, Los Angeles, CA.

Legendre, Geraldine, Yoshiro Miyata & Paul Smolensky (1990) Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

McGrayne, Sharon Bertsch (2011) *The theory that would not die.* New Haven: Yale University Press.

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.
Oliveira e Silva, Giselle (1982) Estudo da regularidade na variacao dos possessivos no portugues do Rio de Janeiro. Ph.D. dissertation, Universidade Federal do Rio de Janeiro.
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.
Shih, Stephanie (2017) Constraint conjunction in probabilistic weighted constraint grammar. *Phonology* 34:243-268.
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.
Stadler, Kevin, Richard A. Blythe, Kenny Smith, and Simon Kirby (2016) Momentum in language change: A model of self-actuating S-shaped curves. *Language Dynamics and Change* 6:171–198.
Storme, Benjamin (forthcoming) Not only size matters: limits to the Law of Three Consonants in French phonology. To appear in *Glossa*.
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.
Westbury, John R., and Patricia A. Keating (1986) On the naturalness of stop consonant voicing. *Journal of Linguistics* 22:145-166.

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 (1969) *A sociolinguistic description of Detroit Negro speech*. Washington, D.C.: Center for Applied Linguistics.

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.